Information Epidemiology and Surveillance in the Google Era

Amaryllis Mavragani

Department of Computing Science and Mathematics

University of Stirling

Ph.D. in Computing Science

May 2021

To my late grandmother, Carla

To the late Thanases Pheidas, my Teacher

Acknowledgements

First, I thank my Supervisor, *Prof. Gabriela Ochoa*, for the opportunity, for her insightful guidance throughout my Ph.D., and for her valuable support.

I would also like to thank *Prof. Konstantinos Tsagarakis* for all his guidance and support throughout these years.

Above all, I thank my husband and daughter, who gave me the motivation to continue trying for the best each and every day.

Finally, I want to thank my best friend, *Dimitra Totikidou*, who stood by me every step of the way, with her advice, patience, long discussions, and happy moments.

Declaration of Authorship

I, AMARYLLIS MAVRAGANI, declare that this Thesis entitled '*Information Epidemiology and Surveillance in the Google Era*' and the works presented in it are my own and are products of my own original research.

The papers presented in this Thesis were written and published while I was a Ph.D. candidate at the Department of Computing Science and Mathematics, University of Stirling, and they have not been previously submitted for any other degree or qualification at the University of Stirling or any other Institution.

In all publications presented in this Thesis, I am the first and leading author; I conceived the idea, designed the research, performed the analysis, and wrote the papers.

Signature

Abstract

Information epidemiology (infodemiology) approaches are increasingly employed in exploring online behavior and in predicting/forecasting diseases/epidemics, providing real time information and the revealed instead of the stated users' interests/preferences that are not otherwise accessible, thus tackling issues of traditional data collection and monitoring.

This Thesis examines how users' Google behavior towards health topics can be useful in public health epidemiology and surveillance. Studying the state of the art in 2017, gaps identified included an up-to-date systematic review, a methodology framework for rigorous data collection and reporting, as well as limited number of approaches in predictions/forecastings and several public health topics that had not been studied before.

To fill the gaps and advance the topic, this Thesis, consisting of 8 interconnected papers, includes: a systematic review of Google Trends in health/medicine categorized by methodology approaches; a methodology framework for rigorous data collection and reporting; six research papers in public health topics, namely COVID-19, STIs, Measles, and asthma, employing basic statistical tools to explore associations, predictability, and forecastings. Several factors limiting the applicability of this approach were also identified and discussed, e.g., lack of small interval health data, misspellings, sudden events.

The collective results could have significant implications for effective policy making, suggesting how multidisciplinary approaches in public health epidemiology and surveillance could make full use of the information and web tools that are available. The latter was especially evident during the COVID-19 pandemic -with open access to real time data- when such approaches were employed for epidemiology and surveillance. During chaotic conditions like in pandemics/epidemics, when policy makers are required to make fast and important decisions, it is vital to proceed with a statistical understanding of Google Trends time series and the users' behavior in accordance with its real determinants, combining medical and non-medical parameters from a variety of research fields, that also take into account the public's awareness and online behavior towards the explored topics.

Keywords: Big Data; Public Health; Digital Epidemiology; Health Informatics; Infodemiology; Infoveillance; Google Trends

Table of Contents

List of Figures

g. **Figure 7.** Online interest in the term “Flu” over the past 5 years (a) worldwide and (b) in the United States.
h. **Figure 8.** Regional online interest in the term “Flu” at metropolitan level over the past 5 years in (a) California, (b) Texas, (c) New York, and (d) Florida.
i. **Figure 9.** Regional online interest in the term “Flu” at city level over the past 5 years in (a) Los Angeles, (b) Dallas, (c) New York, (d) Miami, (e) India, and (f) Greece.
j. **Figure 10.** Customized time range (a) from archive and (b) over the past week

[C] Integrating Smart Health in the US Health Care System: Asthma Monitoring in the Google Era

a. **Figure 1.** Equations for Holt-Winters exponential smoothing
b. **Figure 2.** Online interest by state in the term "asthma" from 2004 to 2015.
c. **Figure 3.** Monthly changes in online interest in the term "asthma" from 2004 to 2015.
d. **Figure 4.** Weekly changes in online interest in the term "asthma" for each year from 2004 to 2015.
e. **Figure 5.** Online interest by state in the term "asthma" per year from 2004 to 2015.
f. **Figure 6.** Google Trends (2004 to 2015) versus forecasts (2005 to 2020) in the United States.
g. **Figure 7.** Google Trends (2004 to 2015) versus forecasts (January 2016 to June 2017) in the United States.
h. **Figure 8.** Google Trends (2004 to 2015) versus forecasts (2005 to 2020) in California.
i. **Figure 9.** Google Trends (2004 to 2015) versus forecasts (2005 to 2020) in Texas.
j. **Figure 10.** Google Trends (2004 to 2015) versus forecasts (2005 to 2020) in Florida.
k. **Figure 11.** Google Trends (2004 to 2015) versus forecasts (2005 to 2020) in New York.

[D] The Internet and the Anti-Vaccine Movement: Tracking the 2017 EU Measles Outbreak

a. **Figure 1.** Worldwide Interest in ‘Measles’, ‘Mumps’, ‘Rubella’, and ‘MMR’ from 2004 to 2017.
b. **Figure 2.** Worldwide Interest by Country in Measles from 2004 to 2017 (gray indicates zero scoring).
c. **Figure 3.** Worldwide Interest by Country in Mumps from 2004 to 2017 (gray indicates zero scoring).

d. **Figure 4.** Worldwide Interest by Country in Rubella from 2004 to 2017 (gray indicates zero scoring).
e. **Figure 5.** Worldwide Interest by Country in MMR from 2004 to 2017 (gray indicates zero scoring).
f. **Figure 6.** Worldwide Online Interest in the term 'Anti Vaccine' from January 2004 to August 2017.
g. **Figure 7.** EU28 Online Interest in the English (blue) and Translated (red) Terms for 'Measles' from January 2004 to August 2017.
h. **Figure 8.** EU28 Population Coverage (%) of the 1st and 2nd Dose of the Vaaccine for Measles from 1980 to 2016.
i. **Figure 9.** EU28 Population Coverage (%) for the 1st Dose in 2016.
j. **Figure 10.** EU28 Population Coverage (%) for the 2nd Dose in 2016.

[E] Forecasting AIDS Prevalence in the United States using Online Search Traffic Data

a. **Fig. 1** Monthly Normalized Google Trends' Data for 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015
b. **Fig. 2** Monthly normalized Google Trends' Data for 'Diagnosis of HIV/AIDS (Topic)' and 'Management of HIV/ AIDS (Topic)'from January 2004 to December 2015
c. **Fig. 3** Online Interest by State for 'AIDS (Search Term)', 'AIDS (Illness)', 'Diagnosis of HIV/AIDS (Topic)', and 'Management of HIV/AIDS (Topic)'from January 2004 to December 2015
d. **Fig. 4** 'AIDS Prevalence' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015 (USA; Alabama–Idaho)
e. **Fig. 5** 'AIDS Prevalence' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015 (Illinois–Montana)
f. **Fig. 6** 'AIDS Prevalence' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015 (Nebraska–South Carolina)
g. **Fig. 7** 'AIDS Prevalence' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015 (South Dakota–Wyoming)
h. **Fig. 8** 'AIDS Diagnoses' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015
i. **Fig. 9** 'AIDS Deaths' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015

j. **Fig. 10** US states categorized by (a) correlations-estimated model's significance and (b) AIDS rates

[F] Infoveillance of Infectious Diseases in USA: STDs, Tuberculosis, and Hepatitis

a. **Fig. 1** Heat map of the online interest in the term 'Chlamydia' by state (2004-2016)
b. **Fig. 2** Online interest heat maps for the term 'Chlamydia' by state by year (2004-2017)
c. **Fig. 3** Heat map of the online interest in the term 'Gonorrhea' by State (2004-2016)
d. **Fig. 4** Online interest heat maps for the term 'Gonorrhea' by state by year (2004-2017)
e. **Fig. 5** Heat map of the online interest in the term 'Syphilis' by State (2004-2016)
f. **Fig. 6** Online interest heat maps for the term 'Syphilis' by state by year (2004-2017)
g. **Fig. 7** Heat map of the online interest in the term 'Tuberculosis' by State (2004-2016)
h. **Fig. 8** Online interest heat maps for the term 'Tuberculosis' by state by year (2004-2017)
i. **Fig. 9** Heat map of the online interest in the term 'Hepatitis' by State (2004-2016)
j. **Fig. 10** Online interest heat maps for the term 'Hepatitis' by state by year (2004-2017)

[G] Tracking COVID-19 in Europe: Infodemiology Approach

a. **Figure 1.** Worldwide heat map for total COVID-19 cases by country (as of March 25, 2020).
b. **Figure 2.** European heat map for total COVID-19 deaths by country (as of March 25th, 2020).
c. **Figure 3.** (a) Cumulative and (b) daily cases, recoveries, and deaths (Italy; February 15-March 24).
d. **Figure 4.** Changes in the Pearson correlation coefficients (*r*) for Italy.
e. **Figure 5.** Changes in the Pearson correlation coefficients (*r*) for Lombardy.
f. **Figure 6.** Changes in the Pearson correlation coefficients (*r*) for Spain.
g. **Figure 7.** Changes in the Pearson correlation coefficients (*r*) for Germany.
h. **Figure 8.** Changes in the Pearson correlation coefficients (*r*) for France.
i. **Figure 9.** Changes in the Pearson correlation coefficients (*r*) for the United Kingdom.

[H] COVID-19 predictability in the United States using Google Trends time series

a. **Figure 1.** Geographical distribution of worldwide COVID-19 cases and deaths as of April 18th

b. **Figure 2.** Geographical distribution of COVID-19 cases and deaths in the US as of April 18th
c. **Figure 3.** Heat maps of the worldwide and US online interest in "Coronavirus (Virus)"
d. **Figure 4.** Heat map of the (a) Pearson and (b) Kendall correlation coefficients by state
e. **Figure 5.** Radar chart of the (a) Pearson and (b) Kendall correlation coefficients by state
f. **Figure 6.** Heat map of β_1 of the predictability analysis models by state
g. **Figure 7.** COVID-19 and Google Trends data from March 4th to April 15th in the US

List of Tables

[E] Forecasting AIDS Prevalence in the United States using Online Search Traffic Data

a. **Table 1** Pearson correlation coefficients between 'AIDS Prevalence' and 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015
b. **Table 2** Pearson correlation coefficients between 'AIDS Diagnoses' and (a) AIDS (Search Term)' and (b) 'AIDS (Illness)' from January 2004 to December 2015
c. **Table 3** Pearson correlation coefficients between 'AIDS Deaths' and (a) (Search Term)' and (b) 'AIDS (Illness)' from January 2004 to December 2015
d. **Table 4** Regression coefficients and R2 for the estimated forecasting models for 'AIDS Prevalence'
e. **Table 5** Estimated Logarithmic forecasting models for USA and selected states
f. **Table 6** Coefficients α and β, and R2 for the estimated forecasting models for 'AIDS Diagnoses'
g. **Table 7** Coefficients α and β, and R2 for the estimated forecasting models for 'AIDS Deaths'

[F] Infoveillance of Infectious Diseases in USA: STDs, Tuberculosis, and Hepatitis

a. **Table 1** Correlations between Google Trends data and Chlamydia cases by state
b. **Table 2** Coefficients α, β, and R^2 of the linear regressions for Chlamydia cases
c. **Table 3** Correlations between Google Trends data and Gonorrhea cases by state
d. **Table 4** Coefficients α, β, and R^2 of the linear regressions for Gonorrhea cases
e. **Table 5** Correlations between Google Trends data and Syphilis cases by state
f. **Table 6** Coefficients α, β, and R^2 of the linear regressions for Syphilis cases
g. **Table 7** Correlations between google trends data and Tuberculosis cases by state
h. **Table 8** Coefficients α, β, and R^2 of the linear regressions for Tuberculosis cases
i. **Table 9** Correlations between Google Trends data and Hepatitis cases by state
j. **Table 10** Coefficients α, β, and R^2 of the linear regressions for Hepatitis cases
k. **Table 11** CDC reported cases for the infectious diseases included in AtlasPlus in 2016
l. **Table 12** CDC reported yearly rates in USA for the examined diseases from 2004 to 2016

Chapter 1: *Context Chapter*

1. Information epidemiology and surveillance

Infodemiology, i.e., information epidemiology, is defined as "*the science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy*" [1]. The first official mention of the term infodemiology according to PubMed, i.e., baring the term on the title of the article, was by Gunther Eysenbach in 2002 in the *American Journal of Preventive Medicine* [2]. However, infodemiology studies, i.e., assessment of health-related topics using Web-based data [3], can be traced back to 1996, while also two more studies in the mid '00s use the term (PubMed); one in 2004 where the quality of hospitals' websites was assessed [4], and one in 2006 showing that flu data from Google correlated with influenza cases [5].

Infodemiology approaches include nowcasting epidemics and outbreaks, surveillance of infectious diseases, and assessment of chronic conditions. Social media and search queries are the most popular sources for retrieving information from Web-based sources. The use of social media is constantly expanding [7], including more features and users, while search query data are of significant value, as they take into account the revealed and not the stated preferences [8-9], though methodology should be designed with caution to ensure the validity of the results [10].

How can the Internet assist in providing accurate as well as real-time assessment of health issues? According to infodemiology, Internet data can be used to inform public health and policy by monitoring the public's online behavior towards diseases, selecting the relevant available information, as well as exploring how the public reacts to health marketing campaigns. Popular social media data sources in infodemiology include Twitter [11-17], Facebook [18-22], Instagram [23-24], while queries from search engines are mostly retrieved by Google Trends [25-32], as well as Yandex [33-35], Baidu [36-37], Bing [38], Yahoo [39], and Daum [40-41]. Other popular sources include websites and platforms [42-45], blogs, forums, and online communities [46-52], while, what has also received attention lately, is mobile apps of certain health categories, e.g., asthma [53] and heart failure self-care management [54]. Several studies have combined two or more sources, as, for example, Facebook and Instagram [55], Facebook and Twitter posts [56], US newspaper media and Facebook [57], and Google and Wikipedia [58].

Internet (real-time) data have been shown to contribute to the analysis and prediction of diseases' outbreaks and epidemics. In particular, one of the most studied topics is that of influenza, where several data sources have been employed to predict and assess flu related topics [39-40, 59-76]. Epidemics and infectious diseases that have been analyzed and assessed using infodemiology and infoveillance approaches include HIV/AIDS [77-79], measles [80-83], and the Zika virus [84-87].

Infodemiology topics have also been the subject of research for several reviews such as in curable sexually transmitted diseases [88] and mental health disorders [89], and for individual data sources, like for search queries and social media [6], mobile phone apps [90], Twitter [91], and Google Trends [92]. Such analyses of several health topics and diseases cover a wide range, including conditions/diseases, epidemics/outbreaks, drugs, mental health, infectious diseases, and cancer.

Popular categories include drugs [39, 93-94] and cannabis/marijuana [95-97], depression/suicide [98-108], as well as smoking/tobacco [109-116], e-cigarette [117-126], and Hookah [127-130]. As far as chronic diseases are concerned, infodemiology and infoveillance studies have been useful in the assessment of diabetes [131-136] and multiple sclerosis [137-138], while other topics include breast cancer [139-142], fitness and diet [143-146], health care performance, evaluation and dissemination [147-148], and HPV [149-154].

Table 1 features an indicative list of health topics and subtopics that have been explored in the fields of infodemiology and infoveillance, and Table 2 consists of selected publications indicating popular data sources, i.e., Google, Twitter, Facebook (FB), Instagram (Insta), Other Social Media (Social), Blogs-Forums-Communities (Blogs), and Websites/Platforms (Websites).

Google and Twitter are very popular sources, with Twitter's advantage being that it can be also used for qualitative analysis, since it is an outlet for official reports and news (e.g., governmental, politicians, celebrities etc.). However, Twitter is not used by all Internet users (profiles, tweets, retweets etc.), contrary to search traffic data, as almost all users employ search engines. Furthermore, there is a significant rise in publications using data from other social media, e.g., Facebook and Instagram, indicative of the younger Internet users' preferences in social media. Researchers in this field should closely follow any future shift in the trends of use of Internet sources.

Table 1. Indicative topics in infodemiology and infoveillance

Category	Topics	References
Cancer	Breast Cancer	139-142
	Skin Cancer	27
	Lung Cancer	155
Conditions/Diseases	Diabetes	131-136
	Multiple Sclerosis	137-138, 156
	Epilepsy	157-158
Drugs	Drug Abuse/Misuse	23, 93-94, 159-160
	Cannabis/Marijuana	95-97
	Adverse Drug Reactions	77, 161-162
	Illicit Drugs	33, 163
	Pharmaceutical Drugs	68, 164
Epidemics/Outbreaks	Influenza	39-40, 59-76
	Zika	84-87
	Measles	80-83
	Dengue	165-166
	H1N1	67, 167
	H7N9	65, 72
	Norovirus	168
Infectious Diseases	HIV/AIDS	77-79
	HPV	149-154
	STDs	30, 169-170
Mental Health	Depression	98-103
	Suicide	104-108
	Schizophrenia	13, 18, 171
	Stress	172-174
Miscellaneous	Radiation	175-176
	Obamacare	177-178
	Abortion	38
	Sudden Infant Death Syndrome	179
	Perinatal Deaths	180
	Fitness and Diet	143-146
	Physical Activity	181-183
Smoking	Smoking/Tobacco	109-116
	Electronic Cigarette/Vaping	117-126
	Hookah	127-130

Table 2. Selected infodemiology and infoveillance studies categorized by data source

	Authors	Google	Twitter	FB	Insta	Social	Blogs	Websites
1	Abbe & Falissard [46]						✓	
2	Abdellaoui et al. [48]						✓	
3	Adawi et al. [184]	✓						
4	Allem et al. [118]				✓			
5	Anderson et al. [185]						✓	
6	Balls-Berry et al. [186]		✓	✓		✓		
7	Baltrusaitis et al. [63]							✓
8	Berlinberg et al. [29]	✓						
9	Bousquet et al. [187]						✓	✓
10	Carrotte et al. [188]	✓	✓		✓	✓		
12	Cherian et al. [160]				✓			
13	Colditz et al. [189]		✓					
14	Edney et al. [144]		✓	✓	✓			
15	García-Díaz et al. [190]		✓					
16	Genes et al. [191]		✓					
17	Gianfredi et al. [192]	✓						
18	Hendriks et al. [55]			✓	✓			
19	Huesch et al. [141]			✓				
20	Jones et al. [142]						✓	
21	Jung et al. [193]							✓
22	Kaminski et al. [194]							
23	Kandula et al. [69]	✓						
24	Konheim-Kalkstein et al. [52]						✓	
25	Kurzinger et al. [195]						✓	
26	Madden et al. [145]	✓						
27	Madden et al. [196]	✓						
28	Martinez-Millana et al. [197]	✓	✓	✓		✓		
29	Martins-Filho et al. [198]	✓						
30	Matsuda et al. [199]						✓	
31	Mejova et al. [22]			✓				
32	Mowery et al. [200]		✓					
33	Mukhija et al. [201]	✓						
34	Muralidhara & Paul [24]				✓			
35	Noll-Hussong [202]	✓						
36	Odlum et al. [14]		✓					
37	Oser et al. [133]						✓	
38	Park & Hong [203]							✓
39	Phillips et al. [25]	✓						
40	Rabarison et al. [204]		✓					
41	Radin & Sciascia [26]	✓						
42	Ricard et al. [98]				✓			✓
43	Roccetti et al. [56]			✓				
44	Samaras et al. [75]	✓						
45	Sanz-Lorente et al. [88]		✓	✓	✓	✓		
48	Seidl et al. [27]	✓						
49	Shi & Salmon [205]					✓		
50	Sinnenberg et al. [134]		✓					
51	Staal et al. [116]							✓
52	Tafti et al. [206]						✓	✓
53	Tana et al. [100]	✓						
54	Tougas et al. [207]		✓	✓	✓			
55	Vasconcellos-Silva et al. [140]	✓						
57	Vickey & Breslin [143]		✓					
58	Winchester et al. [208]	✓	✓			✓		
59	Wood et al. [209]	✓						
60	Zhang et al. [129]				✓			

2. Aims and Motivation

2.1. Google Trends in Infodemiology

Internet data can provide a large amount of information that could not be accessed through traditional surveillance methods such as surveys and registries. New methods and approaches are constantly explored in order to take full advantage of online sources. As infodemiology data can be retrieved in real time and thus allow the nowcasting of the users' search patterns and online behavior, the detection, monitoring, and prediction of epidemics and outbreaks can benefit from the analysis of Google queries, which is the topic of this thesis, i.e., using Google Trends data in infodemiology and infoveillance.

Google Trends [210] is popular in addressing health issues and topics. It is an open online tool that mainly shows what was and is trending, providing both real-time and archived information on Google queries from 2004 onwards. Google Trends data have been employed to analyze the users' behavior towards various health topics, seasonal diseases, and outbreaks, as well as forecast disease prevalence and epidemics. Google Trends main advantage is that it uses the revealed and not the stated users' preferences [8-9], so we can obtain information that would be difficult or impossible to retrieve otherwise. In addition, as data are available in real time, it solves issues that arise with traditional, time-consuming, survey methods.

2.2. Gaps in literature

The first step towards implementing this Thesis, was studying the state of the art of the field. As identified by reviewing the existing literature, Google Trends, despite its limitations, was suggested to be a promising tool for the monitoring of the users' search patterns and online behavior, with studies having explored methods of assessing seasonality, and also showing that Google query data correlate with official health data on several health topics. However, proceeding with the prediction/forecasting of diseases and outbreaks had not been assessed much.

Up to 2017 when this Ph.D. started, the only systematic review in the literature in this topic was that of Nuti et al. [211], which considered publications published until 2013 (in specific, January 3rd, 2014), and consisted of the systematic reporting of the existing publications and presenting the standard information for each selected article. As identified by Nuti et al. [211], the publications from 2009 to 2013 had increased seven-fold, indicating

the trend in Google Trends literature, and suggesting that the topic was promising in health informatics research. However, what the international literature lacked, was a review that, apart from recording the relevant publications and their basic information, also proceeded with an in-depth reporting of the approaches, methods, and tools used when employing Google Trends data in health research. The latter is the topic of the first publication of this Thesis, a systematic review entitled '*Assessing the Methods, Tools, and Statistical Approaches in Google Trends Research: Systematic Review*' [A] and published in the *Journal of Medical Internet Research.*

An important gap identified when reviewing the Google Trends literature was the lack of a uniform data collection and reporting methodology framework, which had resulted in differences and mistakes in the procedure of data selection and collection, like incorrect keyword selection, and appropriate region and period selection, which are crucial for a sound methodology basis. This gap was addressed in the second publication of this Thesis, entitled '*Google Trends in Infodemiology and Infoveillance: Methodology Framework*' [B]. This paper, published in *JMIR Public Health and Surveillance*, consists of a concise overview of how Google Trends data are retrieved and adjusted along with the available features, followed by a step-by-step framework for the key data collection methodology points (keyword(s), region(s), period, and category) when using Google Trends data in infodemiology, where a sound data collection methodology is necessary and crucial for ensuring the validity of the results.

Reviewing the literature also identified several gaps that existed in Google Trends research up to 2016. At first, the quantitative analysis of the selected articles in Google Trends showed that the forecasting of diseases and epidemics was still not assessed in many publications; in specific in only 9 out of 104 examined articles. Though official health data and online search traffic data correlate, the most important step towards health assessment using Google Trends is that of finding methods of predicting, forecasting, and nowcasting diseases' prevalence, outbreaks, and epidemics.

Furthermore, the Systematic Review paper [A] identified several ideas that could be further explored in order for the topic to be brought forward. For example, exploring the association of Google Trends time series with other Google Trends times series, how to use real time Google queries for monitoring outbreaks and epidemics, how to proceed with prediction/forecasting, and the limitations of Google Trends research based on data collection.

Towards these directions, six research papers have been published. Two of these papers considered data from EU (and UK) countries, while the rest (four) considered data from the US. The US, as identified in the Systematic review paper [A], is the most popular country in terms of publications in Google Trends health research. The reasons are threefold. First, USA has a very detailed regional breaking down of Google queries, meaning that the assessment by state is possible. Second, it is an English-speaking country. Third, the official governmental health organization, i.e., the Centers for Disease Control and Prevention (CDC), has a wide variety of freely accessible official health data.

In the first research paper, entitled '*Integrating Smart Health in the US Health Care System: Asthma Monitoring in the Google Era'* [C] and published in *JMIR Public Health and Surveillance*, an alternative method for monitoring asthma in the US using Google Trends data is proposed.

The measles outbreak in Europe in 2017, provided an opportunity for exploring how to monitor the users' online behavior towards the disease, and examine its relationship with the increase in measles cases, also exploring if the rise of the Anti-Vaccine Movement is associated with the decrease in immunization. It is entitled "*The internet and the Anti Vaccine Movement*" and is published in *Big Data and Cognitive Computing* [D].

The next two research papers explore the relationships between Google Trends data and the prevalence of the infectious diseases included in CDC's Atlas Plus: '*Forecasting AIDS Prevalence in the United States using Online Search Traffic Data*' [E], and '*Infoveillance of Infectious Diseases in USA: STDs, Tuberculosis, and Hepatitis*' [F], both published in the *Journal of Big Data.* These two publications, following the same methodology, identify both successful cases of monitoring, as well as cases for which this specific methodology approach is not successful, along with factors that can generally affect the validity of the methods or the results.

Finally, during the COVID-19 pandemic, it was essential to explore novel methods for early disease detection, taking advantage of the opportunity of having daily and openly available data. Two papers addressing the challenges of COVID-19 monitoring and predictability in Europe and the US, respectively, are included in this Thesis, namely '*Tracking COVID-19 in Europe: Infodemiology Approach*' that is published in *JMIR Public Health and Surveillance* [G], and '*COVID-19 predictability in the United States using Google Trends time series*' that is published in *Scientific Reports* [H].

2.3. Objectives

The overarching aim of the research is to explore how users' Google behavior towards health topics can be useful in public health epidemiology and surveillance, which can further build towards multidisciplinary approaches and policy making.

In specific, the main aim of this Thesis is to fill in the gaps in the state of the art, and explore new topics, approaches, and limitations in this line of research, as detailed above. The measurable objectives of this Thesis include:

1. Comprehensive state of the art and methods used up to that point -addressed in the Systematic Review paper [A]
2. Methodology framework for rigorous data collection and reporting -addressed in the Methodology Paper [B]
3. Exploring forecasting online behavior (as measured by Google queries) -addressed in the Asthma paper [C]
4. Calculating or modelling associations between online search traffic data and official health data -addressed in the rest of the research papers [D], [E], [F], [G], and [H]
5. Limitations and pitfalls in this line of research like lack of small interval health data, misspellings, sudden events -mainly addressed in [C], [D] and [F]

3. Main Findings

As mentioned in the *Aims and Motivation* section, this Thesis consists of eight interconnected publications, forming a structured presentation of the topic, i.e., a systematic literature review paper, a methodology paper, and six research papers. Towards implementing this Ph.D. and considering the gaps identified when reviewing the literature in early 2017, the overall findings of this Thesis indicate that Google Trends data are valuable in assisting with the monitoring and forecasting of the online behavioral changes towards health topics.

In specific, this Thesis presents an up-to-date systematic review of the topic, along with in-depth reporting of the approaches employed in this line of research up to the end of 2016, identifying the main methodology directions and research/statistical approaches, e.g., seasonality, correlations, modelling, and forecasting. Moreover, this Thesis presents a Google Trends in infodemiology data collection and reporting methodology framework, that filled in the gap of a solid data selection methodology basis that could affect the results and conclusions reached if data selection and collection is not performed appropriately. Gaps in methodology (e.g., forecasting) as well as topic-wise (e.g., asthma, measles) that were identified in the beginning of this Ph.D., are also addressed in this Thesis, while factors, like misspellings and sudden events, that could compromise the validity of the results, are also presented and discussed. Finally, during the COVID-19 pandemic, an opportunity to explore the potential of Google Trends time series in early disease detection and monitoring when daily data are available was presented, and COVID-19 monitoring and predictability approaches in the US and Europe were explored.

The papers of this Ph.D. Thesis are concisely presented in the following subsections, including a brief background, methodology, and main results and conclusions for each of them.

3.1. Assessing the Methods, Tools, and Statistical Approaches in Google Trends Research: Systematic Review [A]

In the Systematic Review paper [A] -which is an extension to the first Systematic Review in Google Trends by Nuti et al. in 2014 [211]-, a systematic reporting of the selected publications is presented, along with an in-depth reporting of the methods, statistical tools, and approaches used in Google Trends studies in health-related topics up to the end of 2016.

The difference in retrieved number of publications for the same years between the two reviews, lies in the different search strategy and inclusion/exclusion criteria (refer to Figure 1 in the Systematic Review paper [A] and to Figure 1 in Nuti et al., [211], for a detailed overview of the criteria).

Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for selecting studies, we searched for the term "Google Trends" in the Scopus and PubMed databases from 2006 to 2016, applying specific criteria for types of publications and topics. A total of 109 published papers were extracted, excluding duplicates and those that did not fall inside the topics of health and medicine or the selected article types. We then further categorized the published papers according to their methodological approach, namely, visualization, seasonality, correlations, forecasting, and modeling. A detailed description of the methods, tools, and analyses used in the selected publications is provided, categorized by approach (Tables 2-6 in [A]).

All the examined papers consisted of some form of time series analysis, and all but two included data visualization. A total of 23.1% (24/104) studies used Google Trends data for examining seasonality, while 39.4% (41/104) and 32.7% (34/104) of the studies used correlations and modeling, respectively. Only 8.7% (9/104) of the studies used Google Trends data for predictions and forecasting in health-related topics; therefore, it is evident that a gap exists in forecasting using Google Trends data.

In this paper, the gaps in Google Trends research in health and medicine were identified, indicating that, though data on reported cases on various health topics and the respective Google Trends data had been correlated in a large number of studies, only few had proceeded with forecasting using online search traffic data.

3.2 Google Trends in Infodemiology and Infoveillance: Methodology Framework [B]

An issue with Google Trends research that was identified when reviewing the literature on the topic, was that of the lack of a uniform data collection and reporting methodology framework, which had resulted in differences or inconsistencies amongst publications in the topic. The Methodology paper [B] provides data collection overview and data collection guidance in using Google Trends in infodemiology, which is crucial for ensuring that the results are not compromised.

Towards this direction, and since as the analysis of online data is based on empirical relationships, and thus, a solid methodological basis of any Google Trends study is crucial for ensuring the value and validity of the results, this paper aimed at presenting, what parameters/features should be considered as well as what steps that should be followed to ensure the validity of the data collection procedure. Observed inconsistencies mainly include incorrect keyword(s) selection, as was in the case of Google Flu Trends, while other factors, like region or period selection that do not match the official health data availability, can compromise the analysis given the normalization of the data.

This paper consists of a concise overview of how Google Trends data are retrieved and adjusted along with the available features, and a step-by-step framework for the key data collection methodology points, i.e., keyword(s), region(s), period, and category. First, we provide an overview of how the data are retrieved and adjusted along with the available features, followed by the methodology framework for choosing the appropriate parameters, along with examples that can also help the new generation of researchers on the topic.

3.3 Integrating Smart Health in the US Health Care System: Infodemiology Study of Asthma Monitoring in the Google Era [C]

In the Asthma paper [C], we aimed at forecasting the online behavior toward asthma and examined the correlations between queries and reported cases to explore the possibility of nowcasting asthma prevalence in the United States using online search traffic data.

Google Trends time series in term “asthma” in the US from 2004 to 2016 were analysed, in order to explore new ways of monitoring the public’s online interest in the topic. At first, the seasonality of the search queries was analyzed, with the results indicating that the queries exhibit similar seasonality for the years included in the examined period. Our analysis shows that search queries exhibit seasonality within each year and the relationships between each 2 years’ queries are statistically significant ($P<.05$).

Moreover, using Holt-Winters exponential smoothing to the retrieved Google Trends time series, five-year forecasting models in all US states for the online behavioral variations of the interest in the term “asthma”, as measured by Google Trends, were estimated, and validated against available Google query data from January 2016 to June 2017. Said forecasting models performed well, indicating that the monitoring of Google queries -providing real time regional information- could assist with increasing and complementing the responsiveness of the US health care system at metro or city level.

Online behavior toward asthma can be accurately predicted, and this method of forecasting Google queries can be potentially used by health care officials to nowcast asthma prevalence by city, state, or nationally, subject to future availability of daily, weekly, or monthly data on reported cases. This method could therefore be used for improved monitoring and assessment of the needs surrounding the current population of patients with asthma.

3.4 The Internet and the Anti-Vaccine Movement: Tracking the 2017 EU Measles Outbreak [D]

In addressing the issue of misinformation due to information overload available nowadays, and given the Measles outbreak in Europe in 2017, in the Measles paper [D] we examine the behavioral changes in terms related to Measles and the Anti-Vaccine Movement from 2004 to 2017, as well identify any associations between online search traffic data on said queries and the Measles cases and immunization percentages. This is especially important for Measles, since it requires the highest immunization percentage out of the vaccine preventable diseases, and since it has been suggesting that conspiracist ideation is related to the rejection of scientific propositions.

We retrieved normalized Google Trends data from 1 January 2004 to 31 August 2017 for the (then) 28 EU countries, for both the English and the respective translated terms (independent searches). For the Worldwide assessment, the keywords ‘Measles’, ‘Mumps’, ‘Rubella’, ‘MMR’, and ‘Anti Vaccine’ were used, as the term ‘Anti Vaccine’ exhibited the highest relative search volumes compared to other variations of the term/topic. Following, we brifly presented the 1st and 2nd MMR doses immunization percentages in Europe as well as for the EU28, and proceeded with examining the association (by calcualting the Pearson correlation coefficients) between online activity, vaccine population coverage and reported cases of Measles in each of the EU28 countries,

The results show that statistically significant positive correlations exist between monthly Measles cases and Google queries in the respective translated terms in most EU28 countries from January 2011 to August 2017. Furthermore, a strong negative correlation ($p < 0.01$) exists between the online interest in the term ‘Anti Vaccine’ and the Worldwide immunization percentages from 2004 to 2016, meaning that , as the online interest in the term ‘Anti Vaccine’ increases, the immunization percentages decrease, which could be indicative of the role that the Internet plays in the spreading of misinformation.

3.5 Forecasting AIDS Prevalence in the United States using Online Search Traffic Data [E]

In the AIDS paper [E], we aimed at exploring the association of online search traffic data on selected keywords and categories with AIDS Prevalence data from the CDC Atlas Plus from 2004 to 2015.

The results indicate that statistically significant correlations exist at both national and US state level between Google Trends data and '*AIDS Prevalence*', while said relationship, i.e., Google queries in the 'AIDS' search term and illness (topic) against CDC data on the variable 'AIDS Prevalence', seems to follow a logarithmic relationship over the examined period.

This paper also points to the direction of taking advantage of the anonymity that the Internet offers, in order to retrieve information and increase awareness and targeted actions (online assistance, awareness, information, etc.) in sensitive topics, like AIDS, HIV, and sexually transmitted infections (STIs), as also identified in the Infectious Diseases [F] paper, and could further assist with health assessment in the US and in other countries and regions with valid and available official health data.

3.6 Infoveillance of Infectious Diseases in USA: STDs, Tuberculosis, and Hepatitis [F]

In the Infectious Diseases paper [F] we aimed at exploring the possibility of using data from Google Trends to identify any associations of online search activity with the STIs included in the CDC AtlasPlus (excluding AIDS), i.e., in Chlamydia, Gonorrhea, Syphilis, Tuberculosis, and Hepatitis in the United States.The data retrieved for Hepatitis are from January 1st, 2004 to December 31st, 2015, while for the rest of the examined diseases, the examined time frame is from January 1st, 2004 to December 31st, 2016.

First, we provide an overview of the online interest variations on each of these diseases for the respective examined periods. Next, we visualize the geographical distribution of the online interest in each disease for all states for each individual year, followed by calculating the Pearson correlations between Google Trends data and the respective CDC data on each disease's cases. Finally, we estimate linear regressions for the examined diseases at both national and state level, in order to examine the possibility of forecasting said diseases using Google Trends data.

The correlations between Google Trends data and CDC data on Chlamydia cases are statistically significant at a national level and in most of the states, while the forecasting exhibits good performing results in many states. For Hepatitis, significant correlations are observed for several US States, while forecasting also exhibits promising results. On the contrary, several factors can affect the applicability of this forecasting method, as in the cases of Gonorrhea, Syphilis, and Tuberculosis, where the correlations are statistically significant in fewer states.

Our results indicate that Google Trends data can assist in the monitoring of the users' online interest and awareness on the subject, as depicted in the geographical distribution of the behavioral variations in the US for the five examined STIs, i.e., Chlamydia, Gonorrhea, Syphilis, Tuberculosis, and Hepatitis.

This study supports previous findings suggesting that the analysis of real-time online data is important in health assessment, as it tackles the long procedure of data collection and analysis in traditional survey methods, and provides us with information that could not be accessible otherwise. Another important finding with serious implications is that said method is neither trivial nor universal, as several factors can affect its applicability and the validity of the results. The main issues arising from this study, are, as discussed in the Methodology paper [B], the incorrect keyword selection (as, for example, in Gonorrhea due to a misspelling), as well as how a sudden and short-lasting event (like in Hepatitis) or low disease rates (like in Tuberculosis) can be factors that affect the generalization of the conclusions.

3.7 Tracking COVID-19 in Europe: Infodemiology Approach [G]

In the COVID Europe paper [G], a preliminary analysis of monitoring the online behavior towards the new coronavirus using Google Trends data in the five most affected Europe countries at the time (Italy, Spain, France, Germany, and the UK) was performed.

Data from Google Trends were retrieved, spanning from January to March 2020 on the Topic (Virus) of "Coronavirus" (which was selected instead of COIVD-19 due to the time of retrieval when the latter term was not widely used yet). First, the interest in the term worldwide and in the aforementioned countries was examined, while regional analysis is then performed for Italy and the Pearson correlation coefficients between COVID-19 cases and deaths and Google Trends time series are calculated. For the detailed European countries' correlation analysis, case and death data from March 2 to 17 were used.

The results indicate that there is a relationship between query and COVID-19 data, and we also identify that a time point exists where said relationship declines, In specific, a critical point, up to which regions not severely affected exhibit the strongest relationship between Google and COVID-19 data, was identified. This suggests that focus should shift towards these regions to make full use of what real time data assessment can offer. The latter is essential for increasing the preparedness and responsiveness of local health institutions, which is the most important aspect in handling the COVID-19 pandemic. Overall, supporting previous research on the topic, this preliminary approach indicates that Google Trends data can be employed for the detection of early outbreaks.

3.8 COVID-19 predictability in the United States using Google Trends time series [H]

Building on [G] in exploring methods of monitoring and detecting early outbreaks of the disease, in the COVID USA paper [H], we assess the predictability of COVID-19 in the US at state level using Google Trends time series.

For Google Trends data retrieval, we used “coronavirus (virus)”, again due to the term “COVID-19” not being so widespread at the time. As a preliminary investigation, Pearson and Kendall rank correlations between the ratio (COVID- 19 deaths)/(COVID-19 cases) and Google Trends data are examined to explore the relationship between Google Trends data and COVID-19 data on cases and deaths in the US, using data from a subset of periods from March 4th to April 15th for the individual state analyses, depending on data vaialblibily on cases (i.e.,timeframe for each state starting after March 4th was the date for whichthe first confirmed state case was identified. Next, COVID-19 predictability analysis is performed, with the employed model being a quantile regression that is bias corrected via bootstrap simulation.

The results indicate that there are statistically significant correlations between Google Trends and COVID-19 data, while the estimated models exhibit strong COVID-19 predictability. In line with previous work that has suggested that online real-time data are valuable in the monitoring and forecasting of epidemics and outbreaks, it is evident that such infodemiology approaches can assist public health policy makers in addressing the most crucial issues: flattening the curve, allocating health resources, and increasing the effectiveness and preparedness of their respective health care systems.

4. Discussion

4.1. Time series analysis, correlation, and forecasting

What the use of Internet sources has to offer, is real-time information. For health data retrieved through traditional methods, such as registries, surveys, etc., analysis and assessment can take quite some time to be performed. Thus, monitoring, prediction, and nowcasting using said methods is not trivial, and it requires up to several months or even years to be completed and made publicly available.

One of the main limitations in the Google Trends statistical analysis, is the lack of availability of respective official health data. Even when such data are available, they are not available in small intervals, meaning that there are cases where researchers need to perform the analysis with very few observations. Therefore, an outlier or leaving out a single point, for example, could make an explored relationship from significant to non-significant and vice versa. Though this is a critical limitation of these approaches, it can be addressed by obtaining small interval official health data, since Google Trends also provides small time interval data, and modeling the relationship would be more efficient and robust.

Nevertheless, Google Trends data have been successfully employed for the monitoring, detection, and prediction of conditions, diseases, and outbreaks, as identified in Tables 2-6 in the Systematic Review paper [A]. However, for forecasting, the literature at the point of writing was scarce (refer to the Systematic Review [A], Table 4), with only nine publications having been identified as proceeding with prediction/forecasting. Approaches mainly included the Autoregressive Integrated Moving Average (ARIMA), Holt-Winters, and cross correlations, while linear and non-linear regressions, that are also standard methods in forecasting [212], along with multivariate regressions and fit lines, were also identified in Google Trends statistical modeling at that point (refer to the Systematic Review [A], Table 5).

For exploring the relationship between variables, the Pearson correlation is the most widely used method in Google Trends research, which was also employed in the six research papers [C-H]. Popular methods in exploring correlations, are the Spearman correlation, auto-correlations, and cross-correlations (refer to the Systematic Review paper [A], Table 3). Even in cases where very limited data in terms of observations are available, correlations can assist in obtaining an understanding of the relationship. For example, in

the Measles paper [D], where the correlations of different sets of years (with same starting and consecutive ending points) were explored, the results were supportive of our hypothesis, in terms of showing that a negative relationship existed that was significant only after a number of years; the latter could be also possibly explained by an outlier or a single point's effect, but, since the relationship is constantly changing towards the same direction, we were able to extract some preliminary findings and conclusions.

In these approaches, when more than one variable is included, cofound parameters could be considered. For example, as mentioned in the limitations of the Asthma paper [C], this study has not accounted for state-by-state confounders that could influence search patterns, such as the socioeconomic status and demographics of different states that might be relevant to asthma prevalence, as this would exceed the scope of this paper.

However, more complicated analyses (for example, using multiple keywords) should consider the relationships between keyword selection, in order for incorrections, like the ones in the Google Flu Trends case, to be avoided. Moreover, it should be noted that both statistical coherence and interpretation of the modeled results could be increased, if the respective dependent variable, based on each individual case, is modeled using rate data (e.g., cases per 100,000 population) or also be normalized (given that Google Trends data are normalized from 0-100), which would also assist in not having very large coefficients. However, even in cases that do not proceed with modeling the relationship rather than correlating the data and normalization would not add to the interpretation of the statistical analysis, such data adjustment could provide a more understandable and easier to visualize presentation of the results.

This line of research is based on empirical relationships that are defined as relationships or correlations supported by experiments or observations, but not necessarily by theory. Thus validation, along with what is suggested in causal inference in epidemiology, may need a large number of studies that cover a wide variety [215]. Nevertheless, the effect in validating the results that small interval data have, has become evident during the COVID-19 pandemic, where daily data are freely and publicly available in real-time, and forecasting and predictability methods using Google Trends data have indicated their potential.

In time series forecasting, the two most widely used methods are exponential smoothing and ARIMA [212], with Holt-Winters exponential smoothing (as employed in the Asthma paper [C]) being regarded as having motivated some of the most successful forecasting methods. Exponential smoothing assigns weights to more recent observations,

in order to consider that data '*get old'*; thus, a more recent observation is assigned a higher weight than an older observation.

Due to the latter, Holt Winters can provide reliable forecasts for a wide variety of time series, which Hyndman and Athanasopoulos [212] regard as a "*great advantage and of major importance to applications in industry*". While exponential smoothing models are based on a description of trend and seasonality in the data, ARIMA models aim at describing the autocorrelations in the data and at addressing the possible stationarity issues that could arise with seasonal time series. Therefore, ARIMA and Holt-Winters should be regarded as complimentary approaches in forecasting [212].

4.2. Causality and causal inference

It is essential to understand that correlation does not imply causation, or, as Pearl [213] puts it, "*one cannot substantiate causal claims from associations alone, even at the population level—behind every causal conclusion there must lie some causal assumption that is not testable in observational studies*". It should be noted that, as the possibility of biased strategies exists [214], information retrieved from Web-based data should be very carefully analyzed and interpreted. In addition, causality does not necessarily imply that one variable causes the other, rather than that they just are related, possibly even through their individual relationships with another variable. Nevertheless, establishing causality from just one study is almost impossible; it is essential to understand that, in order to answer causal questions, the latter need to be asked in many studies covering a wide range of diversity and "*not within studies, but between them*" [215]. In Google Trends studies, which aim at establishing empirical relationships based on statistical findings as, e.g., in observational epidemiology studies, the exploring of the predictability of certain variables is possible. However, it is important to note that causal relationships are classified as such only when prediction under intervention employing said relationships is possible [216].

While in most cases causality can be argued by changing variables and studying their effect, this is not always applicable due to ethical, cost, or pragmatical reasons [217-218]. If only observational data are available, distinguishing cause from effect is a crucial issue. Inferring causality from such data is regarded as a very important topic [218] and is, in many cases, the only way to explore causal relationships [217]. This has been a matter of heated discussions in research, while several approaches have been made in order to understand and model how such relationships, i.e., making assumptions on data coming from observation rather than experiment [219], could be explored [213, 217-218, 220-221].

According to Hoyer [218], linear causal models are mostly used; however, does not imply that the relationships are indeed linear, but are employed because such modeling approaches are less complicated.

However, unlike exact/formally modelled/formulated sciences (like pure mathematics or theoretical physics), disciplines like social sciences or epidemiology that do not provide numerically precise results and where the latter can also significantly depend on human behavior and thus have unpredicted outcomes or high levels of uncertainty, it is crucial to adopt a very wide range of tools and approaches [222]. However, it is crucial to note that, though the appropriate statistical tools and tests are important for the rigor of analysis, interpretation of formal results should be performed with caution -as all formal modeling/methods have their limitations-, while, at the same time, making sure that interpretation is not treated as a "*black box*" [222].

Causal inference in epidemiology can be categorized as a case of scientific reasoning, that aims at establishing a cause-effect relationship with time order. As Rothman and Greenland argue about causal inference in epidemiology, it "*is better viewed as an exercise in measurement of an effect rather than as a criterion-guided process for deciding whether an effect is present or not*" [223]. Nevertheless, specific criteria establishing the data/evidence validity, do not exist, and measuring said effects are subject to measurement errors [223]. Approaches have explored causal inference by observational data based on their availability and that have not necessarily been collected in controlled manners, mainly considering continuous or discrete numerical data and not applicable in binary data [224].

Despite that scientific work, independent of nature (i.e., experimental or observational), is always incomplete and could be changed/modified or even contradicted in the future [225], exploring and building on our current knowledge is essential for the advancement of research. As Greenland [222] argues, all methodologies have limits or flaws; however, one methodology is not essential for every application, while even a flawed methodology could provide good performance in some cases [222].

In Google Trends studies, however, one of the main issues that rise and is different in each individual case, is identifying the time-order factor of causal inference mentioned above. Though in some cases the time-order (which comes first) for the observational data correlation is easy to determine (e.g., in elections/voting), this is not always the case with, for example, health/disease data.

Nevertheless, as mentioned above, even though monotonically increasing/decreasing trends will tend to exhibit a high correlation, it is important to note

that, even if no causal relationship between two variables can be identified, correlations are useful for forecasting [212]. In this context, results using Google Trends data are mainly based on empirical relationships, i.e., relationships or correlations that are supported by experiments or observations but not necessarily by theory.

To elaborate, observational studies, including ones that use data that already been collected -as in these cases of Google Trends-, mainly try to address the research question by using only what can be observed, i.e., without trying to change the parameters. This, by definition, results in difficulties -if at all possible- to show causal inference, which is even more true for online search traffic data, where the studied population is completely unknown -contrary to other observational studies that may be drawing their conclusions by other, known or more well-defined populations, e.g., from surveys.

Taking into account the limitations posed by the very nature of data and the study design, observational studies in general and from Google Trends in specific, deal with real-word problems, with the ultimate aim to inform public health and policy, either providing general assessments of the public interest and behavior or preliminary approaches identifying issues/topics which can be further studied in more controlled environments.

4.3. The Asthma paper [C]

In the Systematic Review [A], several approaches exploring seasonal diseases were identified. However, monitoring asthma had not been explored up to that point. In the US, CDC surveillance of asthma in order to gather more information on asthma prevalence, is based on the Behavioral Risk Factor Surveillance System (BRFSS) Prevalence Data (telephone survey) and on the National Health Interview Survey (NHIS) Prevalence Data (probability sample survey). In the Asthma paper [C], accounting for the (then) current asthma surveillance methods/tools and taking into account that the aforementioned methods of asthma monitoring are based on qualitative data and take too long to assess, an alternative way to those surveillance methods is proposed, suggesting that such approaches can assist officials with raising awareness, and that city-level analysis of online queries could have a positive effect in the health care system's preparedness at local level, provided the availability of small interval data (e.g., daily).

This paper's novelty and contribution, points to the direction that, throughout the years, the behavioral users' patterns are consistent and can be well forecasted. However, this was only the first step. Future approaches should model the relationship with official small interval data (e.g., admissions, hospitalizations), which, when modelled would, in

turn, provide the opportunity for Google queries to be monitored independently and possible nowcast asthma regionally/locally.

As mentioned above, the Holt-Winters exponential smoothing forecasts are weighted averages of past observations, with weights decaying exponentially as the observations get older, i.e., the older observations are assigned smaller weights than the recent observations. In the Asthma paper [C], we used the additive Holt-Winters exponential smoothing with trend and additive seasonal component, in order to explore the time series' seasonality as well as estimate five-year forecasts for the asthma Google queries at both national and state level. As shown in Figures 8-11, (the models' parameters and coefficients can be found in Appendix 1, Tables A3 and A4), the states forecasting models (depicted in the respective red lines) perform well.

Furthermore, additional analysis was performed in order to a) provide a visual representation of the seasonality of the online queries on asthma (see Figure 4 in the Asthma paper [C]), and b) to argue on the similarity and the significance of the correlation of the (individually retrieved weekly data of 52 observations each, one for each examined year) queries' time series amongst the examined years (Table 1). Alternatively, an autocorrelation function can be employed to explore the seasonality as well as to provide visual representations of the effect, which, in this paper, are present in the 52 figures (USA, 50 states, DC) that depict the respective initial time series along with the estimated forecasts.

Asthma is a seasonal disease, thus, to be able to proceed with any sort of prediction/forecasting (not in Google Trends queries but in general), we need short-interval data on asthma admissions, hospitalizations, or prescriptions. If, for example, daily data were available, lag analysis could be performed to model the response period, from Google search to admission and vice versa. This is important especially for asthma, given the limited data availability and the current methods for asthma surveillance methods. Currently, the literature consists of other approaches in the wider topic of exploring the online behavior towards aspects of asthma or its seasonality [226-228], while Souza Pinto et al. [229] performed a similar approach of exploring the Google Trends data predictability in asthma (along with four other respiratory diseases) for 54 countries, suggesting that this approach has found ground to be further explored.

Since asthma is a seasonal disease that is linked to particular seasons of the year and peaks in asthma exacerbations, as mentioned in the Asthma paper [C] and is also supported by previous work on the topic [230-233], various climatic contexts as well as other related keywords can be used in order to provide alternative approaches to monitoring the online

behavior towards asthma, as identified in this paper, along with not accounting for state-to-state confounders that could influence the online users' behavior in the topic. However, the aim of this paper was to explore if Google queries exhibit seasonality and if forecasting them is possible, for providing insight for future research, while it also suggests that this approach could be valuable in future research and policy makers, if small interval data are available.

Several other associations can be explored. For example, since prior work has suggested that racial characteristics have been identified as significant in asthma prevalence and syphilis as noted in the Asthma [C] and Infectious Diseases [F] papers, respectively, future infodemiology approaches could test this association. Nevertheless, and as Google Trends does not provide demographic information/profiles on the users' searches, this hypothesis can be further explored by either changing the parameters or by using other sources/tools, like, for example, Twitter or other social media platforms, which offer more qualitative data and demographics from the users' profiles (although only the users' stated data are provided).

However, it should be noted that correlations with small number of observations, as, for example, between NHIS & BRFFS data and Google Trends data in the Asthma paper [C] where further analysis could not be performed despite the statistical significance of the correlations, should be carefully interpreted. Apart from single points due to sudden events or outliers that could affect the correlation, the relationship's strength could be also affected if even one point is removed.

In this paper, since only yearly NHIS and BRFSS data are available and as asthma is a seasonal disease, in order to be able to model the relationship between asthma incidence or prevalence and then proceed with any sort of prediction/forecasting (not only in Google Trends, but in general), short-interval data on asthma are needed; regional/local daily data would be ideal. Such data are very hard to obtain and are not freely available in most cases. If, for example, daily data were available, lag analysis could be performed in order to model the response period, from Google search to admission or vice versa (as the time-order relationship could be hard to identify). The latter highlights the need for data to be open and available in short intervals and not only on an annual basis, in order to take full advantage of the very detailed breaking down of Google Trends data, both in time as well as regionally.

4.4. The Measles paper [D]

In the Measles paper [D], a preliminary investigation and identification of the topic is presented, also providing an overview of how Internet data can be used to address the issue of monitoring the Measles outbreaks as well as the rise of the Anti-Vaccine movement. However, this analysis cannot go to the root of the issue, as country analysis can only be performed by native (or, at least, fluent) speakers. For example, in Greek, the translated measles term with the accent is not used for online searching, thus a time series retrieved would be a time series of zeros with very few exceptions of single non-zero values/points. Not only would this not add to the analysis, but it could be more complicated and also affect its replicability. This paper is merely identifying the issue, presenting the relationship, and proposing an idea of how future country-specific analyses can further investigate the topic. This has been identified as a limitation of this paper; however, future approaches have provided both country- [234] and language-specific [235-236] analyses.

4.5. The AIDS paper [E]

In the AIDS paper [E], data on the variables 'AIDS Prevalence', 'AIDS Diagnoses', and 'AIDS Deaths' were retrieved from the CDC AtlasPlus. Said data were only available for 12 years, thus this paper provides a preliminary analysis of how data on the 'AIDS Prevalence' variable can be plotted against Google Trends data in order to explore their potential relationship.

In this paper, modeling only the relationships between the respective dependent and independent variables based on the significance of their respective calculated Pearson correlation coefficients is performed. As mentioned in the paper, the linear and polynomial relationships between the variables were at first explored, and indeed it was the case that some states exhibited higher R^2 for these relationships. Nevertheless, as in both the US as well as in most states the relationship was tending to logarithmic, the estimated models for all categories and all individual states were presented following a logarithmic relationship independent of which relationship (linear, polynomial, logarithmic) exhibited the highest R^2, in order to have a consistent presentation, suggesting that this relationship will be more evident moving forward in time.

However, and given, as in the Asthma paper [C], the limited availability of small interval data, it should be noted that these approaches are sensitive to correlation analyses, as a single point could change both the nature as well as the significance of the relationship, as discussed above. Nevertheless, as depicted in Figures 1 and 2 in the AIDS paper [E], the

online interest, which is based on monthly data (144 observations), show that the online interest in the topic is high and remains relatively high throughout the examined period (meaning that very low relative search volumes are not observed) and that the interest in the topic can be monitored using online search traffic data. Nevertheless, the spikes observed in the early years in November are due to the World AIDS Day on December 1st, and do not affect the results, as they are only very few observations within a large dataset. The visual representation of the time series, though, could be confusing as to the seasonality of the series, which afterwards becomes smoother.

Regional and small interval data availability, as indicated in the Asthma paper [C], is crucial in this line of research. During the COVID-19 pandemic, however, where daily (and regional) data were available, this issue could be overcome. Early research on the topic suggested that, though statistically significant correlations between Google Trends data and COVID-19 data exist, the relationship between the two tends to decrease in both strength and significance in regions that have been affected by COVID-19 moving forward in time, because the interest in the virus decreases. Thus, the critical point, after which the online interest starts declining, should be identified in each individual case to proceed with regional nowcasting. This effect, though not particularly evident in the Asthma [C] and AIDS [E] papers given the lack of availability for short intervals, is the case with several Google Trends cases, which can also be seen in the negative correlations or several models' negative coefficients signs.

4.6. The Infectious Diseases paper [F]

Another important issue discussed in this collection of papers, is including and present cases with non-successful results, in order to elaborate on the fact that some methodologies may not be as simple to design, and any results should be very carefully interpreted. In the Infectious Diseases paper [F], it is suggested that the applicability of this method is not that trivial or universal, and that several factors need to be taken into account when using online data in this line of research. In this analysis, examples of cases where simple approaches in disease forecasting and surveillance do not work, are highlighted. In fact, what is emphasized in this paper is how the suitability of this method along with the respective forecasting results can be affected by low rates or other factors, therefore being a good example of why the selection of keywords and the interpretation of the results when using online search traffic data are crucial for the robustness of the analysis.

Overcoming the keyword selection issue is detailed in the Methodology paper [B], in the Keyword Selection subsection of the Methodology Framework section, along with examples on how to use the "+" feature (Figure 4), translations (Figure 5), and quotes (Figure 6). In this paper, the aim was to show how, contrary to the Asthma paper [C] for example, simple approaches (i.e., single search term keywords) and inappropriate keyword selections (along with other identified factors, like sudden events) can affect the results.

In the Gonorrhea case, for example, selecting "Topic" (or "Disease", "Condition", "Illness", depending on the selected topic) instead of "Search term", would include the related to the topic Google queries, and thus overcome the issue of misspellings. In the AIDS paper [E], towards this direction, we proceeded with the analysis of both 'AIDS (Illness)' and 'AIDS (Search term)', in order to provide a more detailed assessment; however, these two analyses in this particular case did not significantly differ. Another feature that could be used in similar cases, like the one of Gonorrhea identified in the Infectious Diseases paper [F], the "+" feature could be used in order to choose an appropriate set of keywords that could provide successful results. However, in this case, the analysis would become unnecessarily complicated, and thus selecting the 'Topic' would be more sensible. Finally, what is also important though not always assessed in Google Trends research papers, is making a qualitative mention or analysis based on the top/rising related queries/topics, as well as discuss peaks and spikes that are attributed to incidents or events and that can affect the results.

Nevertheless, in the Infectious Diseases paper [F] we wanted to provide examples where the simple methodological approach we used in the AIDS paper [E] does not work, and to indicate how non-appropriate data collection can affect the results. In this paper [F], Tuberculosis was identified as a low-rates case and thus the simple approach of modeling employed method could provide some forecasting opportunities.

Previous approach of providing successful modeling for Tuberculosis can be found in Zhou et al. [237] (referenced in [A], [D], [E], and [F]. In addition, Kostova et al. [238] (referenced in [A]) explored how media can influence online health information in several infectious diseases, including Tuberculosis. Finally, a more recent approach in Tuberculosis forecasting using a combination of keywords, can be found in the paper entitled "*Forecasting tuberculosis using diabetes-related google trends data*" [239].

However, future approaches could, building on what is suggested above, adjust/normalize the data in order to compare the coefficients in such analyses, aiming at exploring if underlying causal relationships exist. Finally, it should be noted that each case

is different and there is no universal threshold to determine, for example, which rates can be considered high enough to proceed with the analysis, as in the case of Tuberculosis, especially when not regional/local analysis is performed. What threshold holds in one approach, may not hold in the other.

Given the above, along with cases of incorrect or not appropriate factors of the data collection procedure, like in Google Flu Trends (refer to the Limitations section) where overestimation of flu cases was performed [240] due to inappropriate keyword selection, it was essential to a) provide a methodology framework for Google Trends data collection (as provided in the Methodology paper [B]), and b) elaborate on case examples of how to identify that the search strategy may not be appropriate (as elaborated on in the Infectious Diseases paper [F]).

4.7. The COVID-19 papers [G, H]

As identified in this Thesis, one of the main issues is the availability of small interval official health data, that pose serious limitations to how the relationship between the two, i.e., Google Trends with actual health data, can be modeled. However, during the COVID-19 pandemic, were detailed regional data for daily cases and deaths were immediately and publicly available, modeling using infodemiology sources for the monitoring and forecasting of several aspects of the pandemic, has shown its potential.

Moreover, during the COVID-19 pandemic, and apart from predictability and forecasting approaches, what has also been evident is how social media platforms can provide more qualitative data that can shift the focus of monitoring and tracking aspects of the pandemic. Such approaches include sentiment analysis, educational purposes, and efforts to measure and raise public awareness, as detailed in the COVID-19 USA paper [H].

In order to explore the characteristics of the Google Trends studies in recent research and, in comparison with the previous approaches as identified in the Systematic Review paper [A], a concise systematic reporting of COVID-19 in this topic is presented. A search in PubMed for ((covid[Title/Abstract]) AND ('google trends'[Title/Abstract])) yields 120 results in 2020 alone. Figure 1 depicts the flow diagram for the procedure of selecting the publications.

Per the exclusion criteria for the article types, letters to the editor, correspondence, preprints, viewpoints, and editorials (26 in total) were not included in further analysis and categorization. Furthermore, 11 papers were not relevant to the topic in discussion, or the use of Google Trends was not significant (e.g., exploring financial aspects of the pandemic

or only employed Google Trends to identify related keywords). Removing the above publications based on article type and content, a total of 83 publications were included for further analysis. Note that research letters and brief reports were included in the analysis.

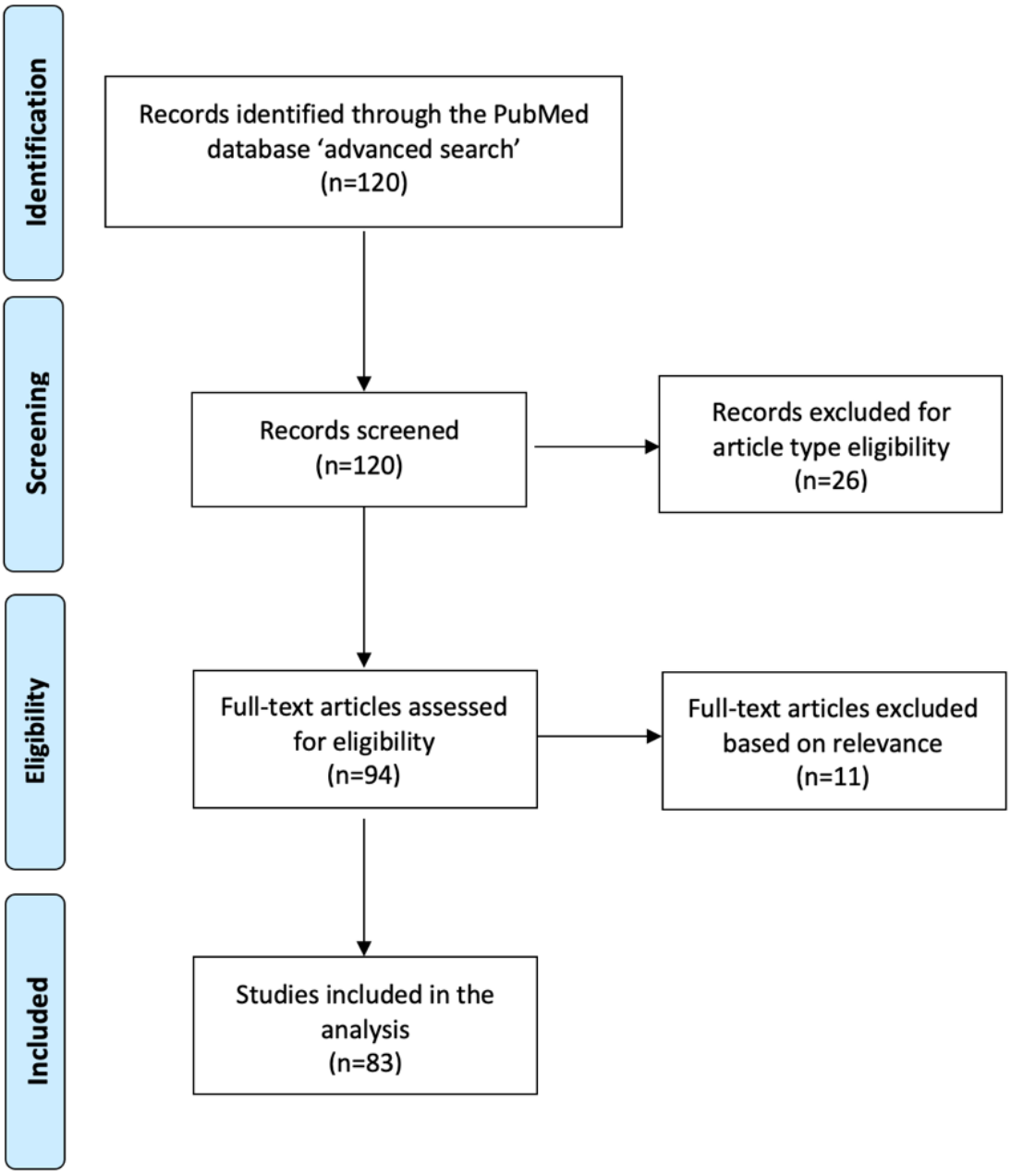


Figure 1. Flow diagram for retrieving the 2020 COVID-19 publications

A systematic reporting of the retrieved publications (in alphabetical order based on the first author) is included in Table 3. Analyses with more than 10 countries are marked as "multicountry analysis" and the respective detailed list of countries are included as notes under the table.

Table 3. Systematic reporting of the selected 2020 COVID-19 publications

First Author	Region	Period	Analysis
Adelhoefer [241]	USA	Aug 1, 2016-Aug 1, 2020	Pearson correlation; ARIMA; Explore the relationship between GT and COVID-19 data
Ahmad [242]	USA	Jan 20-April 20, 2020	Time-lagged cross correlations to gauge the association between GT and COVID-19 data
Arshad Ali [243]	Multicountry analysis[1]	Jan 21-July 21, 2020	Spearman's correlation coefficient (ρ); explore the relationship between GT and COVID-19 data
Asseo [244]	USA, Italy	March 4-Aug 25, 2020	Pearson correlations
Ayyoubzadeh [245]	Iran	Feb 10-March 18, 2020	Linear regression and long-short memory models to estimate COVID-19 cases
Azzam [246]	USA	Jan 1, 2004-20 July 2020	Unpaired t-test among GT data; Multivariable regression analysis to examine the relationship among GT data
Badell-Grau [247]	Australia, Germany, Italy, Spain, UK, USA	Nov 1, 2019-April 17, 2020	Spearman rho correlation to explore the relationship between GT data and COVID-19 data
Bento [248]	USA	Jan 1-March 18, 2020	Poisson models to examine the effect of the first COVID-19 case announcements on Google queries
Bettencourt-Silva [249]	Worldwide	Feb-June 2020	Explore trends in queries related to social determinants of health
Boserup [250]	USA, Puerto Rico, Guam	Sept 29, 2019-April 5, 2020	Assessment of public awareness in COVID-19
Brodeur [251]	USA, UK, Austria, Belgium, France, Ireland, Italy, Luxembourg, Portugal, Spain	Jan 1-April 10, 2020	Exploring the online interest in related to wellbeing terms, in order to examine the effects of the lockdowns
Burnett [252]	Multicountry analysis[2]	Feb 14, 2020-May 13, 2020	Correlation analysis; Mann Kendall test; among GT data on suicide and COVID-19
Cherry [253]	USA, Brazil, Spain, Italy, France	January-May 17, 2020	Spearman correlation between GT data and COVID-19 new daily cases per million
Choi [254]	USA, South Korea	June 2019-June 2020	Explored shift towards cultural and creative enrichment during COVID-19 lockdowns
Ciofani [255]	USA	Jan 1-May 24, 2020	Time-lag correlation analysis between the increase in COVID-19 cases and the utilization of Google for medical information
Cousins [256]	Multicountry analysis[3]	Jan 21-April 2, 2020	Univariate regression; Pearson correlation; root-mean-square error (RMSE); explore the predictability of regional COVID-19 case rates
Du [257]	USA, UK, Canada, Australia	Jan-March 24, 2020	Predictability analysis between prevalence rates of COVID-19 and GT data in fear-related emotions, seeking health-related knowledge etc.
Effenberger [258]	Multicountry analysis[4]	Dec 31, 2019-April 1, 2020	Pearson correlation with lag between GT data and COVID-19 data
Englund [259]	USA	Jan 1-Aug 1, 2020	Time series analysis to see how interest in hydroxychloroquine changes relative to media and news
Gazendam [260]	Worldwide	Feb 5-April 22, 2020	Relationship between public interest in hydroxychloroquine and chloroquine with the related active clinical trials of COVID-19 therapies
Ghosh [261]	India	March 10-May 23, 2020	Independent sample t-test, locally weighted scatterplot smoothing LOWESS, linear regression; relationship of GT data with policy measures
Greiner [262]	USA	March 1-April 5, 2020	Pearson correlations between GT and COVID-19 cases. Bivariate correlations between interest in preventive measures and SAH order delay
Gupta [263]	Worldwide	2004-2020	Epidemiological and hair restoration surgical data were combined with Google Trends data to analyze behavioral variations in trends
Halford [264]	USA	March 3, 2019-April 18, 2020	ARIMA for 18 related to suicide terms during the COVID-19 pandemic
Hamulka [265]	Worldwide, Poland	Jan 1-Oct 31, 2020	Spearman rank correlation between GT data and COVID-19 cumulative cases and deaths
Hartwell [266]	USA	May 1-May 31, 2020	Bivariate correlations between public interest in preventive measures (GT data), changes in confirmed COVID-19 cases after policy measures' expirations, COVID-19 case-fatality rates, and by-state presidential voting
Higgins [267]	Worldwide, Italy, Spain, USA	Jan 9-April 6, 2020	GT were compared to real-world confirmed cases and deaths of COVID-19
Hoerger [268]	USA	April 21, 2019-Aprill 19, 2020	Using Google Trends to track population-level mental health-related Google searches in the United States, Mann Whitney U test
Hong [269]	USA	Jan 21-March 18, 2020	Assessment of interest in telehealth and telemedicine; Pearson correlations of GT data with cumulative COVID-19 cases
Hou [270]	Multicountry analysis[5]	Dec 1, 2019-April 11, 2020	Assessment of real-time public awareness and behavioral responses to the COVID-19 epidemic using GT data
Hu [271]	USA, UK, Canada, Ireland, Australia	Dec 31-Feb 24, 2020	Dynamic series analysis to explore the change in trend of the online interest in COVID-19; Spearman correlations with COVID-19 infections

Husain [272]	USA	Dec 31-March 24, 2020	Monitor the public interest in COVID-19 using GT data
Husnayain [273]	South Korea	Dec 5, 2019-May 31, 2020	Spearman time-lag correlations between GT data and new COVID-19 cases; Single and multiple regressions; Prediction model with lags of new COVID-19 cases with GT as one of the variables
Husnayain [274]	Taiwan	Dec 5-Feb 8, 2020	Correlation of handwashing and mask related GT data with COVID-19 cases with lags
Jimenez [275]	Spain	Feb 20-May 20, 2020	Linear relationship between Google Trends data on COVID-19 cases with a lag; Pearson correlations; time lag correlations
Kardeş [276]	USA	March 15-August 29, 2020	To monitor the interest in 78 drugs and substances search terms related to COVID-19
Kardeş [277]	USA	Jan 1,2016-Sept 6, 2020	Generalized estimating equations with gamma distribution to examine the change in online interest in rheumatology related terms
Knipe [278]	Worldwide, Italy, USA, UK, Spain	Jan 1-March 30, 2020	Explore the online interest in COVID-19 related impact: mental distress, social and economic stressors, and mental health treatment-seeking
Kurian [279]	USA	Jan 22-April 6, 2020	Lag and lead Pearson correlations between Google Trends data
Kutlu [280]	Turkey, Italy	March 31, 2019-June 1, 2020	Pearson correlations of GT and COVID-19 cases
Landy [281]	USA	Dec 13, 2019-May 14, 2020	Explore public interest (using GT data) during the COVID-19 pandemic
Li [282]	China	Jan 2-Feb 12, 2020	Lag correlation between GT and confirmed/suspected COVID-19 cases
Lim [283]	Malaysia	Jan 22-March 26, 2020	Spearman's rank correlation with the number of new and total cases and total deaths
Lin [284]	Multicountry analysis[6]	Dec 20, 2019-April 19, 2020	Spearman rank-order correlations between GT data in insomnia, depression, and suicide related keywords, and COVID-19 cases and deaths and the number of days with increases in search volume for insomnia; forecasting future values with bootstrap CIs
Lippi [285]	Italy	Feb-May 2020	Spearman correlations between GT data and newly diagnosed COVID-19 cases with lags
Mavragani [286]	Italy, Spain, France, Germany, UK	January-March, 2020	Pearson correlations between Google Trends data and COVID-19 cases and deaths
Mavragani [287]	USA	March-April, 2020	Pearson correlation, Kendall rank correlation, between Google Trends and COVID 19 data, quantile regression predictability models
Mayasari [288]	Worldwide	Dec 31, 2019-April 15, 2020	Spearman rank between to measure the relationships between Google Trends and COVID 19 data
Muselli [289]	Worldwide, Italy	Dec 1, 2019-March 16, 2020	Spearman's rho correlation between GT data with Health Communication Strategies and official COVID-19 data
Niburski [290]	USA	March 1-April 30, 2020	Twitter, Google Trends, and amazon purchases in substances related to Trump's statements
Niu [291]	Italy	Jan 1-April 10, 2020	Prediction model using GT data on mask, pneumonia, thermometer, ISS, disinfection
Nsoesie [292]	Nigeria, Kenya, South Africa, UK, USA, India, Australia, Canada	Dec 2019-October 2020	Exponential growth model to characterize and compare the start, peak, and doubling time of COVID-19 misinformation topics
Paguio [293]	Worldwide	Dec 30, 2019-March 30, 2020	Pearson correlation between global online interest in COVID-19 and interest in CDC-recommended routine vaccines.
Pang [294]	Worldwide	Jan 1-April 30, 2020	Trend analysis of search terms for exploring the public interest in facial rejuvenation-related issues.
Panuganti [295]	USA	Jan 1-April 8, 2020	Spearman correlations of GT data (and Twitter data) related to smell symptoms, with COVID-19 incidence
Peng [296]	Multicountry analysis[7]	Jan 10-April 23, 2020	Proposed model with Google Trends data and Random Forest Classification for the prediction of the epidemic alert levels
Pier [297]	USA	March 29-May 16, 2020	t-test to explore the increase in the online interest in otolaryngology related terms
Prasanth [298]	India, USA, UK	Feb 24-May 20, 2020	Forecasting trends in daily and cumulative cases using long short term memory network optimized by GWO; compared results with ARIMA
Rajan [299]	USA	Oct 1, 2019-June 15, 2020	Correlations using moving averages of GT data on gastrointestinal symptoms and confirmed COVID-19 case count and daily new cases
Rokhmah [300]	Indonesia	Jan 1, 2019-June 6, 2020	Time lag correlations; Spearman rank correlations; between GT data on alternative and herbal medicine and practical activity with COVID-19 cases
Rovetta [301]	Worldwide	Feb 20-May 6, 2020	Explore behavior related to the COVID-19 pandemic (infodemic, conspiracies, etc.)
Rovetta [302]	Italy	January-March, 2020	Exploring interest in infodemic monikers
Rovetta [303]	Italy	June 11-Aug 2, 2020	Correlations between GT data and COVID-19 cases; Model using Support-Vector Machine (SVM) model, linear regression (LR), and Decision Tree (DT) to evaluate breakout prediction

Schnoell [304]	Australia, Brazil, Canada, Germany, Italy, South Africa, South Korea, Spain, UK, USA	Jan 1-June 19, 2020	Spearman rank correlation analysis between GT and COVID-19 data; Intraclass correlation for reliability analysis of the selected keywords
Senecal [305]	USA, UK, Spain, Italy	June 1-May 31, 2020	Correlation of GT data on chest pain symptoms and the reported decrease in ACS presenting
Singh [306]	India	February 1-April 27, 2020	GT data distribution check by Q–Q plots; one-way analysis of variance (or non-parametric Kruskal–Wallis test) to compare GT datasets
Sinyor [307]	Worldwide, USA	April 5, 2015-April 4, 2020	Generalized linear model to explore the change between GT datasets; interrupted time series regression using separate models for each keyword
Sousa-Pinto [308]	Multicountry analysis[8]	May 2015-May 2020	Compare GT data related to COVID-19; ARIMA to predict GT data on four diseases (asthma, COPD, diabetes, hypertensions, Chronhs disease)
Sousa-Pinto [309]	France, Germany, Italy, USA, Portugal, Spain, UK, Brazil,	2015-2020	Pearson correlations between COVID-19 data and GT data in several terms
Strzelecki [310]	Portugal, Poland	Jan 20-June 15, 2020	Visualization to measure the online interest; Pearson correlations between GT and COVID-19 spread
Subhash [311]	USA	May 2015-May 2020	Time series analysis to explore the relationship between sports related GT data and the (timing of the) COVID-19 pandemic
Sulyok [312]	Multicountry analysis[9]	Jan 23-Mar 13, 2020	Spearman rank cross-correlation analyses with lag between GT data and COVID-19 cases; Generalized additive models for COVID-19 cases
Sycinska-Dziarnowska [313]	USA, UK, Poland, Italy, Sweden	Jan 1-Aug 23, 2020	Descriptive comparison between GT data; Spearman correlations for GT keywords
Szmuda [314]	Multicountry analysis[10]	Dec 31, 2019-April 13, 2020	Spearman correlations of GT data with news and COVID-19 cases and deaths
Tijerina [315]	USA	Jan 1, 2015-May 21, 2020	Two-sample t-tests to compare GT datasets; Percentage changes in interest between GT datasets
Uvais [316]	India	March 12- June 13, 2020	Bivariate correlations between mental health GT data and COVID-19 cases
Venkatesh [317]	India	Jan 23-April 15, 2020	Spearman and lag correlations between COVID 19 cases and GT data
Walker [318]	France, Spain, Iran, UK, USA, Netherlands, Germany, Italy,	Dec 1, 2019-March 25, 2020	Spearman rank correlations for loss of smell queries with daily COVID-19 cases and deaths
Walker [319]	UK	Jan 31-April 12, 2020	Spearman's rank correlations; cross-correlation with lag; explore the relationship between GT data and COVID-19 cases
Xie [320]	China	Jan 10-Feb 29, 2020	Kendall's Tau-B rank test to check the bivariate correlation between epidemic trends and rumors
Younis [321]	USA	March 5-April 5, 2020	Pearson correlations; cross-correlations were performed between the time-varying reproduction number and social media activity or mobility
Yuan [322]	USA	March 1-April 15, 2020	Pearson correlations and general linear model to correlate and predict trends, respectively
Zattoni [323]	Worldwide	Jan 9-May 25, 2020	Joint point regression analysis; paired t-test to compare the change in interest in pornography

Notes

[1]South Africa, USA, Brazil, Peru, Chile, Mexico, Iran, Pakistan, Saudi Arabia, Russia, UK, Spain, Italy, Turkey, Germany, India, Bangladesh

[2]Australia, Austria, Belgium, Canada, France, Germany, Ireland, Italy, the Netherlands, New Zealand, Portugal, Russia, Spain, Sweden, UK, USA

[3]from 50 states and 166 county-based designated market areas

[4]China, Korea, Japan, Iran, Italy, Austria, Germany, UK, USA, Egypt, Australia, Brazil

[5]Japan, South Korea, Singapore, Italy, France, Spain, UK, USA Brazil, South Africa, India

[6]Iran, Spain, Italy, USA, France, Brazil, UK, Germany, Turkey, Canada, Russia, Japan, South Korea, Australia, Thailand, New Zealand, Singapore, Taiwan, Hong Kong

[7]202 countries

[8]Austria, Belgium, Bulgaria, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Netherlands, Norway, Poland, Portugal, Spain, Sweden, Switzerland, Ukraine, United Kingdom, Egypt, South Africa, Canada, USA, Argentina, Brazil, Chile, Colombia, Ecuador, Peru, Venezuela, India, Israel, Japan, Philippines, Turkey, Australia, New Zealand

[9]Belgium, France, Germany, Hungary, Ireland, Italy, Netherlands, Norway, Spain, Sweden, Switzerland, UK

[10]Albania, Andorra, Austria, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Holy See, Hungary, Iceland, Ireland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Moldova, Monaco, Montenegro, Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, Russia, San Marino, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Ukraine and the United Kingdom

The region popularity in terms of number of analyses is (>10), as expected, USA (50), Italy (27), UK (23), Spain (19), France (12), and Germany (12), and Australia (10), while Worldwide analysis (including those where the number of countries was very large,

i.e., >100) was also performed in several publications. Note that a country can appear in the analysis of more than one publication. These findings are consistent with the region popularity identified in the Systematic Review [A], which identified USA, UK, Australia, Canada, Germany, and Italy as the most popular countries for Google Trends analyses. Spain, the only differentiation evident in these lists, is added in the COVID-19 popularity most probably due to that it was one the first and most affected EU countries during the first wave of the pandemic.

Most of the selected papers, as depicted in Table 3, proceed with correlations, mainly between Google Trends data and COVID-19, or Google Trends time series with other Google Trends times series on selected queries. Spearman and Pearson correlations are the most popular approaches, while bivariate and cross-correlations have also been performed in several analyses.

Moreover, given the availability of daily data for COVID-19 cases and deaths, lag analysis was performed in numerous approaches. Furthermore, modeling, forecastings, and predictions have also been conducted in several cases, indicating the usefulness of Google Trends data when official health data are available in short time-intervals, with such approaches including linear models, linear and multivariate regressions, and ARIMA.

As evident by the above systematic reporting of the COVID-19 Google Trends studies, the analysis methods and approaches of these publications are consistent with the findings of the Systematic Review paper [A]. The two COVID papers included in this Thesis [G, H], had a significant contribution to the international literature on the topic, identifying -and confirming- not only the relationship and predictability potential that has been suggested to exist between Google Trends and health data, but also indicating and elaborating on the behavioral "turning point" in the online interest, which also provides a behavioral explanation on negative correlations or coefficients that have been observed in modeling Google Trends against diseases in general.

4.8. On the negative relationship between health data and Google Trends

In Google Trends time series analysis, the calculated correlations are merely indicative of the relationship of Google Trends data with health data, as, for example, in the Asthma paper [C]. The online interest is not always a monotonic function, nor does a Google Trend time series need to exactly follow the time series of a disease's cases/prevalence/deaths. The point in most of these approaches is not to identify which comes first, i.e., search or case (which, given the nature of empirical relationships, the time-order relationship may be

hard or even impossible to identify), rather than correlate data and then try to model the relationship.

Though counterintuitive in several cases (e.g., with negative correlations), Google searches and diseases/outbreaks do not always have a positive correlation, nor is this correlation constant throughout the weeks/months/years examined. What is suggested in the Asthma paper [C] is that Google queries can be well forecasted by state, but we cannot proceed with any further correlation or modelling, as actual small interval state, metro, or city level asthma data are not available. The correlation is merely calculated in order to see if a relationship exists. The next step is to proceed with modelling, as identified in the four steps of Google Trends research in the Systematic Review paper [A]. Once a relationship has been established, it is easier to explore any further statistical analysis and modelling, as suggested by Hyndman & Athanasopoulos [212].

The latter does not mean that a statistically insignificant correlation indicates that no modeling can be performed (as in the case of Tuberculosis in the Infectious Diseases paper [F] where other approaches have successfully proceeded with the monitoring of this disease), nor that all significant correlations can be modeled. Any further statistical analysis is case specific, and there is no universal way to determine how to proceed, before a preliminary investigation is performed (either with correlations or with basic time series analysis).

From a behavioral point of view, and as identified in the COVID papers [G, H], this can be explained as follows. First, online interest starts to increase and reaches a peak as the number of cases/deaths become high. However, after a certain period, the interest has an inverse course, which could also indicate that the public is overwhelmed by information overload and decreases its information "intake".

The latter is a dynamic relationship that can change back and forth, with spikes in Google queries simultaneously observed with a decline in diseases cases, incidence, prevalence, and vice versa. When daily (or weekly) official health data are openly and immediately available, like in the case of COVID-19, focus shifts towards regional analysis in order to make full use of what real time data assessment can offer, and identify early outbreaks or increase the preparedness and responsiveness of local health institutions.

The decrease in interest is counterintuitive, and can also be observed with a lag before (or after) the case/deaths/incidence time series start exhibiting a downward trend. This can also partly explain the differences in signs among states, but a more in-depth explanation is that the outliers could affect the estimation of the results, especially in the

case of a small sample. In general, linear regression, Pearson correlation, or the Kendall rank correlation, are not as resistant to outliers, as, for example, quantile regressions.

Another example of the above is the negative relationships between AIDS prevalence and retrieved Googe Trends data in the AIDS paper [E]. Since prevalence measures the part of the population that is already living with AIDS, the negative relationship could be attributed to the decreasing interest in online searches as time progresses, as discussed in the COVID-19 Europe paper [G] in which said effect exhibited quicker due to the nature of such a pandemic. Other parameters affecting the sign of the relationship, as shown in the Asthma paper [C] could be the availability of yearly data, which does not only offer a smaller number of observations, but also significantly limits the options for further analysis using Google data since asthma is a seasonal disease and any possible variations would more probably exhibit amongst months than amongst years.

Taking into account that exploring epidemics or even seasonal diseases that may be affected by other (external) factors, the relationship between Google Trends data and official incidence/prevalence data should be regarded as a dynamic process that constantly evolves, and it is probable that there exist several data anomalies (e.g., due to campaigns, raising awareness, sudden events, or other external factors); therefore, formal statistical tools such as the Pearson and Kendall rank correlations should be carefully interpreted.

4.9. Multidisciplinarity in epidemiology and surveillance

As Salathe argued back in 2018 and as is now widely known, traditional epidemiology's and digital epidemiology's aims align [324], which essentially are to use available information for disease prevention and to inform public health and policy. The difference is that in digital epidemiology the focus is on the "*why*" data are generated rather than the data's format, defining the topic as "*epidemiology that uses data that was generated outside the public health system, i.e., with data that was not generated with the primary purpose of doing epidemiology*" [324], urging us to examine the processes from a more wide and overall perspective as to what all this increasingly available information has to offer to epidemiology and surveillance.

The main aim of digital epidemiology would be for its routine implementation from the public health authorities' side [324]. Digital epidemiology in general and Big Data sources in specific, despite having provided successful approaches and insights over the past years during emergency conditions, and given that we are now in an excessive and

continuous data generating era, analysis and interpretation should be done with caution in order to avoid "*ill-informed decision making*" [325].

Therefore, public health should not be viewed solely as a medical field of research, rather than a combination of a wide variety of topics and expertise in order to obtain a broader understanding of what drives successful epidemiology and surveillance approaches. Multidisciplinary approaches mainly aim at addressing real word problems by providing different perspectives and achieving consensus amongst health/medical and other disciplines [326].

Though more focus on quantitative methods has been given in epidemiology and surveillance, Smith et al. argue that different methods and study designs are not "*superior in obtaining evidence*" to one another, that public health challenges should be approached by a multidisciplinary point of view, and that different methods and expertise should be viewed as complimentary to each other [327].

The role of political, social, and legal variables being incorporated into epidemiology and surveillance is essenital in order to achieve successful implementation of public health and surveillance research. It is important to understand, as Rayner notes, that public health should be interdisciplinary with "*strong understanding of political process*" [328]. Moreover, as argued by Kivits et al., in order to address health inequalities, approaches need to also consider "*social, territorial, economic and political perspectives*" [329], as also supported by Eckmanns et al. that further note that "*key ethical, legal, political and societal concerns must not be sidelined in contemporary efforts to accrue maximal data reserves"* [325]. What is also imperative to note is that, as evident during the COVID-19 pandemic, multidisciplinary approaches involving medicine, healthcare, economics, diplomacy and policy are critical [330], while, supporting the above, Kivits et al. note that other non traditional medical/health fields of research, like social, political, and legal sciences are also needed in order to understand and interpret health related research [329].

Can the same epidemiology model work in all regions if social and political variables are not considered? The short answer is no. There are overpopulated parts in the world where social distance is at least challenging, there are developing countries with serious economic inequalities, there are regions where people do not have the luxury of self-isolating, citizens in less privileged regions who must go out to work every day. There are Low- and Middle-Income Countries in South America or Asia, where the models that work for Europe or North America would not.

Interesting such examples could come from the COVID-19 pandemic. Even for countries that are amongst the more privileged ones, cultural differences can play a significant role in how epidemiology models should be built, as, e.g., Scandinavia vs. southern Europe, which would need different approaches for epidemiology models to be successfully implemented given the diversity in, e.g., social distancing. Modeling the spread of the disease by examining how far the aerosol can travel in controlled environments and how many meters of distance in between us we need to minimize the spread of the virus, may work in some regions; however, not incorporating social, societal, and cultural differences in said models may deem them unsuccessful in other regions.

In sense, when building epidemiology models, we cannot leave out the human factor and proceed with solely health/medical variables. During COVID-19 is when we understood in full scale why multidisciplinarity is what public health needs in chaotic times like in a pandemic, when other than strictly health and medical variables needed to be incorporated into epidemiology and surveillance model in order to address the situation in a quick and effective way. Another very important aspect, as noted above, is factoring in the macroeconomic variable in public health and surveillance. What happens in less privileged settings than what we have experienced in western countries, affects us all. Global health is essentially referring to what is happening now rather than a location [330]. Or, as Hinchman et al. interestingly put it, "*Global health is local health*" [330].

As previously outlined as the main goal of this Thesis, multidisciplinary approaches are indeed what we need in order to make better sense of how online behavior and official health data are related, and what multidisciplinary analyses have to offer to public health and surveillance. However, as noted above, these approaches are to be considered as supplemental in epidemiology and surveillance, highlighting, again, the multidisciplinary nature of public health, while interpretation of the results should be done with caution.

As a takeaway message, for more effective policy making and thus successful epidemiology and surveillance approaches to be implemented, all different perspectives, approaches, and -in sense- available information should be considered when making informed decisions. Public health officials should incorporate social, political, and technological on top of strictly medical aspects when building policies for increased preparedness and successful implementation of epidemiology and surveillance in the future.

5. Limitations

5.1. General limitations

Despite the advantages that Internet data sources have to offer, several limitations have been identified in this Thesis concerning the Google Trends research like the necessary openness and availability of official health data in order to take full advantage of said data. Even in countries with strong open data policies, retrieving and analyzing data on diseases' prevalence is time consuming and involving several health officials before they are fully accessible. This means that data are not available in real time and can even take more several months or years for the final official analysis to be publicly available. Even when data have become available, they do not usually consist of short time interval data, e.g., daily, thus the analysis and forecasting of diseases and conditions becomes even trickier.

Moreover, the sample cannot be shown to be representative. In this sense, the use of a search engine is almost universal, i.e., employed by almost all Internet users, and the representation is becoming less of an issue as Internet penetration and age of users increase, as we are moving forward in time.

In addition, even there are more than one (worldwide) search engines available, and thus not all queries (data) on the respective selected topic can be retrieved. Apart from this, several countries have their own search engines, like Yandex (Russia), Daum (South Korea), and Baidu (China). Thus, the analysis of online search traffic data from these countries is not trivial.

Another limitation of using Web-based sources is that data can be affected by sudden incidents or events, which, especially in nowcasting or when the number of observations is low, could provide biased results. The latter needs special caution during the analysis process and geographical representation of the data, as, for example, identifying any outliers, as in the case of Hepatitis [F], which was due to a food poisoning incident in Hawaii.

In addition, as per the correlation investigation, it should be noted that careful analysis is required in order to ensure that association is not inferred by correlation, as it is likely that two monotonic functions could exhibit statistically significant correlation but without association. Nevertheless, as mentioned in the Discussion section, correlations are useful in forecasting even if no causal relationship between two variables exists [212], which is crucial in analyses like in Google Trends research, that are based on empirical relationships.

Furthermore, the forecasting results and interpretation may likely benefit from using a ratio-based scale of official population data rather than count data. All the above call out for multidisciplinarity for this part of health informatics to move forward, and to also minimize any incompleteness that may arise from single-focus data analyses.

5.2. Google Trends normalization

A factor that should be considered, is the normalization of Google Trends data over the selected time frame, which is rather a limitation as to the analysis approach. Though a descriptive data overview can be found in the Methodology paper [B] of this Thesis, i.e., on filtering, eliminations, etc., the entire algorithm for the data adjustment is not reported, as Google Trends does not officially disclose more detailed information on the procedure.

The "Trends Help" section of the platform, describes the procedure as follows [331]: *"Google Trends normalizes search data to make comparisons between terms easier. Search results are normalized to the time and location of a query by the following process: Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity. Otherwise, places with the most search volume would always be ranked highest. The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics. Different regions that show the same search interest for a term don't always have the same total search volumes."*.

The normalization of the data can be explained as follows. Exploring Google queries in a term/topic is, in sense, the interest in said term/topic in terms of its proportion of all Google queries in all terms/topics for the selected region and period. Likewise, the Google queries in a term/topic in a specific region, is the interest in said term/topic in terms of its proportion of all Google queries in all terms/topics for the same region and period [332].

The latter is also evident in the Hepatitis case in the Infectious Diseases paper [F], where the 2016 low relative interest in all other states in Hepatitis apart from the state of Hawaii, is probably due to this Hawaii raw scallops incident in August 2016. This example is important to have a better understanding of the regional aspect of the data adjustment, as it indicates that the procedure is indeed not affected by population (Hawaii ranks 41st in the US in terms of population) rather than relativity of queries, indicating that the Google queries sampling can be significantly affected by increased queries in a topic in even low population regions. The latter's effect as represented in Figure 10 (2016) in the Infectious Diseases paper [F], also makes evident why heat maps, like the ones used in this paper, apart from providing visualizations that make the paper reader friendly, they also provide a visual and

easy way of understanding -and even to a wider audience- how a sudden event or incident can affect the results.

As the Google queries time series retrieved on a term/topic are normalized over the selected period and region taking values between 0 and 100, for any further correlation or statistical analysis and modeling, the region and period selection should exactly match that for which health data are available/used. Respectively, when correlating (or modeling) Google queries on different terms or topics or regions, the selected time frame should, again, be exactly the same. However, though normalization can be helpful in cases like the one mentioned above, it limits the range of selection of analyses, as it does not provide exact volumes rather than normalized values. In sense, the normalization, though making the analysis approach trickier, provides the researcher with the opportunity of looking into the respective examined topic with a different scope, meaning that the comparison of regions and periods is more trivial.

5.3. Anonymity and revealed data

Since Google Trends data provide us with the revealed and not the stated users' preferences, patterns, and behavior, they have an advantage over traditional data collection methods in terms of anonymity, as identified in the AIDS paper [E]. For example, it is more probable to search for AIDS information or testing online before being diagnosed, than to seek information from the GP.

Moreover, this has a significant effect in mental health issues, where several such topics are still considered taboo and where individuals that suffer from mental illness may not seek help from a professional, rather than online self-help. As identified in the Systematic Review [A] as well as in Table 1 of this Context Chapter consisting of indicative topics in infodemiology and infoveillance, Internet data have been valuable in exploring or predicting suicidal behaviors or pointing to the direction of increasing awareness of "hot" topics in the subject.

Since Internet penetration is increasing and since querying is universal, it is understandable that online information is the first step/action people take towards this direction (contrary to filling in a questionnaire, for example). Even in popular social media like Twitter, Facebook, or Reddit, for example, though many users exist and though there are ways to achieve anonymous interaction (liking, commenting, following, etc.), the latter is made in lower volumes that querying.

Nevertheless, there are some cases where the anonymity that Google querying provides will not hold, as, for example, in cases of drug trafficking where law enforcement officials can have access to user logs. However, the latter is not the case with soft drugs, for individual users (not dealers), or prescription drugs, where general interest in aspects of drugs (legal or illegal) can be explored and monitored (refer to Table 1 of this Chapter for topics including drug abuse, cannabis, illicit drugs, pharmaceutical/prescription drugs). Nevertheless, governmental/official access to personal data is not the same in every country's legislation, and this is one of the reasons why in countries with severe governmental internet monitoring, restrictions, censorship, or where internet penetration is low, Google Trends data may not provide accurate results [8].

5.4. Additional features and techniques

Another issue for discussion that arises and that should be considered in future research, is how new approaches, like Natural Language Processing (NLP) or Autocomplete, could have an impact on Google querying. Autocomplete is, in some cases, assistive to Google querying, as it indicates how the keyword/sentence can be completed based on real, common, and trending queries, making it less possible to make spellings mistakes for example, while it could also group together similar searches, if its prediction is correct.

However, in cases like Gonorrhea, where the misspelled term is more often used than the correct one, Autocomplete could have an impact on the Google Trends querying, since it bases its predictions in real previous searches made by the users, also taking into account the location and continuing towards the direction of selecting the incorrect spelling, for example, is possible.

Nevertheless, previous research in flu prediction using Google Trends, indicates that basic NLP techniques (including word removal and stemming) does not improve the models' performance, highlighting "*the quality of the signal enclosed in the data*", also arguing that text processing has a negative effect on information [334]. Nevertheless, all such tools and sources, like Google Trends, are dynamic, meaning that they change constantly, providing with us with additional features, and research up to this point has adapted to such changes, and, most probably, this will be the case in the future as well.

5.5. Google Flu Trends and Controversies

A serious case of controversy due to the overestimation of flu cases, as mentioned in the Methodology [B] and AIDS [E] papers, was the case of Google Flu Trends [240], a tool

provided by Google (2008-2015), similar to Google Trends, that only focused on flu related Google query data to estimate influenza like illness incidence. The issue with said overestimation was the data collection methodology and the keyword selection in specific, which was (and still is) the case with many approaches using Google Trends data. However, this was short afterwards addressed by Lampros et al. [334], who qualitatively explained the shortcomings of the previous approach, identifying the incorrect choice in keywords and providing a more successful approach to forecasting flu cases.

The issue of incorrect or incomplete data methodology that can affect or even compromise the results, as with the Google Flu Trends, along with several such cases being identified when studying the current state of the art, is what showed that the literature was lacking a framework in order to ensure that the basis of any Google Trends approach, i.e., the data collection, is performed correctly, and also that it is reported in detail and uniformly so that all studies can be replicated.

Though Google Flu Trends is not available now and there have been some controversies on the subject several years ago, research on the topic has since then progressed. Moreover, recent research has indicated that use of a combination of variables (electronic health records and Google Trends time series) can successfully predict flu incidence [333], while previous work has also included, apart from health and Google Trends data, also confounding parameters, like weather data [335], that could affect the development of the disease. Nevertheless, in epidemics, where historical data are not available, such analyses could be assisted by assigning weights on observations [335], as also suggested by recent work on the topic comparing forecasting methods (including ARIMA, Holt Winters, and multiple linear regressions), that indicate that smoothing (Holt Winters) performs well for forecasting aspects related to the COVID-19 pandemic.

Over the years, the use of Internet sources in general and Google Trends in specific, especially after the lack of success of the Google Flu Trends project, has been the case of controversial discussions. However, this is quite common for new/emerging research topics, especially in cases like Google Trends, where discussion about its potential success at its first stages exceeded what the first results had to exhibit in terms of immediate success and was therefore quickly negatively judged. For example, in syndromic surveillance, that uses a combination of clinical and non-clinical (alternative) data for early disease detection and spread minimization, it has been indicated that such alternative sources, including Internet sources (and Google Trends data), can present crucial issues for the analysis, influenced by factors that cannot be predicted or quantified, and could potentially be

irrelevant to the topic in interest [337]. However, it is suggested that optimal approaches may be the ones incorporating data from a variety of sources [337], which is also supportive of several approaches so far in the topic of infectious diseases syndromic surveillance methods using Internet data [338].

Research on the topic is still relatively young and constantly evolving, and, despite shortcomings, like the case in every new research area, Big Data in health informatics have come a long way over the past decade and have indicated very promising results. Nevertheless, and as mentioned in the Asthma [C] and Infectious Diseases [F] papers, multidisciplinary approaches combining traditional data with Big Data [240, 326, 335, 339] can further explore the potential that real time assessment can provide. In sense, future research in forecasting infectious diseases should incorporate different methods, data sources, and structures [335], to take full advantage of what such combinations can offer. As noted by Vandenbroucke et al. [215], in public health decision making, the value of causal inference theory is evident, as it is essential that all available evidence is considered; however also indicating the need for "*pluralistic views of causality*" and its assessment [215].

6. Thesis Contribution

This Thesis has suggested that the monitoring of search traffic data from Google can help improve our understanding and analysis of online behavior and behavioral changes and patterns towards diseases and health topics. The collection of the Thesis' publications closely follows the IMRD structure, i.e., consisting of an introductory paper dealing with systematically reviewing the relative literature, methods, and approaches; one methodology paper outlining the available features and proposing the appropriate procedure for data collection; and six research papers in using Google Trends data in various contexts and diseases/conditions, while the Context Chapter summarizes and analyzes several aspects, collective findings, and limitations that could not be part of the papers. Based on the above, the synthesis of this Chapter along with the publications, forms a coherent and structured approach of the examined topic.

The collective contribution of this thesis to the international literature can be summarized as follows. The Review [A] and Methodology [B] papers have a significant novel contribution to the progressing of infodemiology and infoveillance research using Google Trends data. The former provided an updated systematic reporting of Google Trends publications, along with a reporting and analysis of the employed tools and methods, and a quantitative identification of the gaps that acts as a guide for current and future research. The Methodology [B] paper was the first guide for the appropriate building and reporting of a sound Google Trends data collection methodology, so that results are not compromised, and has provided useful guidance for other researchers to follow. The research papers [C], [D], and [E], using methods or combinations of approaches identified in the Systematic Review paper [A] and in novel topics, have contributed to suggesting how to monitor the users' online behavior and exploring if said data can be further employed for the prediction and nowcasting of seasonal conditions/diseases and outbreaks/epidemics, while [F] also provided specific cases and examples as to how to detect factors that affect the applicability using Google query data in health assessment. The two COVID-19 papers and with daily data being available, suggest novel approaches of monitoring epidemics, that further elaborate on the usefulness of Google Trends data in disease surveillance.

Said publications have received over 1,300 citations in Google Scholar as of April 2025. A detailed presentation of the contribution of the publications of this Thesis to the international literature in the topic in terms of citations, is presented in the Appendix.

7. Concluding Remarks - Future Research

Using Internet sources in information epidemiology and disease surveillance is indicated to be of value over the past decade, and the results of this Thesis also point to this direction. Data sources cover a wide variety of tools, social media, platforms, websites, blogs, and search engines, while the studied topics vary from chronic or seasonal disease prevalence to nowcasting epidemics and outbreaks. Despite the limitations that arise when using online search traffic data, Google query data present benefits that can tackle several of the issues that arise with time-consuming traditional assessment methods and examine the users' online behavioral patterns. The analysis of online search traffic data is universal, in the sense that Internet penetration has increased to a point where the majority of people have access to and use the world wide web, and querying is one of the most frequently used Internet features, and thus health informatics research benefits from the use of said sources.

Empirical relationships have been shown to exist between Internet data and official data, with infodemiology using the vast amount of information available online for the assessment of public health and policy matters. Internet sources also have the benefit of the assessment of larger populations, in contrast to most traditional methods that are based on data retrieved from significantly smaller groups, as, for example, surveys.

In order to take full advantage of the benefits of Google Trends data, a wide variety of approaches/analyses and studies are required. For example, exploring seasonal diseases or chronic conditions to increase awareness and the responsiveness of the health system, would benefit from the assessment per city or metropolitan area, as discussed in [C]. However, this can only be possible subject to official data intervals and availability, which are also very important for nowcasting outbreaks and epidemics. For the latter, as evident during the COVID-19 pandemic, Internet sources can actively participate in the procedure of tracking the spread of the disease and have been suggested to be valuable in also monitoring other aspects of the pandemic.

Furthermore, there is a significant rise in publications from other (than Twitter) social media such as Facebook and Instagram. This could be showing the younger Internet users' preferences in the use of social media and could be revealing a trend of the use of said platforms. Researchers in this field should closely follow any potential shifts in Internet usage along with the correspondence with age of users, to ensure -to the point that this is

possible with Web-based data- that the sample is representative, and that research aims change along with what is trending.

As research on the subject is expanding very quickly and new researchers are getting involved in research in the topic, what is necessary is to focus on reviews (both systematic and scoping), to act as guides to both the experienced as well as new researchers. As the search by source yields many results, focus should be given to future reviews that focus not only in infodemiology and infoveillance in general, but also on source employed (e.g., systematic reviews on Facebook research), topic explored (e.g., cancer in infodemiology), and even more specific combination, i.e., source and topic/disease (e.g., mobile apps for asthma).

What is also important, is the analysis of the methods that have been employed up till now. Though this has been addressed in review paper [A] of this Thesis, the period covered was until the end of 2016, and, since the large corpus of literature is after this point, a more in-depth categorization is needed, to further assess the successfulness and diversity of the new approaches over the years. Finally, though several papers with combination of sources exist, what the international literature lacks are studies aiming at combining infodemiology data, methods, and approaches, as identified in the Systematic Review paper [A]. This, however, this would require a higher number of researchers and resources. Apart from the required significant work towards the direction of exploring novel aspects and approaches of monitoring Internet data in digital epidemiology, it is essential that a unified methodology checklist and reporting for all Google Trends studies to follow is developed, to minimize incorrections and inconsistencies.

Present results could have significant implications for effective policy making, while multidisciplinary analyses are crucial to understand the nature of the relationship between Internet and health data. Considering that policy makers are required, as in the case of the COVID-19 pandemic, to make important decisions during periods with chaotic conditions, it is vital to progress with a statistical understanding of Google Trends time series and the users' behavior in accordance with its real determinants. In a sense, such approaches should not be strictly medical to estimate robust prediction models, rather than a combination of medical and non-medical parameters from several research fields, that also take into account the public's awareness and online behavior towards the explored topics.

Chapter 2: *Publications*

As mentioned in the Context Chapter, this Thesis consists of a portfolio of eight (8) interconnected published papers in the field of infodemiology and infoveillance using Google Trends data.

[A] Mavragani A, Ochoa G, Tsagarakis KP. Assessing the Methods, Tools, and Statistical Approaches in Google Trends Research: Systematic Review. *Journal of Medical Internet Research*, **2018**;20(11):e270

[B] Mavragani A, Ochoa G (2019) Google Trends in Infodemiology and Infoveillance: Methodology Framework. *JMIR Public Health and Surveillance*, **2019**;5(2):e13439

[C] Mavragani A, Sampri A, Sypsa K, Tsagarakis KP. Integrating Smart Health in the US Health Care System: Infodemiology Study of Asthma Monitoring in the Google Era. *JMIR Public Health and Surveillance*, **2018**;4(1):e24

[D] Mavragani A & Ochoa G (2018) The Internet and the Anti-Vaccine Movement: Tracking the 2017 EU Measles Outbreak. *Big Data and Cognitive Computing*, **2018**;2(1):2

[E] Mavragani A, Ochoa G. Forecasting AIDS Prevalence in the United States using Online Search Traffic Data. *Journal of Big Data*, **2018**;5:17

[F] Mavragani A, Ochoa G. Infoveillance of Infectious Diseases in USA: STDs, Tuberculosis, and Hepatitis. *Journal of Big Data*, **2018**;5:30

[G] Mavragani A. Tracking COVID-19 in Europe: Infodemiology Approach. *JMIR Public Health and Surveillance*, **2020**;6(2):e18941

[H] Mavragani A, Gkillas K. COVID-19 predictability in the United States using Google Trends time series. *Scientific Reports*, **2020**;10:20693

The *Multimedia Appendices* for [A], [B], and [C] can be found in the *Appendix*.

Review

Assessing the Methods, Tools, and Statistical Approaches in Google Trends Research: Systematic Review

Amaryllis Mavragani[1], BSc, MSc; Gabriela Ochoa[1], BSc, MSc, PhD; Konstantinos P Tsagarakis[2], DipEng, PhD

[1]Department of Computing Science and Mathematics, University of Stirling, Stirling, Scotland, United Kingdom

[2]Department of Environmental Engineering, Democritus University of Thrace, Xanthi, Greece

Corresponding Author:
Amaryllis Mavragani, BSc, MSc
Department of Computing Science and Mathematics
University of Stirling
Stirling, Scotland, FK94LA,
United Kingdom
Phone: 44 7523782711
Email: amaryllis.mavragani1@stir.ac.uk

Abstract

Background: In the era of information overload, are big data analytics the answer to access and better manage available knowledge? Over the last decade, the use of Web-based data in public health issues, that is, infodemiology, has been proven useful in assessing various aspects of human behavior. Google Trends is the most popular tool to gather such information, and it has been used in several topics up to this point, with health and medicine being the most focused subject. Web-based behavior is monitored and analyzed in order to examine actual human behavior so as to predict, better assess, and even prevent health-related issues that constantly arise in everyday life.

Objective: This systematic review aimed at reporting and further presenting and analyzing the methods, tools, and statistical approaches for Google Trends (infodemiology) studies in health-related topics from 2006 to 2016 to provide an overview of the usefulness of said tool and be a point of reference for future research on the subject.

Methods: Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for selecting studies, we searched for the term "Google Trends" in the Scopus and PubMed databases from 2006 to 2016, applying specific criteria for types of publications and topics. A total of 109 published papers were extracted, excluding duplicates and those that did not fall inside the topics of health and medicine or the selected article types. We then further categorized the published papers according to their methodological approach, namely, visualization, seasonality, correlations, forecasting, and modeling.

Results: All the examined papers comprised, by definition, time series analysis, and all but two included data visualization. A total of 23.1% (24/104) studies used Google Trends data for examining seasonality, while 39.4% (41/104) and 32.7% (34/104) of the studies used correlations and modeling, respectively. Only 8.7% (9/104) of the studies used Google Trends data for predictions and forecasting in health-related topics; therefore, it is evident that a gap exists in forecasting using Google Trends data.

Conclusions: The monitoring of online queries can provide insight into human behavior, as this field is significantly and continuously growing and will be proven more than valuable in the future for assessing behavioral changes and providing ground for research using data that could not have been accessed otherwise.

(J Med Internet Res 2018;20(11):e270) doi:10.2196/jmir.9366

KEYWORDS

big data; health assessment; infodemiology; Google Trends; medicine; review; statistical analysis

Introduction

Big data are characterized by the 8 Vs [1]: volume (exponentially increasing volumes) [2], variety (wide range of datasets), velocity (high processing speed) [3], veracity, value [4,5], variability, volatility, and validity [1]. Big data have shown great potential in forecasting and better decision making [1]; though handling these data with conventional ways is inadequate [6], they are being continuously integrated in research [7] with novel approaches and methods.

The analysis of online search queries has been of notable popularity in the field of big data analytics in academic research [8,9]. As internet penetration is continuously increasing, the use of search traffic data, social media data, and data from other Web-based sources and tools can assist in facilitating a better understanding and analysis of Web-based behavior and behavioral changes [10].

The most popular tool for analyzing behavior using Web-based data is *Google Trends* [11]. Online search traffic data have been suggested to be a good analyzer of internet behavior, while Google Trends acts as a reliable tool in predicting changes in human behavior; subject to careful selection of the searched-for terms, Google data can accurately measure the public's interest [12]. Google Trends provides the field of big data with new opportunities, as it has been shown to be valid [13] and has been proven valuable [14,15], accurate [16], and beneficial [17] for forecasting. Therefore, great potential arises from using Web-based queries to examine topics and issues that would have been difficult or even impossible to explore without the use of big data. The monitoring of Web-based activity is a valid indicator of public behavior, and it has been effectively used in predictions [18,19], nowcastings [20], and forecasting [17,21,22].

Google Trends shows the changes in online interest for time series in any selected term in any country or region over a selected time period, for example, a specific year, several years, 3 weeks, 4 months, 30 days, 7 days, 4 hours, 1 hour, or a specified time-frame. In addition, different terms in different regions can be compared simultaneously. Data are downloaded from the Web in ".csv" format and are adjusted as follows: "*Search results are proportionate to the time and location of a query: Each data point is divided by the total searches of the geography and time range it represents, to compare relative popularity. Otherwise places with the most search volume would always be ranked highest. The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics. Different regions that show the same number of searches for a term will not always have the same total search volumes* " [23].

Healthcare is one of the fields in which big data are widely applied [24,25], with the number of publications in this field showing a high increase [26]. Researchers have placed a significant focus on examining Web-based search queries for health and medicine related topics [27]. Data from Google Trends have been shown to be valuable in predictions, detection of outbreaks, and monitoring interest, as detailed below, while such applications could be analyzed and evaluated by government officials and policy makers to deal with various health issues and disease occurrence.

The monitoring and analysis of internet data fall under the research field of infodemiology, that is, employing data collected from Web-based sources aiming at informing public health and policy [28]. These data have the advantage of being real time, thus tackling the issue of long periods of delay from gathering data to analysis and forecasting. Over the past decade, the field of infodemiology has been shown to be highly valuable in assessing health topics, retrieving web-based data from, for example, Google [29,30], Twitter [31-34], social media [35,36], or combinations of ≥2 Web-based data sources [37,38].

As the use of Google Trends in examining human behavior is relatively novel, new methods of assessing Google health data are constantly arising. Up to this point, several topics have been examined, such as epilepsy [39,40], cancer [41], thrombosis [42], silicosis [43], and various medical procedures including cancer screening examinations [44,45], bariatric surgery [46], and laser eye surgery [47].

Another trend rising is the measurement of the change in interest in controversial issues [48,49] and in drug-related subjects, such as searches in prescription [50] or illicit drugs [51,52]. In addition, Google Trends data have been used in examining interest in various aspects of the health care system [53-55].

Apart from the above, Google Trends data have also been useful in measuring the public's reaction to various outbreaks or incidents, such as attention to the epidemic of Middle East Respiratory Syndrome [56], the Ebola outbreak [57], measles [58], and Swine flu [59], as well as the influence of media coverage on online interest [60]. Google queries for the respective terms have been reported to increase or peak when a public figure or celebrity is related [61-65].

Google Trends has also been valuable in examining seasonal trends in various diseases and health issues, such as Lyme disease [66], urinary tract infection [67], asthma [30], varicose vein treatment [68], and snoring and sleep apnea [69]. Furthermore, Deiner et al [70] showed that indeed there exists the same seasonality in Google Trends and clinical diagnoses. What has also been reported is that seasonality in Google searches on tobacco is correlated with seasonality in Google searches on lung cancer [71], while online queries for allergic rhinitis have the same seasonality as in real life cases [72]. Thus, we observe that, apart from measuring public interest, Google Trends studies show that the seasonality of online search traffic data can be related to the seasonality of actual cases of the respective diseases searched for.

As mentioned above, Google queries have been used so far to examine general interest in drugs. Taking a step further, Schuster et al [73] found a correlation between the percentage change in the global revenues in Lipitor statin for dyslipidemia treatment and Google searches, while several other studies have reported findings toward this direction, that is, correlations of Web-based searches with prescription issuing [74-76]. The detection and monitoring of flu has also been of notable popularity in health assessment. Data from Google Flu Trends have been shown to correlate with official flu data [77,78], and Google data on the relevant terms correlate with cases of influenza-like illness [79].

In addition, online search queries for suicide have been shown to be associated with actual suicide rates [80,81], while other examples indicative of the relationship between Web-based data and human behavior include the correlations between official data and internet searches in veterinary issues [82], sleep deprivation [83], sexually transmitted infections [84], Ebola-related searches [85], and allergies [86,87].

Furthermore, Zhou et al [88] showed how the early detection of tuberculosis outbreaks can be improved using Google Trends

data; while suicide rates and Google data seem to be related, the former are suggested to be a good indicator for developing suicide prevention policies [89]. In addition, methamphetamine criminal behavior has been shown to be related to meth searches [90]. Finally, recent research on using Google Trends in predictions and forecasting include the development of predictive models of pertussis occurrence [91], while online search queries have been employed to forecast dementia incidence [92] and prescription volumes in ototopical antibiotics [93].

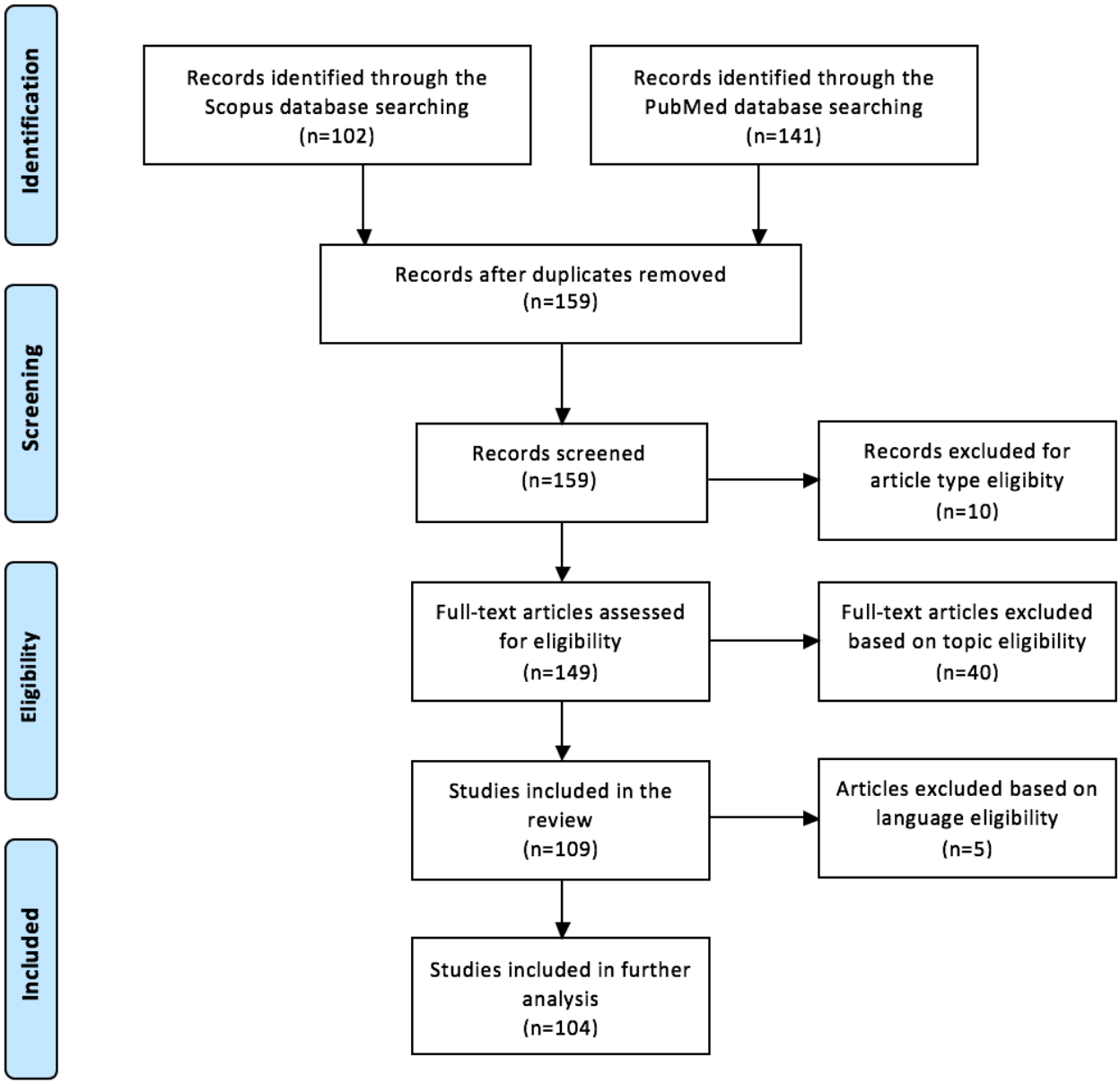
Given the diversity of subjects that Google Trends data have been used up for until this point to examine changes in interest and the usefulness of this tool in assessing human behavior, it is evident that the analysis of online search traffic data is indeed valuable in exploring and predicting behavioral changes.

In 2014, Nuti et al [27] published a systematic review of Google Trends research including the years up to 2013. This review was of importance as the first one in the field, and it reported Google Trends research up to that point. The current review differs from Nuti et al's in two ways. First, it includes 3 more full years of Google Trends research, that is, 2014, 2015, and 2016, which account for the vast majority of the research conducted in this field for the examined period based on our selection criteria. Second, while the first part of our paper is a systematic review reporting standard information, that is, authors, country, region, keywords, and language, the second part offers a detailed analysis and categorization of the methods, approaches, and statistical tools used in each of this paper. Thus, it serves as a point of reference in Google Trends research not only by subject or topic but by analysis or method as well.

Methods

The aim of this review was to include all articles on the topics of health and medicine that have used Google Trends data since its establishment in 2006 through 2016. We searched for the term "Google Trends" in the Scopus [94] and PubMed [95] databases from 2006 to 2016, and following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (Figure 1), the total number of publications included in this review was 109.

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram of the selection procedure for including studies.



First, we conducted a search in Scopus for the keyword "Google Trends" in the "Abstract-Title-Keywords" field for "Articles," "Articles in press," "Reviews," and "Conference papers" from 2006 to 2016. Out of the available categories, we selected "Medicine," "Biochemistry Genetics and Molecular Biology," "Neuroscience," "Immunology and Microbiology," "Pharmacology, Toxicology, and Pharmaceuticals," "Health Profession," "Nursing," and "Veterinary." The search returned 102 publications. Second, we searched for the keyword "Google Trends" in PubMed from 2006 to 2016, which provided a total of 141 publications. Excluding the duplicates, which numbered 84 in total, 159 publications met our criteria. Excluding the ones that did not match the criteria for article type (10 publications) and the ones that did not fall inside the scope of health and medicine (40 publications), a total of 109 studies were included in this review. Note that 5 studies were written in a language other than English and were therefore not included in the quantitative part or in the detailed analysis of the methods of each study. Figure 2 depicts the number of publications by year from 2009 to 2016: 2 in 2009, 3 in 2010, 2 in 2011, 1 in 2012, 12 in 2013, 21 in 2014, 28 in 2015, and 40 in 2016.

The selected studies are further analyzed according to their methodologies, and the gaps, advantages, and limitations of the tool have been discussed so as to assist in future research. Thus, we provide a more detailed categorization of the examined papers according to the main category that they belong to, that is, visualization and general time series analysis, seasonality, correlations, predictions or forecasting, modeling, and statistical method or tool employed. Note that a study can fall into >1 category. The categorization by individual medical field is not applicable due to the high number of individual topics. Table 1 consists of the description of each parameter used to classify each study.

Figure 2. Google Trends' publications per year in health-related fields from 2009 to 2016.

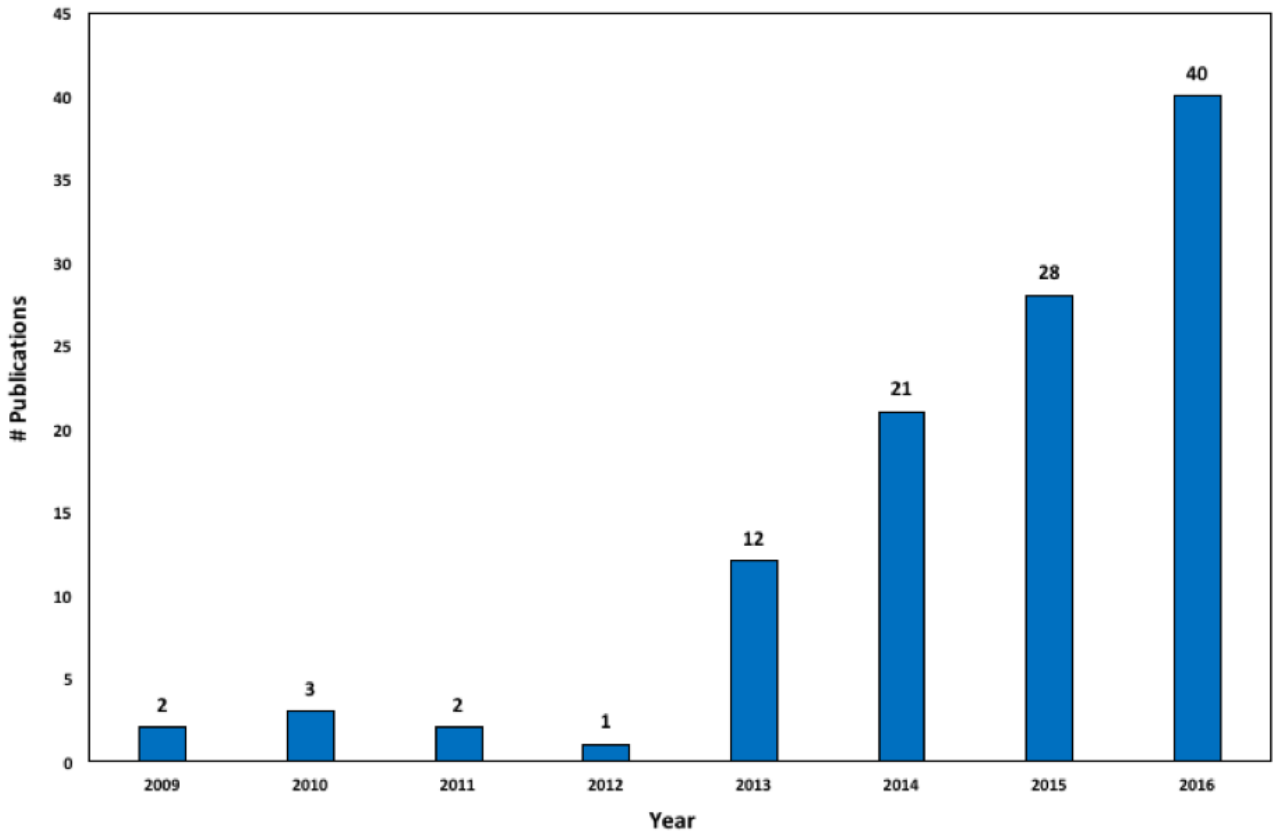


Table 1. Description of the parameters used for classification.

Parameter	Description
Authors	Includes the surname of the authors, date of publication, and link to the reference list (eg, *Smith et al, 2016 [57]*).
Period	Refers to the time-frame for which Google Trends data were retrieved and used in the study (eg, *2004-2015*).
Region	Refers to the country or countries or region (eg, *USA; Worldwide; Oceania*) that Google Trends data were extracted for.
Language	Refers to the language in which the Google Trends search was conducted (eg, search for the Italian word *Si*).
Keywords	Basic keywords are included in this category, mostly referring to the health topic examined and important keywords used to describe it.
Visualization (V)	Includes any form of visualization, that is, figures, maps, and screenshots (eg, screenshots of the Google Trends website).
Seasonality (S)	Studies that have explored the seasonality of the respective topic are included.
Correlations (C)	Studies that have examined correlations are included in this category. Correlations may be between Google Trends data and official data, among Google Trends time series, or between Google Trends and other Web-based sources' time series.
Forecasting (F)	This category includes studies that conducted forecasting of either Google Trends time series or diseases, outbreaks, etc, using Google Trends data, independent of the method used.
Modeling (M)	Studies in this category conducted some form of modeling using Google Trends data.
Statistical Tools (St)	This category includes the studies that used statistical tools or tests, eg, *t* test. Tools and methods for statistical modeling, (eg, *regression*), are not included in this category but only in the category of Modeling.

Results

Multimedia Appendix 1 consists of the first classification of the selected studies [27,39-57,59-93,96-144]; there are 104 in total, as the studies of Kohler et al [145], Orellano et al [146], Cjuno et al [147], Tejada-Llacsa [148], and Yang et al [149] are written in German, Spanish, or Chinese, and thus are not included in the more detailed categorization and analysis.

All the examined papers involve, by definition, time series analysis, and almost all include some form of visualization. Only 8.7% (9/104) studies used Google Trends data for predictions and forecasting, and 23.1% (24/104) used them for examining seasonality, while correlations and modeling were performed in 39.4% (41/104) and 32.7% (34/104) studies, respectively. As the category of forecasting and predictions exhibits the least number of studies, it is evident that a gap exists in the literature for forecasting using Google Trends in health assessment.

As is evident in Multimedia Appendix 1, Google queries have been employed up to this point in many countries and several languages. Figure 3 shows a worldwide map by examined country for assessing health and medicine related issues using Google Trends data up to 2016. Worldwide, the studies that explore topics related to the respective terms number 23 in total. As far as individual countries are concerned, US data have been employed in the most (60) studies, while other countries that have been significantly examined include the United Kingdom (15), Australia (13), Canada (9), Germany (8), and Italy (7).

The four most examined countries are English-speaking ones. The reasons for this could include that Google Trends, though not case-sensitive, does take into account accents and spelling mistakes; therefore, for countries with more complicated alphabets, the analysis of Web-based data should be more careful. In addition, other factors that could play a significant role and are taken into account when choosing the countries to be examined using online search traffic data are the availability of official data, the openness of said data, any internet restrictions or monitoring in countries with lower scores in freedom of press or freedom of speech, and internet penetration.

The rest of the analysis consists of the further breaking down of the initial categorization to include the respective methods that were used for examining seasonality, correlations, forecasting, and performing statistical tests and estimating models, along with a concise introduction to each of these methods and how they were used to assess health issues.

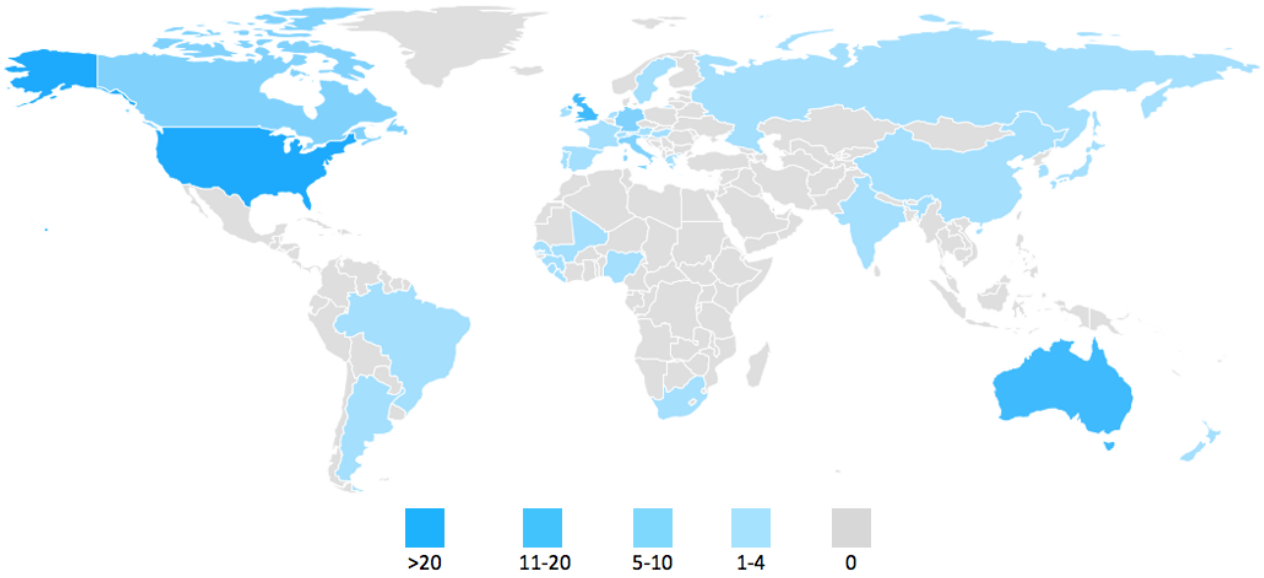
Table 2 shows the methods used to explore seasonality; Tables 3 and 4 present the methods used to examine correlations and perform predictions and forecasting, respectively. Finally, Tables 5 and 6 list the modeling methods and other statistical tools employed in health assessment using Google Trends.

The most popular way to explore seasonality is to use visual evidence and examine and discuss peaks, as shown in Table 2. Furthermore, several studies have used cosinor analysis [8,69,134,138,142], which is a time series analysis method for seasonal data using least squares.

Apart from seasonality [122], analysis of variance (ANOVA) has been also used for geographical comparisons between regions or countries [49,51,68,93] and between differences in monthly data [41]. It is a test used for examining if significant differences between means exist. In the case of 2 means, *t* test is the equivalent to ANOVA.

The Kruskal-Wallis test is also a popular method for examining seasonality using Google Trends [57,68,113]. It is a nonparametric, independent of distribution test, for continuous as well as ordinal-level dependent variables, employed when the one-way ANOVA assumptions do not hold, that is, for examining statistically significant differences between ≥3 groups. It uses random sample with independent observations, with the dependent variable being at least ordinal.

Figure 3. Countries by number of Scopus and PubMed publications using Google Trends.



Other methods of exploring seasonality include the nonparametric tests (independent of distribution) Wilcoxon signed rank [18,113] and Mann-Whitney U test [67], which are used for comparing data in different seasons or time periods when the equivalent parametric *t* tests cannot be used. The latter has been also used by some studies to compare weekly data [105] and differences among regions [113].

For examining correlations (Table 3), the vast majority of the studies used the Pearson correlation coefficient, which examines the strength of association between 2 quantitative, continuous variables, employed when the relationship is linear. The Spearman rho (rank-order) correlation, the second most used method, is the nonparametric version of the Pearson correlation, has also been used to explore seasonality between time series [70]. Spearman correlation coefficient (denoted by ρ or r_s) measures the levels to which 2 ranked variables (ordinal, interval, or ratio) are related to each other.

Cross-correlations are used for examining the relationship of 2 time series, while simultaneously exploring if the data are periodic. It is often employed in correlating Google Trends data with observed data [50,82,90,135] and between different Google search terms [80], while it can be also used for examining linear and temporal associations of seasonal data [71]. Cross-correlations have been also used in forecasting, where Wang et al [92] showed that cross-correlations of new dementia cases with Google Trends data can assist with the forecasting of dementia cases, and Solano et al [80] forecasted the suicide rates 2 years ahead using Google queries. The autocorrelations are basically cross-correlations for one time series, that is, a time series cross-correlated with itself.

The Kendall's tau-b test correlation coefficient is a nonparametric alternative to Pearson and Spearman correlations and is used to measure the strength and direction of the relationship between 2 (at least ordinal) variables. It has been employed by 1 study [138] to examine the correlations between Google Trends data and the results of a paper interview survey.

The Spearman-Brown prediction (or prophecy) formula is used to predict how reliable the test is after changing its length. It has also been employed by only 1 study [65] to explore the relationship between railway suicide and Google hits.

The generalized linear model estimates the linear relationship between a dependent and ≥1 independent variables. It was used by Domnich et al [79] to predict influenza-like illness morbidity, with the exploratory variables being "Influenza," "Fever," and "Tachipirin search volumes," along with the Holt-Winters method and the autoregressive moving average process for the residuals. Holt-Winters is a method employed in exploring the seasonality in time series, and for predictions, the autoregressive moving average (also called the Box-Jenkins model) is a special case of the autoregressive integrated moving average, used for the analysis of time series and predictions.

Autoregressive integrated moving average is a commonly used method for time series analysis and predictions [55,63,86,92,141], the latter having also been assessed by linear regressions and modeling [88,91]. Multivariable regressions are used to estimate the relationship of ≥2 independent variables with a dependent one. In Google Trends, they have been used to relate Ebola searches, reported cases, and the Human Development Index [85] and to study the relationship between climate and environmental variables and Google hits [125].

Hierarchical linear modeling is a regression of ordinary least squares that is employed to analyze hierarchically structured data, that is, units that are grouped together, and it has been employed by 1 study so far [83].

The Mann-Kendall test, which is the nonparametric alternative test to the independent sample, has been used to show the statistical differences of peaks [43] and to detect trends [140]. Finally, the *t* test is used to compare 2 sample means of the same population, and it has been employed for comparing Google searches with the baseline period [105] and to examine the statistical differences of peaks [41].

Table 2. Methods for exploring seasonality with Google Trends in health assessment.

Number	Authors	Method	Description
1	Bakker et al, 2016 [96]	Morlet Wavelet Analysis	To test the seasonality of Google Trends data in the examined countries
2	Braun and Harreus, 2013 [104]	Visual evidence	N/A[a]
3	Crowson et al, 2016 [93]	Seasonal peaks	N/A
4	Deiner et al, 2016 [70]	Spearman correlation	Correlating the seasonality of clinical diagnoses with Google Trends data
5	El-Sheikha, 2015 [113]	Kruskal-Wallis test	To show seasonality for different months
6	Garrison et al, 2015 [116]	Least-squares sinusoidal model	Variability in outcomes (supported also from a comparison with searches in Australia)
7	Harsha et al, 2014 [68]	Kruskal-Wallis test	Seasonal (monthly) comparisons
8	Harsha et al, 2015 [119]	Kruskal-Wallis test	Seasonal (monthly) comparisons
9	Hassid et al, 2016 [120]	Pearson correlation	To examine seasonal variations across symptoms
10	Ingram and Plante, 2013 [122]	Cosinor analysis; analysis of variance	To test the seasonal variation of the normalized Google Trends data; to compare the seasonal increase among the examined countries
11	Ingram et al, 2015 [69]	Cosinor analysis	To test the seasonal variation of the normalized Google Trends data
12	Kang et al, 2015 [72]	Visual observation	N/A
13	Leffler et al, 2010 [125]	Correlations	Showing correlations among the 4 seasons for the 39 examined terms
14	Liu et al, 2016 [127]	Seasonal model and a null model	Seasonality explained the searches significantly better with an F-test
15	Phelan et al, 2016 [133]	Correlograms (autocorrelations plots)	Visual interpretation for exploring seasonal peaks
16	Plante and Ingram, 2014 [134]	Cosinor analysis	To test the seasonal variation of the normalized Google Trends data
17	Rossignol et al, 2013 [67]	Mann-Whitney U test; Harmonic Product Spectrum	Comparison of summer vs winter hits; evaluation of seasonality
18	Seifter et al, 2010 [66]	Visual evidence	N/A
19	Sentana-Lledo et al, 2016 [138]	Cosinor analysis	To test the seasonal variations of the Google Trends data
20	Takada, 2012 [139]	Visual evidence	N/A
21	Telfer and Woodburn, 2015 [140]	Two-way Wilcoxon signed rank test	To explore differences between winter and summer
22	Toosi and Kalia, 2015 [142]	Visual evidence; cosinor analysis	To identify differences in seasonality between countries
23	Willson et al, 2015 [86]	Visual evidence	N/A
24	Zhang et al, 2015 [71]	Periodograms; ideal pass filter	To study the periodograms; to extract seasonal components

[a]N/A: not applicable.

Many studies have employed Google Trends for visualizing the changes in online interest or discussing peaks and spikes [60,62,123,124]. Brigo and Trinka [40] and Brigo et al [39] have studied the search volumes for related terms, Chaves et al [109] and Luckett et al [128] have explored terms related to the studied topic, and Davis et al [110] have examined related internet searches. Other approaches include the reporting of the polynomial trend lines [46] and investigation of statistically significant differences in yearly increases [119]. In addition, "Google Correlate" has been used to explore related terms [91,138].

Finally, several studies have used other sources of big data, namely, Google News [43,63,80], Twitter [43,54,61,63,108], Yandex [52], Baidu [121], Wikipedia [43,63], Facebook and Google+ [54], and YouTube [43,54,63]. Google is the most popular search engine. However, other Web-based sources are used or even preferred to Google in some regions; therefore, many studies use data from these sources to examine general interest in the respective subjects, compare them to Google Trends data, or use them together as variables.

Table 3. Methods of exploring correlations using Google Trends in health assessment.

Number	Authors	Method	Description
1	Alicino et al, 2015 [85]	Pearson correlation	Ebola-related Google Trends data with Ebola cases
2	Arora et al, 2016 [81]	Spearman correlation	Suicide search activity vs official suicide rates (and per age)
3	Bakker et al, 2016 [96]	Correlations	Between Google Trends data and reported cases
4	Bragazzi et al, 2016 [99]	Pearson correlation	Between Google Trends data and epidemiological data
5	Bragazzi, 2013 [98]	Autocorrelation; Pearson correlation	For the time series for multiple sclerosis (MS); between MS terms
6	Bragazzi et al, 2016 [101]	Autocorrelation; Partial Autocorrelation	To compute correlation of the time series with its own values
7	Bragazzi et al, 2016 [102]	Pearson correlation	Status epilepticus terms with etiology and management related terms
8	Bragazzi et al, 2016 [43]	Pearson correlation	Google searches for Silicosis with Normalized Google News, Google Scholar, PubMed Publications, Twitter traffic, Wikipedia
9	Bragazzi et al, 2016 [63]	Pearson correlation	Among Google Trends data and other data generating sources
10	Bragazzi, 2014 [103]	Pearson correlation; autocorrelation and partial autocorrelation	Nonsuicidal self-injury and related terms; nonsuicidal self-injury plots showed regular cyclical pattern
11	Cavazos-Regh et al, 2015 [107]	Pearson correlation	Among Google Trends data for noncigarette tobacco and prevalence
12	Cho et al, 2013 [78]	Pearson correlation	Google flu-related queries with surveillance data for different influenza seasons
13	Crowson et al, 2016 [93]	Pearson correlation	Between the selected keywords. Between medical prescriptions data and Google Trends data
14	Deiner et al, 2016 [70]	Spearman correlation	For correlating seasonality of clinical diagnoses with Google Trends data
15	Domnich et al, 2015 [79]	Pearson correlation	Among the examined search terms and influenza-like illness
16	Foroughi et al, 2016 [115]	Rank correlations; cross-country correlations; Pearson correlations	For search volumes; for the search volumes for cancer; for the weekly search volumes between countries
17	Gahr et al, 2015 [75]	Pearson correlation	Among annual prescription volumes and Google Trends data
18	Gamma et al, 2016 [90]	Cross-correlations	Cross-correlations between search volumes and crime statistics
19	Gollust et al, 2016 [117]	Multinomial Logit Models	To relate health insurance rates
20	Guernier et al, 2016 [82]	Spearman correlation; cross-correlation	Correlating the examined search terms with notifications of tick paralysis cases record; with lag values from −7 to +7 months
21	Hassid et al, 2016 [120]	Pearson correlation	Between Google Trends data and National Inpatient Sample data
22	Johnson et al, 2014 [84]	Pearson correlation	Pearson correlations to explore the relation of Google Trends data and sexually transmitted infection reported rates
23	Kang et al, 2013 [77]	Pearson correlation	To explore the association of (and among) search terms with surveillance data
24	Kang et al, 2015 [72]	Spearman correlation	Google Trends data for allergic rhinitis and related Google Trends terms and real world epidemiologic data for the United States
25	Koburger et al, 2015 [65]	Spearman-Brown correlation	To explore relations among Google Trends data and railway suicides
26	Ling and Lee, 2016 [126]	Pearson correlation	Between disease prevalence and Google Trends data
27	Mavragani et al, 2016 [76]	Pearson correlation	Between Google Trends data and published papers and Google Trends data with prescriptions
28	Phelan et al, 2016 [133]	Linear Regression	To examine if there is significant correlation between searches and time

Number	Authors	Method	Description
29	Poletto et al, 2016 [56]	Pearson correlation	Between Google Trends data and number of alerts published by ProMED mail and the number of Disease Outbreak News published by the World Health Organization
30	Pollett et al, 2015 [91]	Pearson correlation	To shortlist related search terms to pertussis
31	Rohart et al, 2016 [135]	Spearman rank correlations; Spearman correlation; cross-correlations	For the diseases examined; correlations between diseases and the investigated search metrics; to identify best lags
32	Shin et al, 2016 [137]	Spearman correlation	Between Google Trends data and the number of confirmed cases of Middle East Respiratory Syndrome and for quarantined cases of Middle East Respiratory Syndrome
33	Schootman et al, 2015 [45]	Pearson correlation	Between Respiratory Syncytial Virus and Behavioral Risk Factor Surveillance System prevalence data for 5 cancer screening tests
34	Schuster et al, 2010 [73]	Correlations	Lipitor Google Trends data and Lipitor revenues
35	Sentana-Lledo et al, 2016 [138]	Kendall's Tau-b test	To explore the correlation of Google Trends data with paper interview survey results
36	Simmering et al, 2014 [50]	Cross-correlations	Between Google Trends data for drugs and drug utilization, to see changes in search volumes following knowledge events
37	Solano et al, 2016 [80]	Correlations; cross-correlations	Between Google Trends data for suicide and national suicide rates; between different search terms
38	Wang et al, 2015 [92]	Pearson correlation	Between Google Trends data and new dementia cases
39	Willson et al, 2015 [86]	Spearman correlation	Between Google Trends data and observed data for aeroallergens
40	Zhang et al, 2015 [71]	Cross-correlations	To examine linear and temporal associations of the seasonal data
41	Zhang et al, 2016 [51]	Pearson correlation	To study pairwise comparisons among searches for different terms in Google Trends

Table 4. Forecasting and predictions using Google Trends in health assessment.

Number	Authors	Method	Description
1	Bakker et al, 2016 [96]	Statistical model	For forecasting chicken poxforce of infection, that is, monthly per capita rate of infection of children 0-14
2	Domnich et al, 2015 [79]	Generalized least squares (maximum likelihood estimates); Holt-Winters	Query-based models to predict influenza-like illness morbidity, with the exploratory variables: Influenza, Fever, Tachipirin; compared for forecasting power with Holt-Winters based on the real data (hold out set)
3	Parker et al, 2016 [132]	Statistical model	For forecasting deaths for 1 year in advance (2015)
4	Pollett et al, 2015 [91]	Prediction model	Tested the predicted model with a left-out dataset for prediction accuracy
5	Rohart et al, 2016 [135]	Linear models	To forecast with 1 or 2 weeks step
6	Solano et al, 2016 [80]	Cross-Correlations	Forecasting for suicides for 2 years without data (2013-14) based on Google Trends data of those years
7	Wang et al, 2015 [92]	Cross-Correlations	To investigate forecasting with lags of 0-12 months
8	Zhang et al, 2016 [51]	Autoregressive Moving Average	To predict Respiratory Syncytial Virus for "dabbing"
9	Zhou et al, 2011 [88]	Dynamic model	To provide real time estimations by correcting the forecasting with the new morbidity data when published

Table 5. Statistical modeling using Google Trends in health assessment.

Number	Authors	Method	Description
1	Alicino et al, 2015 [85]	Multivariate regression	For relating Ebola Google Trends data, number of Ebola Cases, and the Human Development Index
2	Bakker et al, 2016 [96]	Statistical model	For forecasting chicken poxforce of infection, that is, monthly per capita rate of infection
3	Bentley and Ormerod, 2009 [59]	Maximum likelihood estimation	Established social model for engaging a new behavior for Web-based searching for flu terms
	Barnes et al, 2015 [83]	Hierarchical linear modeling	Three levels: 3 Mondays, 6 years, 47 search terms
4	Bragazzi, 2013 [98]	Multiple linear regression	To confirm multiannual long-term trends
5	Domnich et al, 2015 [79]	Generalized linear model, autoregressive moving average process	Query volume-based models to predict influenza-like illness morbidity
6	El-Sheikha, 2015 [113]	Linear regression	To show the global, regional, and country level interest for the search term
7	Fenichel et al, 2013 [114]	Moving average, generalized linear model	Google Trends data as a variable in predicting loses in flights
8	Garrison et al, 2015 [116]	Seasonal model	Best fit combination of a straight line and a sinusoid
9	Gollust et al, 2016 [117]	Multinomial logit models	To relate health insurance rates
10	Haney et al, 2014 [55]	ARIMA[a]	Radiology residency interest
11	Harsha et al, 2014 [68]	Linear model	Statistical justification of annual increase in search volumes
12	Harsha et al, 2015 [119]	Linear model	Statistical justification of annual increase in search volumes and of the Web-based interest related to applications for interventional radiology
13	Leffler et al, 2010 [125]	Multivariable Linear Regressions	For studying the effect of climatic and environmental variables to internet searches
17	Linkov et al, 2014 [46]	Polynomial trend lines	Fitted spline polynomial trend lines per time without statistical reporting
18	Liu et al, 2016 [127]	Seasonal model	Best fit combination of a straight line and a sinusoid
19	Majumder et al, 2016 [129]	Linear Smoothing	To adjust HealthMap to using Google Trends, model fits
20	Noar et al, 2013 [64]	Linear Regression	To estimate the slope coefficient for changes in the magnitude of the effect size of Google Trends data and media search increases
21	Parker et al, 2016[132]	L1-regularization on Google Trends	To build a model for forecasting deaths in each state
22	Phelan et al, 2014 [49]	Linear Regression	To estimate the relation between news reports and search activity
23	Phelan et al, 2016 [133]	Linear Regression	To examine if there is a significant correlation between searches and time
24	Pollett et al, 2015 [91]	Linear Regression	Prediction model for pertussis cases based on Google Trends data of the most related terms
25	Rohart et al, 2016 [135]	Linear models	To forecast with 1 or 2 weeks step
26	Scatà et al, 2016 [136]	Epidemic model	Google Trends data is a measure of awareness, along with other sources
27	Schuster et al, 2010 [73]	Generalized Linear models	Google Trends data for the examined drugs, Google Trends data and changes in annual revenues, and Google Trends data vs resource utilization
28	Stein et al, 2013 [47]	Regression Fit Lines	To examine differences in queries
29	Telfer and Woodburn, 2015 [140]	Visual decomposition; local regression	Figures 4, 6 and 8; regression-based decomposition of the time series for the search terms
30	Troelstra et al, 2016 [141]	ARIMA	To account for dependency between data points in time series for “quit smoking” searches
31	Willson et al, 2015 [86]	ARIMA	To quantify the effect of the observed (pollen) counts with the levels of search activity

Number	Authors	Method	Description
32	Willson et al, 2015 [87]	ARIMA	To quantify the effect of the observed (pollen) counts with the levels of search activity
33	Yang et al, 2015 [144]	Prediction model (ARGO[b])	To predict influenza-like illness
34	Zhou et al, 2011 [88]	Dynamic Modeling	For forecasting tuberculosis incidents using Google Trends data

[a]ARIMA: autoregressive integrated moving average.

[b]ARGO: autoregression with Google search data.

Table 6. Statistical tests and tools using Google Trends in health assessment.

Number	Authors	Method	Description
1	Bragazzi et al, 2016 [43]	Mann-Kendall test	To show the statistical difference of peaks from the remaining period
2	Bragazzi et al, 2016 [63]	ARIMA[a]	To show increased web searches due to an event, and correct seasonality
3	Campen et al, 2014 [105]	Independent samples *t* test; Mann-Whitney U test with Bonferroni correction	For comparing searches with baseline period; for multiple weekly data comparisons
4	Crowson et al, 2016 [93]	ANOVA[b] (Post-hoc Tukey test)	To compare grouped geographical federal regions of the United States (Northeast, Midwest, South, West)
5	El-Sheikha, 2015 [113]	Wilcoxon rank test; Mann-Whitney	To study the change of interest at different time periods; to compare Web-based interest between the Northern and Southern hemispheres
6	Gahr et al, 2015 [75]	Coefficients of determination	To determine the amount of variability between annual prescription volumes and Google search terms
7	Harsha et al, 2014 [68]	ANOVA (Tukey-Kramer post hot test)	For the comparisons of US regions
8	Murray et al, 2016 [41]	ANOVA; *t* test	To explore differences in months' means per year; for the statistical differences of peaks compared with the remaining hits
9	Noar et al, 2013 [64]	Augmented Dickey-Fuller tests	To test for nonstationarity of the time series
10	Phelan et al, 2014 [49]	ANOVA	To explore differences among countries
11	Rohart et al, 2016 [135]	Mean Square Error for Prediction	To assess prediction accuracy
12	Telfer and Woodburn, 2015 [140]	Mann-Kendall trend tests	To detect trends significantly larger than the variance in the data for search terms
13	Troelstra et al, 2016 [141]	ARIMA	Studied the effect of smoking cessation policies with ARIMA interrupted time series modeling (Multimedia Appendix 1)
14	Zhang et al, 2015 [71]	Augmented Dickey-Fuller test	To detect whether or not the extracted seasonal components of the studied trends were stationary
15	Zhang et al, 2016 [51]	ANOVA	To examine the search interest for dabbing between groups of legal status states in the United States

[a]ARIMA: autoregressive integrated moving average.

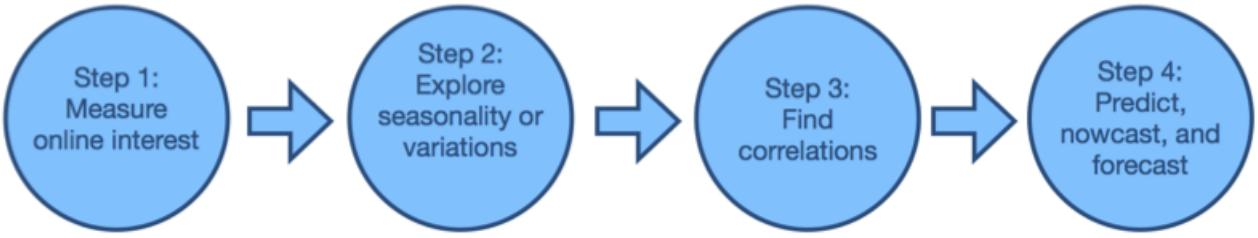
[b]ANOVA: analysis of variance.

Discussion

Principal Findings

With internet penetration constantly growing, users' Web-based search patterns can provide a great opportunity to examine and further predict human behavior. In addressing the challenge of big data analytics, Google Trends has been a popular tool in research over the past decade, with its main advantage being that it uses the revealed and not the stated data. Health and medicine are the most popular fields where Google Trends data have been employed so far to examine and predict human behavior. This review provides a detailed overview and classification of the examined studies (109 in total from 2006 through 2016), which are then further categorized and analyzed by approach, method, and statistical tools employed for data analysis.

Figure 4. The four steps toward employing Google Trends for health assessment.



The vast majority of studies using Google Trends in health assessment so far have included data visualization, that is, figures, maps, or screenshots. As discussed in the analysis, the most popular way of using Google Trends data in this field is correlating them with official data on disease occurrence, spreading, and outbreaks. The assessment of suicide tendencies and (prescription or illegal) drug-related queries has been of notably growing popularity over the course of the last years. As is evident, the gap in the existing literature is the use of Google Trends for predictions and forecasting in health-related topics and issues. Though data on reported cases of various health issues and the respective Google Trends data have been correlated in a large number of studies, only a few have proceeded with forecasting incidents and occurrences using online search traffic data.

In research using Google Trends in health and medicine from 2006 to 2016, the ultimate goal is to be able to use and analyze Web-based data to predict and provide insight to better assess health issues and topics. The four main steps, based on the presentation of the papers published up to this point in assessing health using Google Trends, are as follows (Figure 4):

1. Measure the general Web-based interest.
2. Detect any variations or seasonality of Web-based interest, and proceed with examining any relations between actual events or cases.
3. Correlate Web-based search queries among them or with official or actual data and events.
4. Predict, nowcast, and forecast health-related events, outbreaks, etc.

Limitations

This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for selecting the examined papers from the Scopus and PubMed databases. Though this includes the majority of papers published on the topic from 2006 to 2016, the studies that are not indexed in these databases or are not indexed based on the selection criteria used in this review were not included in further analysis. In addition, as is evident in Figure 2, research using Google Trends data has shown a significant increase from each year to the next since 2013. This review included studies published in Google Trends research through 2016. However, there are several studies published in 2017 and 2018 that are not included. This review provides, at first, an overall description of each examined study, which is standard review information. The second part is a classification and assessment of the methodology, tools, and results of each study. Though the first part mainly reports what is included in the methodology of each study, the second part could include a bias, as it is the authors' assessment and categorization of the methods employed based on the results obtained after a very careful and thorough examination of each individual study.

Conclusions

This review consists of the studies published from 2006 to 2016 on Google Trends research in the Scopus and PubMed databases based on the selected criteria. The aim of this review was to serve as a point of reference for future research in health assessment using Google Trends, as each study, apart from the basic information, for example, period, region, language, is also categorized by the method, approach, and statistical tools employed for the analysis of the data retrieved from Google Trends. Google Trends data are being all the more integrated in infodemiology research, and Web-based data have been shown to empirically correlate with official health data in many topics. It is thus evident that this field will become increasingly popular in the future in health assessment, as the gathering of real time data is crucial in monitoring and analyzing seasonal diseases as well as epidemics and outbreaks.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Publication details and categorization.

[PDF File (Adobe PDF File), 256KB - jmir_v20i11e270_app1.pdf]

References

1. Al Nuaimi E, Al Neyadi H, Mohamed N, Al-Jaroodi J. Applications of big data to smart cities. J Internet Serv Appl 2015 Dec 1;6(1):1-15. [doi: 10.1186/s13174-015-0041-5]

2. Hilbert M, López P. The world's technological capacity to store, communicate, and compute information. Science 2011 Apr 01;332(6025):60-65 [FREE Full text] [doi: 10.1126/science.1200970] [Medline: 21310967]
3. Philip Chen C, Zhang C. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences 2014 Aug;275:314-347. [doi: 10.1016/j.ins.2014.01.015]
4. Jin X, Wah BW, Cheng X, Wang Y. Significance and Challenges of Big Data Research. Big Data Research 2015 Jun;2(2):59-64. [doi: 10.1016/j.bdr.2015.01.006]
5. Fosso Wamba S, Akter S, Edwards A, Chopin G, Gnanzou D. How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics 2015 Jul;165:234-246. [doi: 10.1016/j.ijpe.2014.12.031]
6. Chang RM, Kauffman RJ, Kwon Y. Understanding the paradigm shift to computational social science in the presence of big data. Decision Support Systems 2014 Jul;63:67-80. [doi: 10.1016/j.dss.2013.08.008]
7. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 2015 Apr;35(2):137-144. [doi: 10.1016/j.ijinfomgt.2014.10.007]
8. Preis T, Moat HS, Stanley HE, Bishop SR. Quantifying the advantage of looking forward. Sci Rep 2012 Apr;2:350 [FREE Full text] [doi: 10.1038/srep00350] [Medline: 22482034]
9. Preis T, Moat HS, Stanley HE. Quantifying trading behavior in financial markets using Google Trends. Sci Rep 2013 Apr;3:1684 [FREE Full text] [doi: 10.1038/srep01684] [Medline: 23619126]
10. Burnap P, Rana OF, Avis N, Williams M, Housley W, Edwards A, et al. Detecting tension in online communities with computational Twitter analysis. Technological Forecasting and Social Change 2015 Jun;95:96-108. [doi: 10.1016/j.techfore.2013.04.013]
11. Google Trends. URL: https://trends.google.com/trends/explore [accessed 2017-11-08] [WebCite Cache ID 6uot1LkyX]
12. Scharkow M, Vogelgesang J. Measuring the Public Agenda using Search Engine Queries. International Journal of Public Opinion Research 2011 Mar 01;23(1):104-113. [doi: 10.1093/ijpor/edq048]
13. McCallum ML, Bury GW. Public interest in the environment is falling: a response to Ficetola (2013). Biodivers Conserv 2014 Feb 14;23(4):1057-1062. [doi: 10.1007/s10531-014-0640-7]
14. Jun S, Park D. Consumer information search behavior and purchasing decisions: Empirical evidence from Korea. Technological Forecasting and Social Change 2016 Jun;107:97-111. [doi: 10.1016/j.techfore.2016.03.021]
15. Jun S, Park D, Yeom J. The possibility of using search traffic information to explore consumer product attitudes and forecast consumer preference. Technological Forecasting and Social Change 2014 Jul;86:237-253. [doi: 10.1016/j.techfore.2013.10.021]
16. Han S, Chung H, Kang B. It is time to prepare for the future: Forecasting social trends. Communications in Computer and Information Scienc 2012:e2012-e2031.
17. Vosen S, Schmidt T. Forecasting private consumption: survey-based indicators vs. Google trends. J. Forecast 2011 Jan 13;30(6):565-578. [doi: 10.1002/for.1213]
18. Choi H, Varian H. Predicting the Present with Google Trends. Economic Record. (SUPPL.1) 2012;88:2-9. [doi: 10.1111/j.1475-4932.2012.00809.x]
19. Mavragani A, Tsagarakis KP. YES or NO: Predicting the 2015 GReferendum results using Google Trends. Technological Forecasting and Social Change 2016 Aug;109:1-5. [doi: 10.1016/j.techfore.2016.04.028]
20. Carrière-Swallow Y, Labbé F. Nowcasting with Google Trends in an Emerging Market. J. Forecast 2011 Nov 20;32(4):289-298. [doi: 10.1002/for.1252]
21. Vicente MR, López-Menéndez AJ, Pérez R. Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? Technological Forecasting and Social Change 2015 Mar;92:132-139. [doi: 10.1016/j.techfore.2014.12.005]
22. Jun S, Yeom J, Son J. A study of the method using search traffic to analyze new technology adoption. Technological Forecasting and Social Change 2014 Jan;81:82-95. [doi: 10.1016/j.techfore.2013.02.007]
23. Google. How Trends data is adjusted URL: https://support.google.com/trends/answer/4365533?hl=en [accessed 2017-11-08] [WebCite Cache ID 6uot8lARg]
24. Fan J, Han F, Liu H. Challenges of Big Data Analysis. Natl Sci Rev 2014 Jun;1(2):293-314 [FREE Full text] [doi: 10.1093/nsr/nwt032] [Medline: 25419469]
25. Yoo C, Ramirez L, Liuzzi J. Big data analysis using modern statistical and machine learning methods in medicine. Int Neurourol J 2014 Jun;18(2):50-57 [FREE Full text] [doi: 10.5213/inj.2014.18.2.50] [Medline: 24987556]
26. Gu D, Li J, Li X, Liang C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. Int J Med Inform 2017 Dec;98:22-32. [doi: 10.1016/j.ijmedinf.2016.11.006] [Medline: 28034409]
27. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of google trends in health care research: a systematic review. PLoS One 2014 Oct;9(10):e109583 [FREE Full text] [doi: 10.1371/journal.pone.0109583] [Medline: 25337815]
28. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]

29. Phillips CA, Barz LA, Li Y, Schapira MM, Bailey LC, Merchant RM. Relationship Between State-Level Google Online Search Volume and Cancer Incidence in the United States: Retrospective Study. J Med Internet Res 2018 Jan 08;20(1):e6 [FREE Full text] [doi: 10.2196/jmir.8870] [Medline: 29311051]
30. Mavragani A, Sampri A, Sypsa K, Tsagarakis KP. Integrating Smart Health in the US Health Care System: Infodemiology Study of Asthma Monitoring in the Google Era. JMIR Public Health Surveill 2018 Mar 12;4(1):e24 [FREE Full text] [doi: 10.2196/publichealth.8726] [Medline: 29530839]
31. Chen T, Dredze M. Vaccine Images on Twitter: Analysis of What Images are Shared. J Med Internet Res 2018 Apr 03;20(4):e130 [FREE Full text] [doi: 10.2196/jmir.8221] [Medline: 29615386]
32. Farhadloo M, Winneg K, Chan MS, Hall JK, Albarracin D. Associations of Topics of Discussion on Twitter With Survey Measures of Attitudes, Knowledge, and Behaviors Related to Zika: Probabilistic Study in the United States. JMIR Public Health Surveill 2018 Feb 09;4(1):e16 [FREE Full text] [doi: 10.2196/publichealth.8186] [Medline: 29426815]
33. Simpson SS, Adams N, Brugman CM, Conners TJ. Detecting Novel and Emerging Drug Terms Using Natural Language Processing: A Social Media Corpus Study. JMIR Public Health Surveill 2018 Jan 08;4(1):e2 [FREE Full text] [doi: 10.2196/publichealth.7726] [Medline: 29311050]
34. van Lent LG, Sungur H, Kunneman FA, van de Velde B, Das E. Too Far to Care? Measuring Public Attention and Fear for Ebola Using Twitter. J Med Internet Res 2017 Jun 13;19(6):e193 [FREE Full text] [doi: 10.2196/jmir.7219] [Medline: 28611015]
35. Wongkoblap A, Vadillo MA, Curcin V. Researching Mental Health Disorders in the Era of Social Media: Systematic Review. J Med Internet Res 2017 Jun 29;19(6):e228 [FREE Full text] [doi: 10.2196/jmir.7215] [Medline: 28663166]
36. Huesch M, Chetlen A, Segel J, Schetter S. Frequencies of Private Mentions and Sharing of Mammography and Breast Cancer Terms on Facebook: A Pilot Study. J Med Internet Res 2017 Jun 09;19(6):e201 [FREE Full text] [doi: 10.2196/jmir.7508] [Medline: 28600279]
37. Lu FS, Hou S, Baltrusaitis K, Shah M, Leskovec J, Sosic R, et al. Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis. JMIR Public Health Surveill 2018 Jan 09;4(1):e4 [FREE Full text] [doi: 10.2196/publichealth.8950] [Medline: 29317382]
38. Roccetti M, Marfia G, Salomoni P, Prandi C, Zagari RM, Gningaye KFL, et al. Attitudes of Crohn's Disease Patients: Infodemiology Case Study and Sentiment Analysis of Facebook and Twitter Posts. JMIR Public Health Surveill 2017 Aug 09;3(3):e51 [FREE Full text] [doi: 10.2196/publichealth.7004] [Medline: 28793981]
39. Brigo F, Igwe SC, Ausserer H, Nardone R, Tezzon F, Bongiovanni LG, et al. Why do people Google epilepsy? An infodemiological study of online behavior for epilepsy-related search terms. Epilepsy Behav 2014 Feb;31:67-70. [doi: 10.1016/j.yebeh.2013.11.020] [Medline: 24361764]
40. Brigo F, Trinka E. Google search behavior for status epilepticus. Epilepsy Behav 2015 Aug;49:146-149. [doi: 10.1016/j.yebeh.2015.02.029] [Medline: 25873438]
41. Murray G, O'Rourke C, Hogan J, Fenton JE. Detecting internet search activity for mouth cancer in Ireland. Br J Oral Maxillofac Surg 2016 Feb;54(2):163-165. [doi: 10.1016/j.bjoms.2015.12.005] [Medline: 26774361]
42. Scheres LJJ, Lijfering WM, Middeldorp S, Cannegieter SC. Influence of World Thrombosis Day on digital information seeking on venous thrombosis: a Google Trends study. J Thromb Haemost 2016 Dec;14(12):2325-2328. [doi: 10.1111/jth.13529] [Medline: 27735128]
43. Bragazzi NL, Dini G, Toletone A, Brigo F, Durando P. Leveraging Big Data for Exploring Occupational Diseases-Related Interest at the Level of Scientific Community, Media Coverage and Novel Data Streams: The Example of Silicosis as a Pilot Study. PLoS One 2016;11(11):e0166051 [FREE Full text] [doi: 10.1371/journal.pone.0166051] [Medline: 27806115]
44. Rosenkrantz AB, Prabhu V. Public Interest in Imaging-Based Cancer Screening Examinations in the United States: Analysis Using a Web-Based Search Tool. AJR Am J Roentgenol 2016 Jan;206(1):113-118. [doi: 10.2214/AJR.15.14840] [Medline: 26700342]
45. Schootman M, Toor A, Cavazos-Rehg P, Jeffe DB, McQueen A, Eberth J, et al. The utility of Google Trends data to examine interest in cancer screening. BMJ Open 2015 Jun 08;5(6):e006678 [FREE Full text] [doi: 10.1136/bmjopen-2014-006678] [Medline: 26056120]
46. Linkov F, Bovbjerg DH, Freese KE, Ramanathan R, Eid GM, Gourash W. Bariatric surgery interest around the world: what Google Trends can teach us. Surg Obes Relat Dis 2014 May;10(3):533-538. [doi: 10.1016/j.soard.2013.10.007] [Medline: 24794184]
47. Stein JD, Childers DM, Nan B, Mian SI. Gauging interest of the general public in laser-assisted in situ keratomileusis eye surgery. Cornea 2013 Jul;32(7):1015-1018 [FREE Full text] [doi: 10.1097/ICO.0b013e318283c85a] [Medline: 23538615]
48. Gafson AR, Giovannoni G. CCSVI-A. A call to clinicans and scientists to vocalise in an Internet age. Mult Scler Relat Disord 2014 Mar;3(2):143-146. [doi: 10.1016/j.msard.2013.10.005] [Medline: 25878001]
49. Phelan N, Kelly JC, Moore DP, Kenny P. The effect of the metal-on-metal hip controversy on Internet search activity. Eur J Orthop Surg Traumatol 2014 Oct;24(7):1203-1210. [doi: 10.1007/s00590-013-1399-3] [Medline: 24390041]
50. Simmering JE, Polgreen LA, Polgreen PM. Web search query volume as a measure of pharmaceutical utilization and changes in prescribing patterns. Res Social Adm Pharm 2014;10(6):896-903. [doi: 10.1016/j.sapharm.2014.01.003] [Medline: 24603135]

51. Zhang Z, Zheng X, Zeng DD, Leischow SJ. Tracking Dabbing Using Search Query Surveillance: A Case Study in the United States. J Med Internet Res 2016 Sep 16;18(9):e252 [FREE Full text] [doi: 10.2196/jmir.5802] [Medline: 27637361]
52. Zheluk A, Quinn C, Meylakhs P. Internet search and krokodil in the Russian Federation: an infoveillance study. J Med Internet Res 2014 Sep 18;16(9):e212 [FREE Full text] [doi: 10.2196/jmir.3203] [Medline: 25236385]
53. Kadry B, Chu LF, Kadry B, Gammas D, Macario A. Analysis of 4999 online physician ratings indicates that most patients give physicians a favorable rating. J Med Internet Res 2011 Nov;13(4):e95 [FREE Full text] [doi: 10.2196/jmir.1960] [Medline: 22088924]
54. Huesch MD, Currid-Halkett E, Doctor JN. Public hospital quality report awareness: evidence from National and Californian Internet searches and social media mentions, 2012. BMJ Open 2014 Mar 11;4(3):e004417 [FREE Full text] [doi: 10.1136/bmjopen-2013-004417] [Medline: 24618223]
55. Haney NM, Kinsella SD, Morey JM. United States medical school graduate interest in radiology residency programs as depicted by online search tools. J Am Coll Radiol 2014 Feb;11(2):193-197. [doi: 10.1016/j.jacr.2013.06.023] [Medline: 24120904]
56. Poletto C, Boëlle P, Colizza V. Risk of MERS importation and onward transmission: a systematic review and analysis of cases reported to WHO. BMC Infect Dis 2016 Aug 25;16(1):448 [FREE Full text] [doi: 10.1186/s12879-016-1787-5] [Medline: 27562369]
57. Hossain L, Kam D, Kong F, Wigand RT, Bossomaier T. Social media in Ebola outbreak. Epidemiol Infect 2016 Jul;144(10):2136-2143. [doi: 10.1017/S095026881600039X] [Medline: 26939535]
58. Mavragani A, Ochoa G. The Internet and the Anti-Vaccine Movement: Tracking the 2017 EU Measles Outbreak. BDCC 2018 Jan 16;2(1):2. [doi: 10.3390/bdcc2010002]
59. Bentley RA, Ormerod P. Social versus independent interest in 'bird flu' and 'swine flu'. PLoS Curr 2009 Sep 3;1:RRN1036. [doi: 10.1371/currents.RRN1036]
60. Kostkova P, Fowler D, Wiseman S, Weinberg JR. Major infection events over 5 years: how is media coverage influencing online information needs of health care professionals and the public? J Med Internet Res 2013 Jul 15;15(7):e107 [FREE Full text] [doi: 10.2196/jmir.2146] [Medline: 23856364]
61. Pandey A, Abdullah K, Drazner MH. Impact of Vice President Cheney on public interest in left ventricular assist devices and heart transplantation. Am J Cardiol 2014 May 01;113(9):1529-1531. [doi: 10.1016/j.amjcard.2014.02.007] [Medline: 24630787]
62. Brigo F, Lochner P, Tezzon F, Nardone R. Web search behavior for multiple sclerosis: An infodemiological study. Multiple Sclerosis and Related Disorders 2014 Jul;3(4):440-443. [doi: 10.1016/j.msard.2014.02.005]
63. Bragazzi NL, Watad A, Brigo F, Adawi M, Amital H, Shoenfeld Y. Public health awareness of autoimmune diseases after the death of a celebrity. Clin Rheumatol 2016 Dec 20:1911-1917. [doi: 10.1007/s10067-016-3513-5] [Medline: 28000011]
64. Noar S, Ribisl K, Althouse B, Willoughby J, Ayers J. Using digital surveillance to examine the impact of public figure pancreatic cancer announcements on media and search query outcomes. Journal of the National Cancer Institute - Monographs 2013:188-194.
65. Koburger N, Mergl R, Rummel-Kluge C, Ibelshäuser A, Meise U, Postuvan V, et al. Celebrity suicide on the railway network: Can one case trigger international effects? J Affect Disord 2015 Oct 01;185:38-46. [doi: 10.1016/j.jad.2015.06.037] [Medline: 26143403]
66. Seifter A, Schwarzwalder A, Geis K, Aucott J. The utility of "Google Trends" for epidemiological research: Lyme disease as an example. Geospat Health 2010 May;4(2):135-137. [doi: 10.4081/gh.2010.195] [Medline: 20503183]
67. Rossignol L, Pelat C, Lambert B, Flahault A, Chartier-Kastler E, Hanslik T. A method to assess seasonality of urinary tract infections based on medication sales and google trends. PLoS One 2013;8(10):e76020 [FREE Full text] [doi: 10.1371/journal.pone.0076020] [Medline: 24204587]
68. Harsha AK, Schmitt JE, Stavropoulos SW. Know your market: use of online query tools to quantify trends in patient information-seeking behavior for varicose vein treatment. J Vasc Interv Radiol 2014 Jan;25(1):53-57. [doi: 10.1016/j.jvir.2013.09.015] [Medline: 24286941]
69. Ingram DG, Matthews CK, Plante DT. Seasonal trends in sleep-disordered breathing: evidence from Internet search engine query data. Sleep Breath 2015 Mar;19(1):79-84. [doi: 10.1007/s11325-014-0965-1] [Medline: 24595717]
70. Deiner MS, Lietman TM, McLeod SD, Chodosh J, Porco TC. Surveillance Tools Emerging From Search Engines and Social Media Data for Determining Eye Disease Patterns. JAMA Ophthalmol 2016 Sep 01;134(9):1024-1030 [FREE Full text] [doi: 10.1001/jamaophthalmol.2016.2267] [Medline: 27416554]
71. Zhang Z, Zheng X, Zeng DD, Leischow SJ. Information seeking regarding tobacco and lung cancer: effects of seasonality. PLoS One 2015 Mar;10(3):e0117938 [FREE Full text] [doi: 10.1371/journal.pone.0117938] [Medline: 25781020]
72. Kang M, Song W, Choi S, Kim H, Ha H, Kim S, et al. Google unveils a glimpse of allergic rhinitis in the real world. Allergy 2015 Jan;70(1):124-128. [doi: 10.1111/all.12528] [Medline: 25280183]
73. Schuster N, Rogers M, McMahon JL. Using search engine query data to track pharmaceutical utilization: a study of statins. The American journal of managed care 2010;16(8):215-219.

74. Skeldon SC, Kozhimannil KB, Majumdar SR, Law MR. The effect of competing direct-to-consumer advertising campaigns on the use of drugs for benign prostatic hyperplasia: time series analysis. J Gen Intern Med 2015 Apr;30(4):514-520 [FREE Full text] [doi: 10.1007/s11606-014-3063-y] [Medline: 25338730]
75. Gahr M, Uzelac Z, Zeiss R, Connemann BJ, Lang D, Schönfeldt-Lecuona C. Linking Annual Prescription Volume of Antidepressants to Corresponding Web Search Query Data: A Possible Proxy for Medical Prescription Behavior? J Clin Psychopharmacol 2015 Dec;35(6):681-685. [doi: 10.1097/JCP.0000000000000397] [Medline: 26355849]
76. Mavragani A, Sypsa K, Sampri A, Tsagarakis K. Quantifying the UK Online Interest in Substances of the EU Watchlist for Water Monitoring: Diclofenac, Estradiol, and the Macrolide Antibiotics. Water 2016 Nov 18;8(11):542. [doi: 10.3390/w8110542]
77. Kang M, Zhong H, He J, Rutherford S, Yang F. Using Google Trends for influenza surveillance in South China. PLoS One 2013;8(1):e55205 [FREE Full text] [doi: 10.1371/journal.pone.0055205] [Medline: 23372837]
78. Cho S, Sohn CH, Jo MW, Shin S, Lee JH, Ryoo SM, et al. Correlation between national influenza surveillance data and google trends in South Korea. PLoS One 2013 Dec;8(12):e81422 [FREE Full text] [doi: 10.1371/journal.pone.0081422] [Medline: 24339927]
79. Domnich A, Panatto D, Signori A, Lai PL, Gasparini R, Amicizia D. Age-related differences in the accuracy of web query-based predictions of influenza-like illness. PLoS One 2015;10(5):e0127754 [FREE Full text] [doi: 10.1371/journal.pone.0127754] [Medline: 26011418]
80. Solano P, Ustulin M, Pizzorno E, Vichi M, Pompili M, Serafini G, et al. A Google-based approach for monitoring suicide risk. Psychiatry Res 2016 Dec 30;246:581-586. [doi: 10.1016/j.psychres.2016.10.030] [Medline: 27837725]
81. Arora VS, Stuckler D, McKee M. Tracking search engine queries for suicide in the United Kingdom, 2004-2013. Public Health 2016 Aug;137:147-153. [doi: 10.1016/j.puhe.2015.10.015] [Medline: 26976489]
82. Guernier V, Milinovich GJ, Bezerra SMA, Haworth M, Coleman G, Soares MRJ. Use of big data in the surveillance of veterinary diseases: early detection of tick paralysis in companion animals. Parasit Vectors 2016 Dec 23;9(1):303 [FREE Full text] [doi: 10.1186/s13071-016-1590-6] [Medline: 27215214]
83. Barnes CM, Gunia BC, Wagner DT. Sleep and moral awareness. J Sleep Res 2015 Apr;24(2):181-188 [FREE Full text] [doi: 10.1111/jsr.12231] [Medline: 25159702]
84. Johnson AK, Mehta SD. A comparison of Internet search trends and sexually transmitted infection rates using Google trends. Sex Transm Dis 2014 Jan;41(1):61-63. [doi: 10.1097/OLQ.0000000000000065] [Medline: 24326584]
85. Alicino C, Bragazzi NL, Faccio V, Amicizia D, Panatto D, Gasparini R, et al. Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes. Infect Dis Poverty 2015 Dec 10;4:54 [FREE Full text] [doi: 10.1186/s40249-015-0090-9] [Medline: 26654247]
86. Willson TJ, Lospinoso J, Weitzel E, McMains K. Correlating Regional Aeroallergen Effects on Internet Search Activity. Otolaryngol Head Neck Surg 2014 Dec 12;152(2):228-232. [doi: 10.1177/0194599814560149] [Medline: 25505261]
87. Willson TJ, Shams A, Lospinoso J, Weitzel E, McMains K. Searching for Cedar: Geographic Variation in Single Aeroallergen Shows Dose Response in Internet Search Activity. Otolaryngol Head Neck Surg 2015 Nov 02;153(5):770-774. [doi: 10.1177/0194599815601650] [Medline: 26340925]
88. Zhou X, Ye J, Feng Y. Tuberculosis surveillance by analyzing Google trends. IEEE Trans Biomed Eng 2011 Aug;58(8):2247-2254. [doi: 10.1109/TBME.2011.2132132] [Medline: 21435969]
89. Fond G, Gaman A, Brunel L, Haffen E, Llorca P. Google Trends ® : Ready for real-time suicide prevention or just a Zeta-Jones effect? An exploratory study. Psychiatry Research 2015 Aug;228(3):913-917. [doi: 10.1016/j.psychres.2015.04.022]
90. Gamma A, Schleifer R, Weinmann W, Buadze A, Liebrenz M. Could Google Trends Be Used to Predict Methamphetamine-Related Crime? An Analysis of Search Volume Data in Switzerland, Germany, and Austria. PLoS ONE 2016 Nov 30;11(11):e0166566. [doi: 10.1371/journal.pone.0166566]
91. Pollett S, Wood N, Boscardin WJ, Bengtsson H, Schwarcz S, Harriman K, et al. Validating the Use of Google Trends to Enhance Pertussis Surveillance in California. PLoS Curr 2015 Oct 19:1-10. [doi: 10.1371/currents.outbreaks.7119696b3e7523faa4543faac87c56c2]
92. Wang H, Chen D, Yu H, Chen Y. Forecasting the Incidence of Dementia and Dementia-Related Outpatient Visits With Google Trends: Evidence From Taiwan. J Med Internet Res 2015 Nov 19;17(11):e264 [FREE Full text] [doi: 10.2196/jmir.4516] [Medline: 26586281]
93. Crowson MG, Schulz K, Tucci DL. National Utilization and Forecasting of Ototopical Antibiotics. Otology & Neurotology 2016;37(8):1049-1054. [doi: 10.1097/MAO.0000000000001115]
94. Scopus. URL: https://www.scopus.com/home.uri [accessed 2017-11-08] [WebCite Cache ID 6uotA2G5x]
95. PubMed. accessed 18/4/2018. URL: https://www.ncbi.nlm.nih.gov/pubmed/ [accessed 2018-09-07] [WebCite Cache ID 72Fb7LRKh]
96. Bakker KM, Martinez-Bakker ME, Helm B, Stevenson TJ. Digital epidemiology reveals global childhood disease seasonality and the effects of immunization. PNAS 2016;113(24):6689 [FREE Full text] [doi: 10.1073/pnas.1523941113] [Medline: 27247405]

97. Borron SW, Watts SH, Tull J, Baeza S, Diebold S, Barrow A. Intentional Misuse and Abuse of Loperamide: A New Look at a Drug with "Low Abuse Potential". J Emerg Med 2017 Jul;53(1):73-84. [doi: 10.1016/j.jemermed.2017.03.018] [Medline: 28501383]
98. Bragazzi NL. Infodemiology and infoveillance of multiple sclerosis in Italy. Mult Scler Int 2013;2013:924029 [FREE Full text] [doi: 10.1155/2013/924029] [Medline: 24027636]
99. Bragazzi NL, Dini G, Toletone A, Brigo F, Durando P. Infodemiological data concerning silicosis in the USA in the period 2004-2010 correlating with real-world statistical data. Data Brief 2017 Feb;10:457-464 [FREE Full text] [doi: 10.1016/j.dib.2016.11.021] [Medline: 28054008]
100. Bragazzi NL, Barberis I, Rosselli R, Gianfredi V, Nucci D, Moretti M, et al. How often people google for vaccination: Qualitative and quantitative insights from a systematic search of the web-based activities using Google Trends. Hum Vaccin Immunother 2017 Feb;13(2):464-469. [doi: 10.1080/21645515.2017.1264742] [Medline: 27983896]
101. Bragazzi N, Bacigaluppi S, Robba C, Siri A, Canepa G, Brigo F. Infodemiological data of West-Nile virus disease in Italy in the study period 2004-2015. Data Brief 2016:839-845 [FREE Full text]
102. Bragazzi NL, Bacigaluppi S, Robba C, Nardone R, Trinka E, Brigo F. Infodemiology of status epilepticus: A systematic validation of the Google Trends-based search queries. Epilepsy Behav 2016 Feb;55:120-123. [doi: 10.1016/j.yebeh.2015.12.017] [Medline: 26773681]
103. Bragazzi NL. A Google Trends-based approach for monitoring NSSI. Psychol Res Behav Manag 2013 Dec;7:1-8 [FREE Full text] [doi: 10.2147/PRBM.S44084] [Medline: 24376364]
104. Braun T, Harréus U. Medical nowcasting using Google Trends: application in otolaryngology. Eur Arch Otorhinolaryngol 2013 Jul;270(7):2157-2160. [doi: 10.1007/s00405-013-2532-y] [Medline: 23632877]
105. van Campen JS, van Diessen E, Otte WM, Joels M, Jansen FE, Braun KPJ. Does Saint Nicholas provoke seizures? Hints from Google Trends. Epilepsy Behav 2014 Mar;32:132-134. [doi: 10.1016/j.yebeh.2014.01.019] [Medline: 24548849]
106. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 2009 Nov 15;49(10):1557-1564 [FREE Full text] [doi: 10.1086/630200] [Medline: 19845471]
107. Cavazos-Rehg PA, Krauss MJ, Spitznagel EL, Lowery A, Grucza RA, Chaloupka FJ, et al. Monitoring of non-cigarette tobacco use using Google Trends. Tob Control 2015 May;24(3):249-255 [FREE Full text] [doi: 10.1136/tobaccocontrol-2013-051276] [Medline: 24500269]
108. Cha Y, Stow CA. Mining web-based data to assess public response to environmental events. Environ Pollut 2015 Mar;198:97-99. [doi: 10.1016/j.envpol.2014.12.027] [Medline: 25577650]
109. Chaves JN, Libardi AL, Agostinho-Pesse RS, Morettin M, Alvarenga KDF. Tele-health: assessment of websites on newborn hearing screening in Portuguese Language. Codas 2015 Dec;27(6):526-533 [FREE Full text] [doi: 10.1590/2317-1782/20152014169] [Medline: 26691616]
110. Davis NF, Gnanappiragasam S, Thornhill JA. Interstitial cystitis/painful bladder syndrome: the influence of modern diagnostic criteria on epidemiology and on Internet search activity by the public. Transl Androl Urol 2015 Oct;4(5):506-511 [FREE Full text] [doi: 10.3978/j.issn.2223-4683.2015.06.08] [Medline: 26816850]
111. Fazeli DS, Carlos RC, Hall KS, Dalton VK. Novel data sources for women's health research: mapping breast screening online information seeking through Google trends. Acad Radiol 2014 Sep;21(9):1172-1176 [FREE Full text] [doi: 10.1016/j.acra.2014.05.005] [Medline: 24998689]
112. DeVilbiss E, Lee B. Brief Report: Trends in U.S. National Autism Awareness from 2004 to 2014: The Impact of National Autism Awareness Month. Journal of Autism and Developmental Disorders 2014;44(12):3271-3273. [doi: 10.1007/s10803-014-2160-4] [Medline: 24915931]
113. El-Sheikha J. Global search demand for varicose vein information on the internet. Phlebology 2015 Sep;30(8):533-540. [doi: 10.1177/0268355514542681] [Medline: 24993972]
114. Fenichel EP, Kuminoff NV, Chowell G. Skip the trip: air travelers' behavioral responses to pandemic influenza. PLoS One 2013 Mar;8(3):e58249 [FREE Full text] [doi: 10.1371/journal.pone.0058249] [Medline: 23526970]
115. Foroughi F, Lam AK, Lim MS, Saremi N, Ahmadvand A. “Googling” for Cancer: An Infodemiological Assessment of Online Search Interests in Australia, Canada, New Zealand, the United Kingdom, and the United States. JMIR Cancer 2016 May 04;2(1):e5 [FREE Full text] [doi: 10.2196/cancer.5212] [Medline: 28410185]
116. Garrison SR, Dormuth CR, Morrow RL, Carney GA, Khan KM. Seasonal effects on the occurrence of nocturnal leg cramps: a prospective cohort study. CMAJ 2015 Mar 03;187(4):248-253 [FREE Full text] [doi: 10.1503/cmaj.140497] [Medline: 25623650]
117. Gollust SE, Qin X, Wilcock AD, Baum LM, Barry CL, Niederdeppe J, et al. Search and You Shall Find: Geographic Characteristics Associated With Google Searches During the Affordable Care Act's First Enrollment Period. Med Care Res Rev 2017 Dec;74(6):723-735. [doi: 10.1177/1077558716660944] [Medline: 27457426]
118. Harorli OT, Harorli H. Evaluation of internet search trends of some common oral problems, 2004 to 2014. Community Dental Health 2014;31(3):188-192. [doi: 10.1922/CDH_3330Harorl?05]
119. Harsha AK, Schmitt JE, Stavropoulos SW. Match day: online search trends reflect growing interest in IR training. J Vasc Interv Radiol 2015 Jan;26(1):95-100. [doi: 10.1016/j.jvir.2014.09.011] [Medline: 25541447]

120. Hassid BG, Day LW, Awad MA, Sewell JL, Osterberg EC, Breyer BN. Using Search Engine Query Data to Explore the Epidemiology of Common Gastrointestinal Symptoms. Dig Dis Sci 2017 Dec;62(3):588-592. [doi: 10.1007/s10620-016-4384-y] [Medline: 27878646]
121. Huang J, Zheng R, Emery S. Assessing the impact of the national smoking ban in indoor public places in china: evidence from quit smoking related online searches. PLoS One 2013 Jun;8(6):e65577 [FREE Full text] [doi: 10.1371/journal.pone.0065577] [Medline: 23776504]
122. Ingram DG, Plante DT. Seasonal trends in restless legs symptomatology: evidence from Internet search query data. Sleep Med 2013 Dec;14(12):1364-1368. [doi: 10.1016/j.sleep.2013.06.016] [Medline: 24152798]
123. Jha S, Wang Z, Laucis N, Bhattacharyya T. Trends in Media Reports, Oral Bisphosphonate Prescriptions, and Hip Fractures 1996-2012: An Ecological Analysis. J Bone Miner Res 2015 Dec;30(12):2179-2187 [FREE Full text] [doi: 10.1002/jbmr.2565] [Medline: 26018247]
124. Lawson MAC, Lawson MA, Kalff R, Walter J. Google Search Queries About Neurosurgical Topics: Are They a Suitable Guide for Neurosurgeons? World Neurosurg 2016 Jun;90:179-185. [doi: 10.1016/j.wneu.2016.02.045] [Medline: 26898496]
125. Leffler CT, Davenport B, Chan D. Frequency and seasonal variation of ophthalmology-related internet searches. Can J Ophthalmol 2010 Jun;45(3):274-279. [doi: 10.3129/i10-022] [Medline: 20436544]
126. Ling R, Lee J. Disease Monitoring and Health Campaign Evaluation Using Google Search Activities for HIV and AIDS, Stroke, Colorectal Cancer, and Marijuana Use in Canada: A Retrospective Observational Study. JMIR Public Health Surveill 2016 Oct 12;2(2):e156 [FREE Full text] [doi: 10.2196/publichealth.6504] [Medline: 27733330]
127. Liu F, Allan GM, Korownyk C, Kolber M, Flook N, Sternberg H, et al. Seasonality of Ankle Swelling: Population Symptom Reporting Using Google Trends. Ann Fam Med 2016 Dec;14(4):356-358 [FREE Full text] [doi: 10.1370/afm.1953] [Medline: 27401424]
128. Luckett T, Disler R, Hosie A, Johnson M, Davidson P, Currow D, et al. Content and quality of websites supporting self-management of chronic breathlessness in advanced illness: a systematic review. NPJ Prim Care Respir Med 2016 Dec 26;26:16025 [FREE Full text] [doi: 10.1038/npjpcrm.2016.25] [Medline: 27225898]
129. Majumder MS, Santillana M, Mekaru SR, McGinnis DP, Khan K, Brownstein JS. 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 2016 Jun 01;2(1):e30 [FREE Full text] [doi: 10.2196/publichealth.5814] [Medline: 27251981]
130. Mattin MJ, Solano-Gallego L, Dhollander S, Afonso A, Brodbelt DC. The frequency and distribution of canine leishmaniosis diagnosed by veterinary practitioners in Europe. Vet J 2014 Jun;200(3):410-419. [doi: 10.1016/j.tvjl.2014.03.033] [Medline: 24767097]
131. Myers L, Jones J, Boesten N, Lancman M. Psychogenic non-epileptic seizures (PNES) on the Internet: Online representation of the disorder and frequency of search terms. Seizure 2016 Aug 01;40:114-122 [FREE Full text]
132. Parker J, Cuthbertson C, Loveridge S, Skidmore M, Dyar W. Forecasting state-level premature deaths from alcohol, drugs, and suicides using Google Trends data. J Affect Disord 2017 Dec 15;213:9-15. [doi: 10.1016/j.jad.2016.10.038] [Medline: 28171770]
133. Phelan N, Davy S, O'Keeffe GW, Barry DS. Googling in anatomy education: Can google trends inform educators of national online search patterns of anatomical syllabi? Anat Sci Educ 2017 Mar;10(2):152-159. [doi: 10.1002/ase.1641] [Medline: 27547967]
134. Plante DT, Ingram DG. Seasonal trends in tinnitus symptomatology: evidence from Internet search engine query data. Eur Arch Otorhinolaryngol 2015 Oct;272(10):2807-2813. [doi: 10.1007/s00405-014-3287-9] [Medline: 25234771]
135. Rohart F, Milinovich GJ, Avril SMR, Lê Cao KA, Tong S, Hu W. Disease surveillance based on Internet-based linear models: an Australian case study of previously unmodeled infection diseases. Sci Rep 2016 Dec 20;6:38522 [FREE Full text] [doi: 10.1038/srep38522] [Medline: 27994231]
136. Scatà M, Di Stefano A, Liò P, La Corte A. The Impact of Heterogeneity and Awareness in Modeling Epidemic Spreading on Multiplex Networks. Sci Rep 2016 Dec 16;6:37105 [FREE Full text] [doi: 10.1038/srep37105] [Medline: 27848978]
137. Shin S, Seo D, An J, Kwak H, Kim S, Gwack J, et al. High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. Sci Rep 2016 Sep 06;6:32920 [FREE Full text] [doi: 10.1038/srep32920] [Medline: 27595921]
138. Sentana-Lledo D, Barbu CM, Ngo MN, Wu Y, Sethuraman K, Levy MZ. Seasons, Searches, and Intentions: What The Internet Can Tell Us About The Bed Bug (Hemiptera: Cimicidae) Epidemic. J Med Entomol 2016 Jan;53(1):116-121. [doi: 10.1093/jme/tjv158] [Medline: 26474879]
139. Takada K. Japanese Interest in "Hotaru" (Fireflies) and "Kabuto-Mushi" (Japanese Rhinoceros Beetles) Corresponds with Seasonality in Visible Abundance. Insects 2012 Apr 10;3(2):424-431 [FREE Full text] [doi: 10.3390/insects3020424] [Medline: 26466535]
140. Telfer S, Woodburn J. Let me Google that for you: a time series analysis of seasonality in internet search trends for terms related to foot and ankle pain. J Foot Ankle Res 2015 Jul;8:27 [FREE Full text] [doi: 10.1186/s13047-015-0074-9] [Medline: 26146521]

141. Troelstra SA, Bosdriesz JR, de Boer MR, Kunst AE. Effect of Tobacco Control Policies on Information Seeking for Smoking Cessation in the Netherlands: A Google Trends Study. PLoS One 2016 Feb;11(2):e0148489 [FREE Full text] [doi: 10.1371/journal.pone.0148489] [Medline: 26849567]
142. Toosi B, Kalia S. Seasonal and Geographic Patterns in Tanning Using Real-Time Data From Google Trends. JAMA Dermatol 2016 Feb;152(2):215-217. [doi: 10.1001/jamadermatol.2015.3008] [Medline: 26719968]
143. Warren KE, Wen LS. Measles, social media and surveillance in Baltimore City. J Public Health (Oxf) 2017 Sep 01;39(3):e73-e78. [doi: 10.1093/pubmed/fdw076] [Medline: 27521926]
144. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. PNAS 2015;112(47):14473 [FREE Full text] [doi: 10.1073/pnas.1515373112] [Medline: 26553980]
145. Köhler MJ, Springer S, Kaatz M. [On the seasonality of dermatoses: a retrospective analysis of search engine query data depending on the season]. Hautarzt 2014 Sep 14;65(9):814-822. [doi: 10.1007/s00105-014-2848-6] [Medline: 25234631]
146. Orellano PW, Reynoso JI, Antman J, Argibay O. Uso de la herramienta Google Trends para estimar la incidencia de enfermedades tipo influenza en Argentina. Cad. Saúde Pública 2015 Apr;31(4):691-700. [doi: 10.1590/0102-311X00072814]
147. Cjuno J, Taype-Rondan A. Estacionalidad de la cefalea en el hemisferio norte y el hemisferio sur: una aproximación utilizando Google Trends. Rev. méd. Chile 2016 Jul;144(7):947-947. [doi: 10.4067/S0034-98872016000700019]
148. Tejada-Llacsa P. Gaceta Sanitaria. 2016. ¿Qué se busca sobre el aborto en Internet? Una evaluación con Google Trends en Perú URL: https://linkinghub.elsevier.com/retrieve/pii/S0213911116300486 [accessed 2018-09-07] [WebCite Cache ID 72FbxshJz]
149. Yang Y, Zeng Q, Zhao H, Yi J, Li Q, Xia Y. Hepatitis B prediction model based on Google trends. Journal of Shanghai Jiaotong University (Medical Science) 2013;33(2):204-208. [doi: 10.3969/j.issn.1674-8115.2013.02.016]

Abbreviations

ANOVA: analysis of variance
ARIMA: autoregressive integrated moving average.
MS: multiple sclerosis

Edited by G Eysenbach; submitted 08.11.17; peer-reviewed by A Benis, J Bian, C Fincham; comments to author 15.03.18; revised version received 07.05.18; accepted 21.06.18; published 06.11.18

Please cite as:
Mavragani A, Ochoa G, Tsagarakis KP
Assessing the Methods, Tools, and Statistical Approaches in Google Trends Research: Systematic Review
J Med Internet Res 2018;20(11):e270
URL: https://www.jmir.org/2018/11/e270/
doi:10.2196/jmir.9366
PMID:

Tutorial

Google Trends in Infodemiology and Infoveillance: Methodology Framework

Amaryllis Mavragani, BSc, MSc; Gabriela Ochoa, BSc, MSc, PhD

Department of Computing Science and Mathematics, Faculty of Natural Sciences, University of Stirling, Stirling, United Kingdom

Corresponding Author:
Amaryllis Mavragani, BSc, MSc
Department of Computing Science and Mathematics
Faculty of Natural Sciences
University of Stirling
University Campus
Stirling, FK94LA
United Kingdom
Phone: 44 7523782711
Email: amaryllis.mavragani1@stir.ac.uk

Abstract

Internet data are being increasingly integrated into health informatics research and are becoming a useful tool for exploring human behavior. The most popular tool for examining online behavior is Google Trends, an open tool that provides information on trends and the variations of online interest in selected keywords and topics over time. Online search traffic data from Google have been shown to be useful in analyzing human behavior toward health topics and in predicting disease occurrence and outbreaks. Despite the large number of Google Trends studies during the last decade, the literature on the subject lacks a specific methodology framework. This article aims at providing an overview of the tool and data and at presenting the first methodology framework in using Google Trends in infodemiology and infoveillance, including the main factors that need to be taken into account for a strong methodology base. We provide a step-by-step guide for the methodology that needs to be followed when using Google Trends and the essential aspects required for valid results in this line of research. At first, an overview of the tool and the data are presented, followed by an analysis of the key methodological points for ensuring the validity of the results, which include selecting the appropriate keyword(s), region(s), period, and category. Overall, this article presents and analyzes the key points that need to be considered to achieve a strong methodological basis for using Google Trends data, which is crucial for ensuring the value and validity of the results, as the analysis of online queries is extensively integrated in health research in the big data era.

(JMIR Public Health Surveill 2019;5(2):e13439) doi:10.2196/13439

KEYWORDS

big data; health; infodemiology; infoveillance; internet behavior; Google Trends

Introduction

The use of internet data has become an integral part of health informatics over the past decade, with online sources becoming increasingly available and providing data that can be useful in analyzing and predicting human behavior. This use of the internet has formed two new concepts: "Infodemiology," first defined by Eysenbach as "the science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy" [1], and "Infoveillance," defined as "the longitudinal tracking of infodemiology metrics for surveillance and trend analysis" [2].

The main limitation of validating this line of research is the general lack of openness and availability of official health data. Data collection and analysis of official health data on disease occurrence and prevalence involve several health officials and can even take years until the relevant data are available. This means that data cannot be accessed in real time, which is crucial in health assessment. In several countries, official health data are not publicly available, and even in countries where data are available, they usually consist of large time-interval data (eg, annual data), which makes the analysis and forecasting of diseases and outbreaks more difficult.

Nevertheless, data from several online sources are being widely used to monitor disease outbreaks and occurrence, mainly from Google [3-7] and social media [8-12]. Twitter has become increasingly popular over the past few years [13-19], while

several other studies have combined data from different online sources such as Facebook and Twitter [20] or Google, Twitter, and electronic health records [21].

Currently, the most popular tool in addressing health issues and topics with the use of internet data is Google Trends [22], an open online tool that provides both real-time and archived information on Google queries from 2004 on. The main advantage of Google Trends is that it uses the revealed and not stated users' preferences [23]; therefore, we can obtain information that would be otherwise difficult or impossible to collect. In addition, as data are available in real time, it solves issues that arise with traditional, time-consuming survey methods. Another advantage is that, as Web searches are performed anonymously, it enables the analysis and forecasting of sensitive diseases and topics, such as AIDS [24], mental illnesses and suicide [25-27], and illegal drugs [28,29].

Despite the limitations of data from traditional sources and owing to the fact that online data have shown to be valuable in predictions, the combination of traditional data and Web-based data should be explored, as the results could provide valid and interesting results. Over the past few years, the diversity of online sources used in addressing infodemiology topics is increasing. Indicative recent publications of online sources and combinations of sources are presented in Table 1.

As discussed above, many studies have used Google Trends data to analyze online behavior toward health topics and to forecast prevalence of diseases. However, the literature lacks a methodology framework that provides a concise overview and detailed guidance for future researchers. We believe such a framework is imperative, as the analysis of online data is based on empirical relationships, and thus, a solid methodological basis of any Google Trends study is crucial for ensuring the value and validity of the results.

Table 1. Recent indicative infodemiology studies.

Author(s)	Keywords	Google Trends	Twitter	Facebook	Other social media (eg, YouTube)	Blogs, forums, news outlets, Wikipedia	Databases, electronic health records	Other search engines (Baidu)
Abdellaoui et al [30]	Drug treatment					✓		
Allen et al [31]	Tobacco waterpipe		✓					
Berlinger et al [32]	Herpes, Vaccination	✓						
Bragazzi and Mahroum [33]	Plague, Madagascar	✓						
Chen et al [18]	Zika epidemic		✓					
Forounghi et al [34]	Cancer	✓						
Gianfredi et al [35]	Pertussis	✓						
Hswen et al [36]	Psychological analysis, Autism		✓					
Jones et al [37]	Cancer					✓		
Kandula et al [38]	Influenza	✓						
Keller et al [39]	Bowel disease, Pregnancy, Medication			✓	✓	✓		
Mavragani et al [7]	Asthma	✓						
Mejova et al [40]	Health monitoring			✓				
Odlum et al [41]	HIV/AIDS		✓					
Phillips et al [42]	Cancer	✓						
Poirier et al [43]	Influenza, Hospitals	✓					✓	
Radin et al [44]	Systematic Lupus Erythematous	✓						
Roccetti et al [20]	Crohn's disease		✓	✓				
Tana et al [25]	Depression, Finland	✓						
Vsconcellos-Silva et al [45]	Cancer	✓						
Wakamiya et al [46]	Influenza		✓					
Wang et al [47]	Obesity	✓						
Watad et al [48]	West Nile Virus	✓			✓	✓		
Xu et al [49]	Cancer, China							✓

We proceed in a step-by-step manner to develop the methodology framework that should be followed when using Google Trends in infodemiology. First, we provide an overview of how the data are retrieved and adjusted along with the available features, followed by the methodology framework for choosing the appropriate keyword(s), region(s), period, and category. Finally, the results are discussed, along with the limitations of the tool and suggestions for future research.

Methodology Framework

Data Overview

Google Trends is an open online tool that provides information on what was and is trending, based on actual users' Google queries. It offers a variety of choices, such as Trending Searches, Year in Search, and Explore. Table 2 describes the features offered by Google Trends and their respective descriptions.

When using Google Trends for research, data are retrieved from the "Explore" feature, which allows download of real-time data from the last week and archived data for specific keywords and topics from January 2004 up to 36 hours before the search is conducted. The data are retrieved directly from the Google Trends Explore page in .csv format after the examined keyword(s) is entered and the region, period, and category are selected. By default, the period is set to "Worldwide," the time frame is set to "past 12 months," and the category is set to "All categories."

The data are normalized over the selected time frame, and the adjustment is reported by Google as follows:

> *Search results are proportionate to the time and location of a query by the following process: Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity. Otherwise, places with the most search volume would always be ranked highest. The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics. Different regions that show the same search interest for a term don't always have the same total search volumes [50]*

The normalization of data indicates that the values vary from 0 to 100. The value 0 does not necessarily indicate no searches, but rather indicates very low search volumes that are not included in the results. The adjustment process also excludes queries that are made over a short time frame from the same internet protocol address and queries that contain special characters. Google does not have a filter for controversial topics, but it excludes related search terms that are sexual. However, it allows retrieval of queries' normalized hits for any keyword entered, independent of filters.

Google Trends allows one to explore the online interest in one term or the comparison of the online interest for up to five terms. It allows a variety of combinations to compare different terms and regions as follows:

- For one term in one region over a specific period, such as for "Asthma" in the United States from January 2004 to December 2014 (Figure 1a)
- For the same term in different regions over the same period, such as for "Tuberculosis" in the United States and United Kingdom from March 24, 2007, to April 7, 2011 (Figure 1b)
- For different terms (up to five) in the same region for the same period, such as for the terms "Chlamydia," "Tuberculosis," and "Syphilis" in Australia from October 5, 2012, to December 18, 2012 (Figure 1c)
- For different terms (up to five) for different regions over the same period, such as comparing the term "Asthma" in the United States, "AIDS" in the United Kingdom, and "Measles" in Canada from June 1, 2017, to July 15, 2018 (Figure 1d)

When the term(s), region(s), period(s), and category are defined, the outputs are a graph of the variations of all examined terms in the online interest over the selected time frame (Figure 1) and their respective heat maps, which are presented separately for all examined regions (Figure 2); all datasets can be downloaded in .csv format.

Apart from the graph, the .csv with the relative search volumes, and the interest heat maps, Google Trends also shows and allows one to download .csv files of (1) the "Top related queries", defined as "Top searches are terms that are most frequently searched with the term you entered in the same search session, within the chosen category, country, or region" (Figure 3a); (2) the "Rising related queries", defined as "terms that were searched for with the keyword you entered...which had the most significant growth in volume in the requested time period" (Figure 3b); (3) the "Top Related Topics" (Figure 3c); and (4) the "Rising Related Topics" (Figure 3d).

Table 2. Google Trends Features and Descriptions.

Feature	Description
Homepage	Provides an overview of what is searched for in a selected region (default: United States)
Explore	Allows exploration of the online interest for specific keywords over selected periods and regions (default: worldwide, 12 months)
Trending Searches	Shows the trending queries for (1) daily search trends and (2) real-time search trends in a selected region (default: United States)
Year in Searches	Show what was trending in a specific region in a specific year (default: United States, previous year)
Subscriptions	Allows subscription for (1) a specific topic in a specific region and sends updates for noteworthy events (via email either once a week or once a month) and (2) trending searches and sends updates about trending searches (via email either as it happens, or once a day, or once a week and includes either "Top Daily Searches," "Majority of Daily Search Trends," or "All Daily Search Trends")

Figure 1. Graphs of the variations in the online interest for the examined terms over the selected time frame in Google Trends.

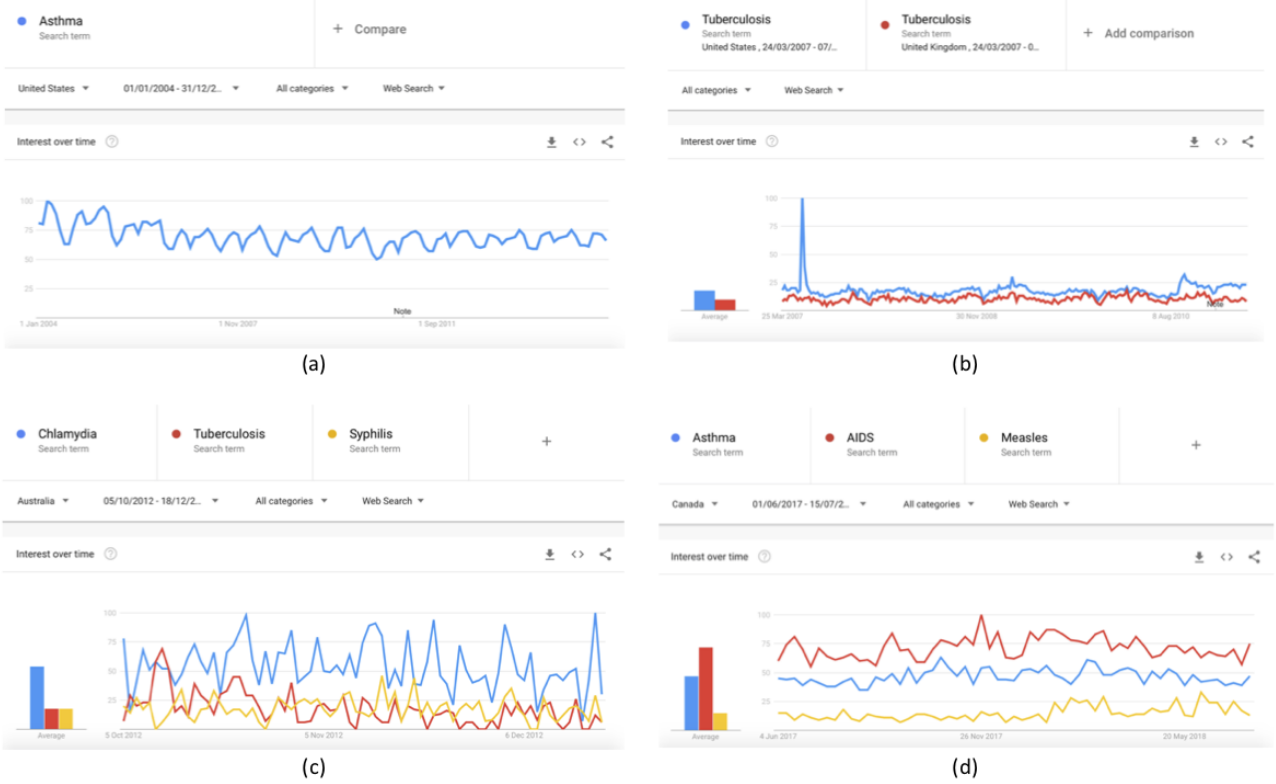


Figure 2. Heat map for (a) “Asthma” in the United States from Jan 2004 to Dec 2014; (b) “Tuberculosis” in the United States and United Kingdom from March 24, 2007, to April 7, 2011; (c) “Chlamydia,” “Tuberculosis,” and “Syphilis” in Australia from Oct 5, 2012, to Dec 18, 2012; (d) “Asthma” in the United States, “AIDS” in the United Kingdom, and “Measles” in Canada from June 1, 2017, to July 15, 2018.

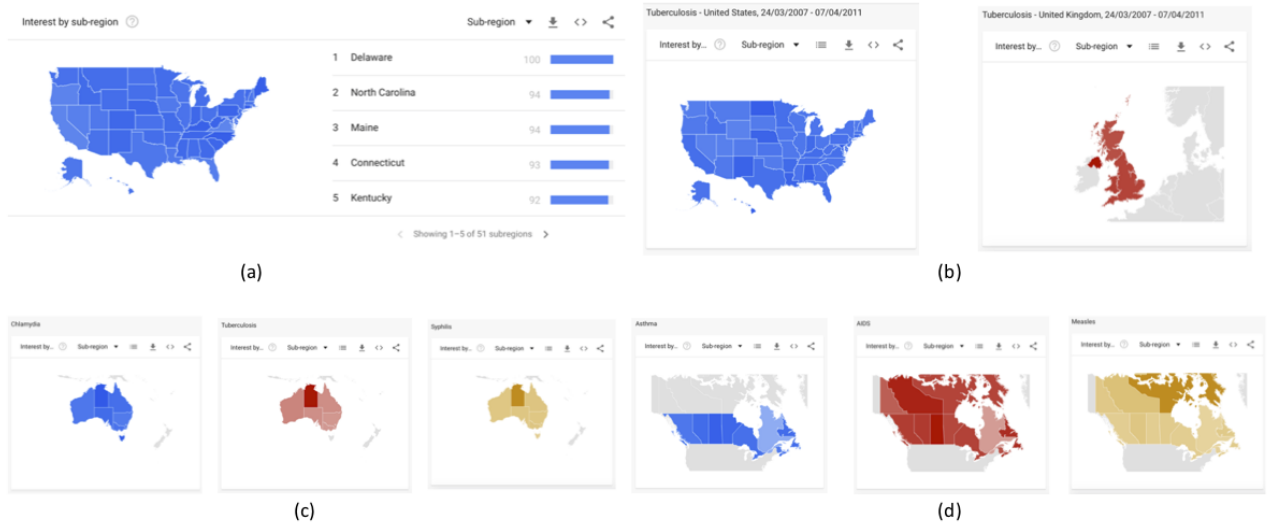
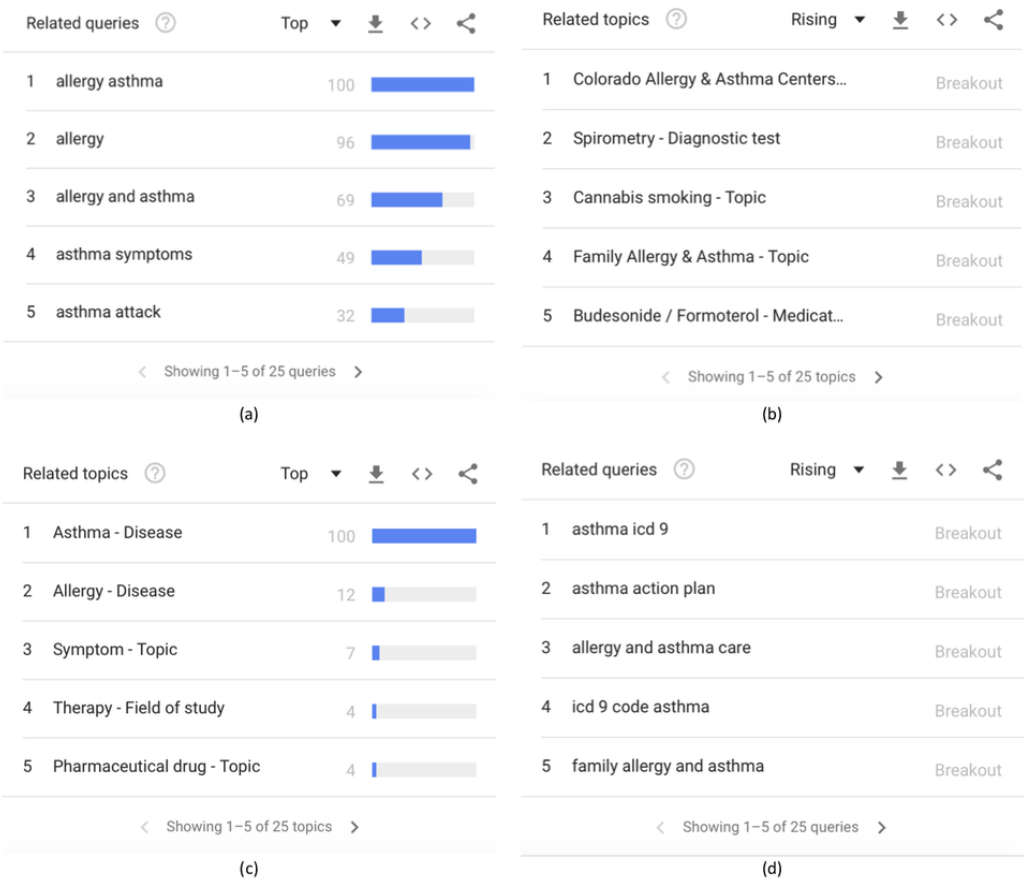


Figure 3. Google Trends' (a) top related queries, (b) rising related topics, (c) top related topics, and (d) rising related queries for "Asthma" in the United States from Jan 1, 2004, to Dec 31, 2014.



Keyword Selection

The selection of the correct keyword(s) when examining online queries is key for valid results [51]. Thus, many factors should be taken into consideration when using Google Trends data in order to ensure a valid analysis.

Google Trends is not case sensitive, but it takes into account accents, plural or singular forms, and spelling mistakes. Therefore, whatever the choice of keywords or combination of keywords, parts of the respective queries will not be considered for further analysis.

To partly overcome this limitation, the "+" feature can be used to include the most commonly encountered misspellings, which are selected and entered manually; however, we should keep in mind that some results will always be missing, as all possible spelling variations cannot be included. In addition, incorrect spellings of some words could be used even more often than the correct one, in which case, the analysis will not be trivial. However, in most of the cases, the correct spelling is the most commonly used, and therefore, the analysis can proceed as usual. For example, gonorrhea is often misspelled, mainly as "Gonorrea," which is also the Spanish term for the disease. As depicted in Figure 4a, both terms have significantly high volumes. Therefore, to include more results, both terms could be entered as the search term by using the "+" feature (Figure 4b). In this way, all results including the correct and the incorrect spellings are aggregated in the results. Note that this is not limited to only two terms; the "+" feature can be used for multiple keywords or for results in multiple languages in a region.

In the case of accents, before choosing the keywords to be examined, the variations in interest between the terms with and those without accents and special characters should be explored. For example, measles translates into "Sarampión," "ošpice," "mässling," and "Ιλαρά" in Spanish, Slovenian, Swedish, and Greek, respectively. As depicted in Figure 5, in Spanish and Greek, the term without the accent is searched for in higher volumes; in Slovenian, the term with the accent is mostly used; and in Swedish, the term without the accent is almost nonexistent. Thus, in Greek searches, the term without accent should be selected, in Slovenian and Swedish searches, terms with accents should be used, while for Spanish, as both terms yield significant results, either both terms using the "+" feature or the term without the accent should be selected.

Another important aspect is the use of quotation marks when selecting the keyword. This obviously applies only to keywords with two or more words. For example, breast cancer can be searched online by using or not using quotes. To elaborate, the term "breast cancer" without quotes will yield results that include the words "breast" and "cancer" in any possible combination and order; for example, keywords "breast cancer screening" and "breast and colon cancer" are both included in the results. However, when using quotes, the term "breast cancer" is included as is; for example, "breast cancer screening," "living with breast cancer," and "breast cancer patient." As shown in Figure 6a, the results are almost identical in this case.

However, this is not always the case. As depicted in Figure 6b, this is clearly different for "HIV test." When searching for HIV test with and without quotes, the results differ in volumes of searches, despite the trend being very similar but not exactly the same.

Finally, when researching with Google Trends, the options of "search term" and "disease" (or "topic") are available when entering a keyword. Although the "search term" gives results for all keywords that include the selected term, "disease" includes various keywords that fall within the category, or, as Google describes it, "topics are a group of terms that share the same concept in any language."

Therefore, it is imperative that keyword selection is conducted with caution and that the available options and features are carefully explored and analyzed. This will ensure validity of the results.

Figure 4. Use of the "+" feature for including misspelled terms for (a) "Gonorrhea" compared to "Gonorrea"; (b) both terms by using the "+" feature.

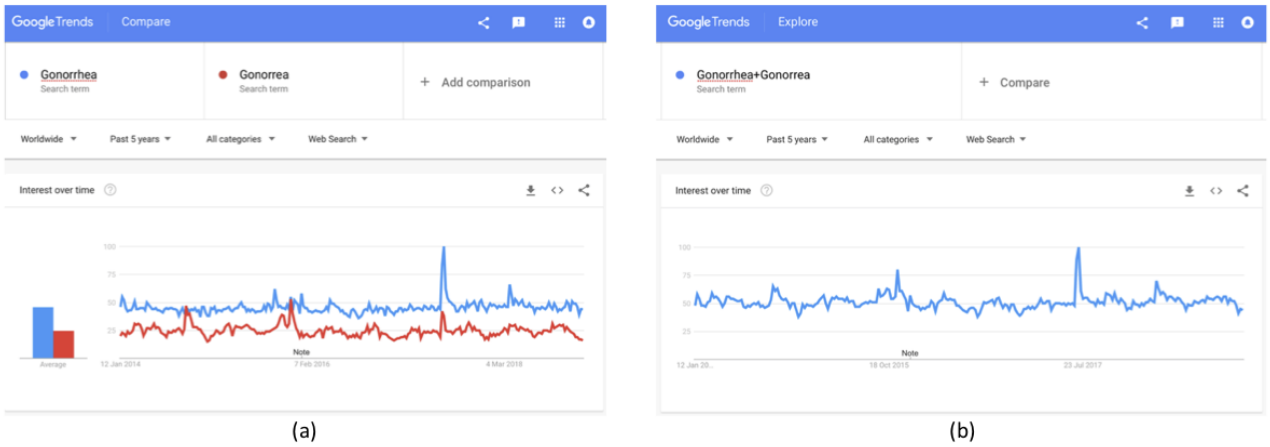


Figure 5. Selection of the correct keyword for measles based on the use of accents in the respective translated terms in (a) Spanish, (b) Slovenian, (c) Swedish, and (d) Greek.

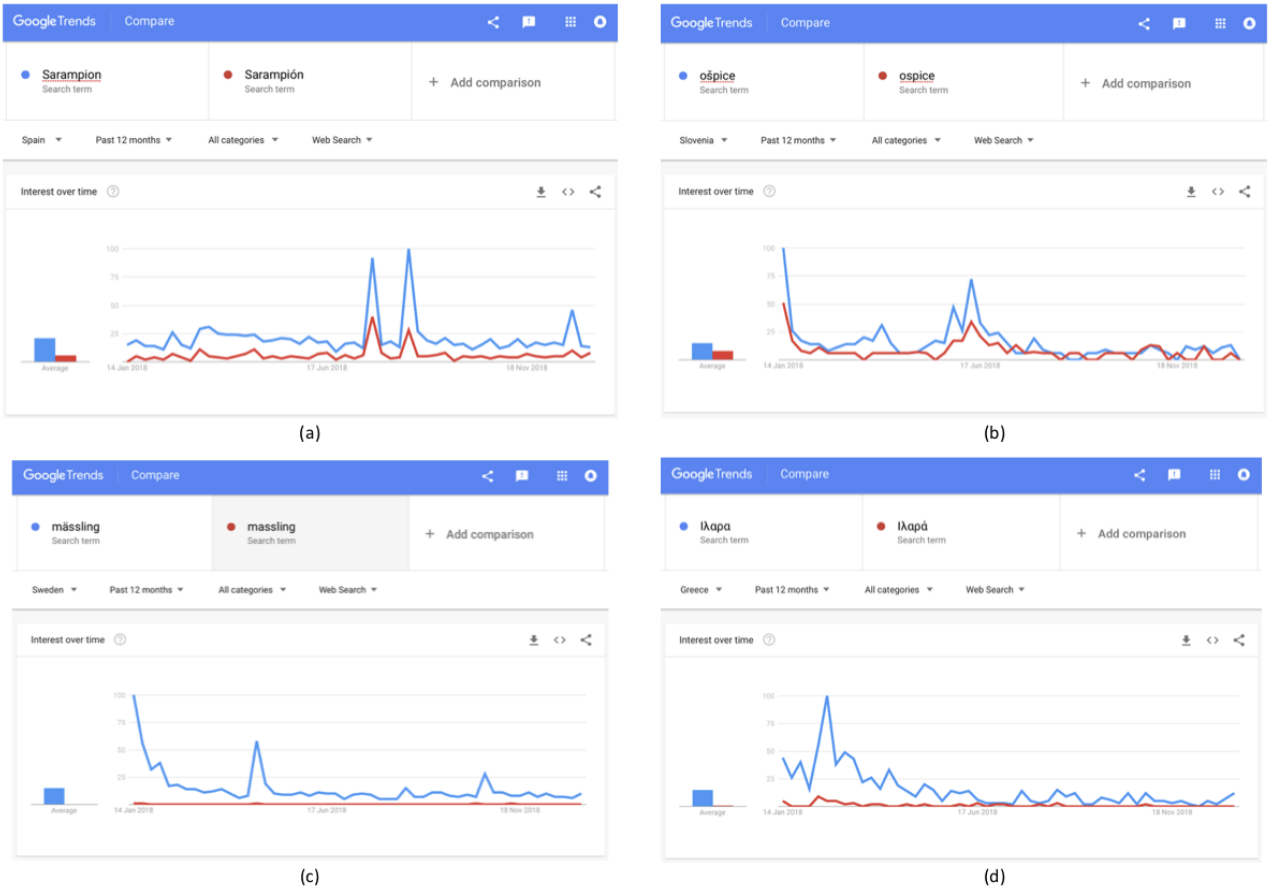
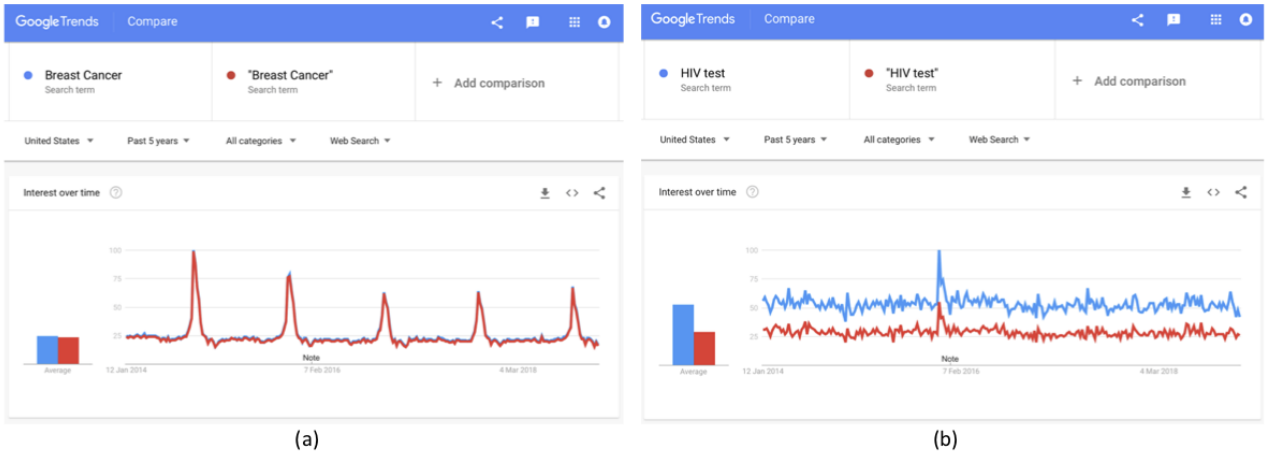


Figure 6. Differences in results with and without quotation marks for (a) "Breast Cancer" and (b) "HIV test.".



Region Selection

The next step is to select the geographical region for which query data are retrieved. The first level of categorization allows data download for the online interest of one or more terms worldwide or by country. The list available includes all countries, in most of which interest in smaller regions can be explored.

For example, in the United States, it is possible to compare results even at metropolitan and city levels. Figure 7a shows the regional online interest in the term "Flu" worldwide, where the United States is the country with the highest online interest in the examined term, followed by the rest of the 33 countries in which the examined term is most popular. Figure 7b shows the heat map of the interest by state in the United States in the term "Flu" over the past 5 years; either as a new independent search or by clicking on the country "USA" in the worldwide map. As shown in the right bottom corner of Figure 7, Google Trends provides the relative interest for all 50 US states plus Washington DC.

In the case of the United States, it is possible to examine the online interest by metropolitan area, as depicted in Figure 8 with the examples of California, Texas, New York, and Florida. The option for examining the online interest at the metropolitan level is not available for all countries, where from the state (or county) level, the interest changes directly to the city level. This includes fewer cities than regions with available metropolitan area data, as, for example, in countries with very large populations like India (Figure 9e) or with smaller populations like Greece (Figure 9f).

Figure 9 depicts the online interest by city in the selected metropolitan areas of Los Angeles in California, Dallas in Texas, New York in New York, and Miami in Florida.

At metropolitan level, by selecting the "include low search volume regions," the total of the included cities is 123 in Los Angeles, 67 in Texas, 110 New York, and 50 in Miami, while in India and Greece, the number of cities remains 7 and 2, respectively.

Figure 7. Online interest in the term "Flu" over the past 5 years (a) worldwide and (b) in the United States.

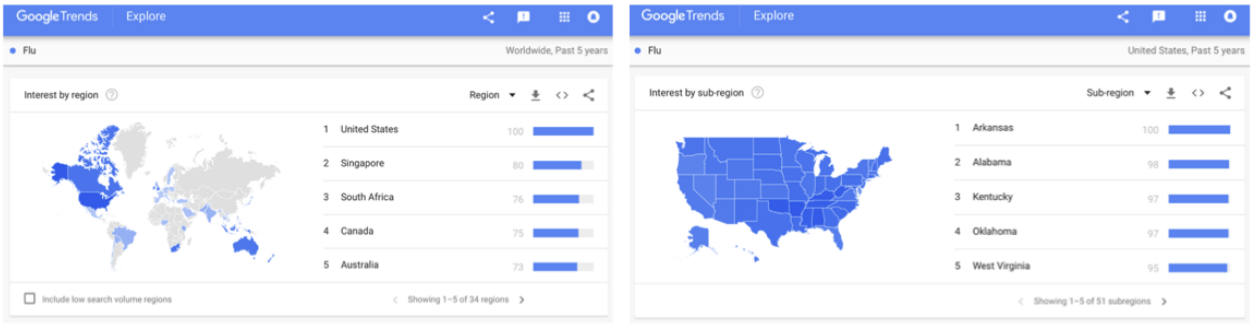


Figure 8. Regional online interest in the term "Flu" at metropolitan level over the past 5 years in (a) California, (b) Texas, (c) New York, and (d) Florida.

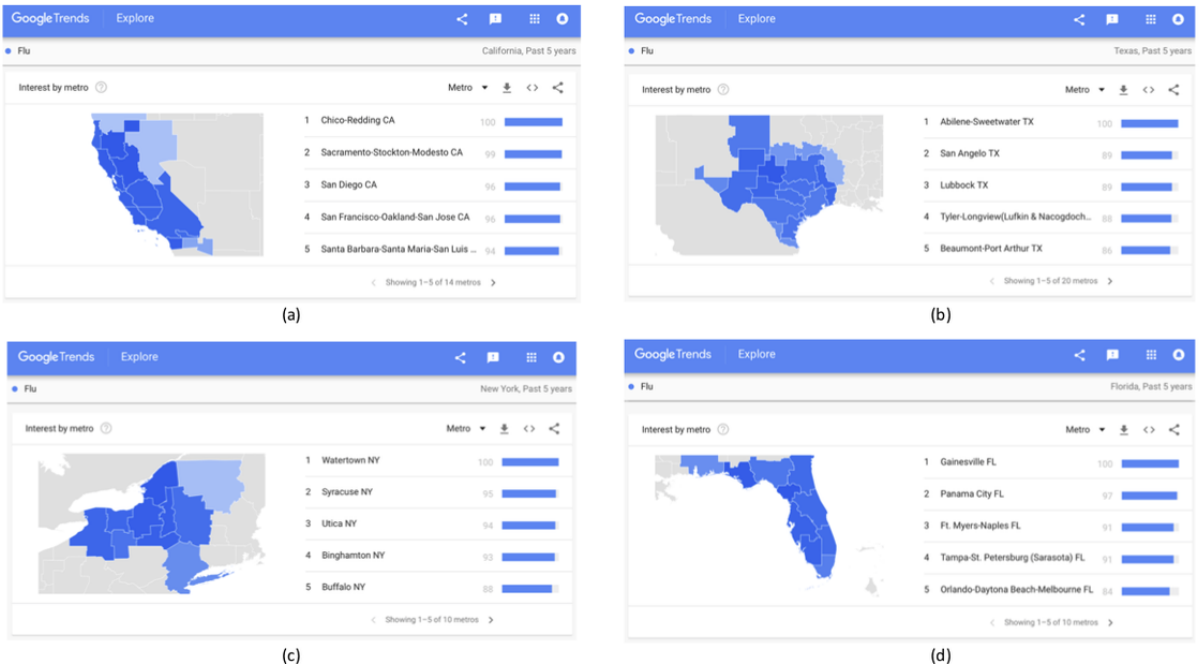
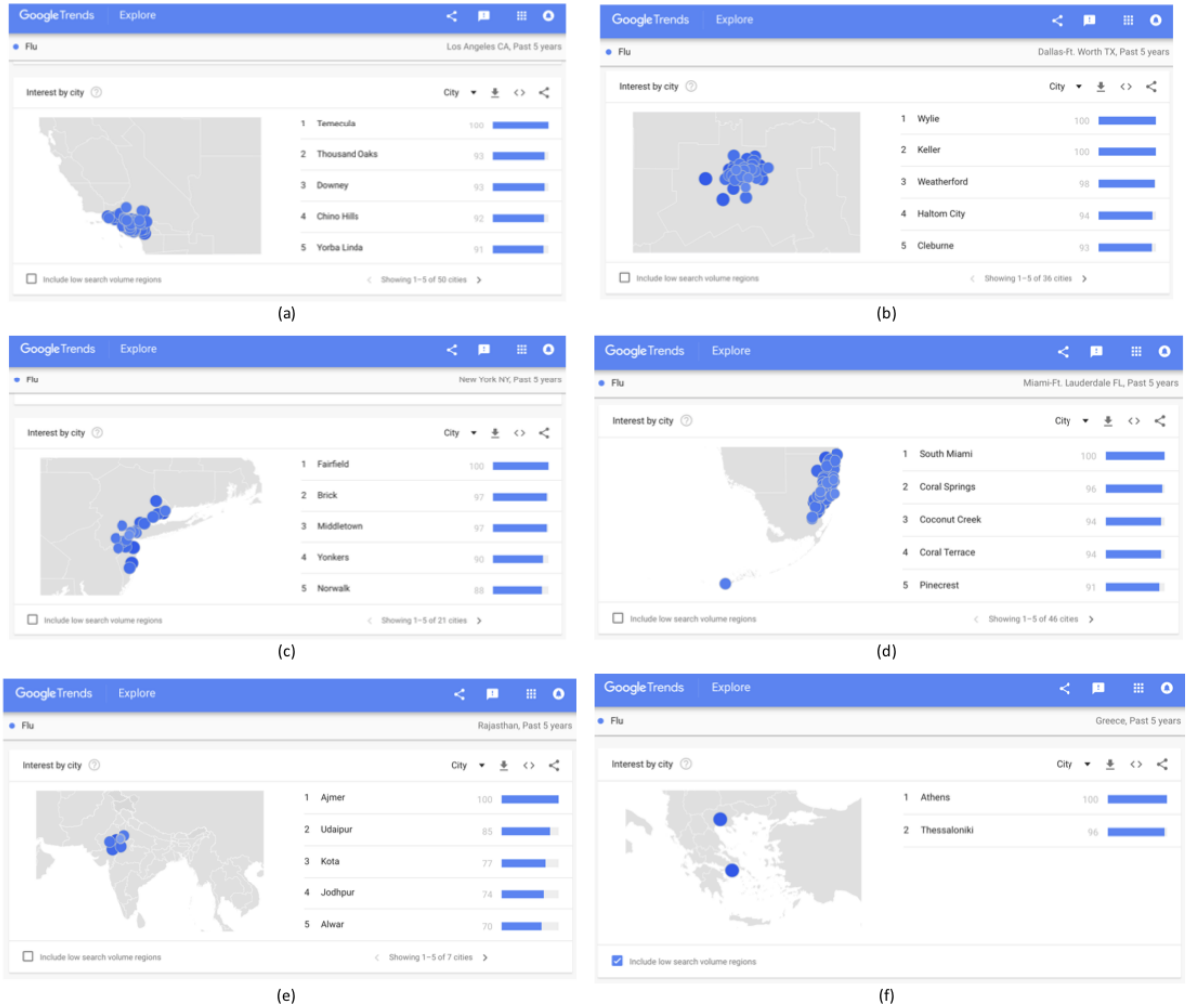


Figure 9. Regional online interest in the term "Flu" at city level over the past 5 years in (a) Los Angeles, (b) Dallas, (c) New York, (d) Miami, (e) India, and (f) Greece.



Period Selection

As the data are normalized over the selected period, the time frame for which Google Trends data are retrieved is crucial for the validity of the results. The selection of the examined time frame is one of the most common mistakes in Google Trends research. The main guideline is that the period selected for Google data should be exactly the same as the one for which official data are available and will be examined. For example, if monthly (or yearly) official data from January 2004 to December 2014 are available, then the selected period for retrieving Google Trends data should be January 2004 to December 2014. Neither 15 datasets for each individual year nor a random number of datasets arbitrarily chosen should be used; a single dataset should be compiled including the months from January 2004 to December 2014. Note that data may slightly vary depending on the time of retrieval; thus, the date and time of downloading must be reported.

Depending on the time frame, the interval for which data are available varies significantly (Table 3), which includes the data intervals for the preselected time frames in Google Trends. Note that the default selection is 12 months.

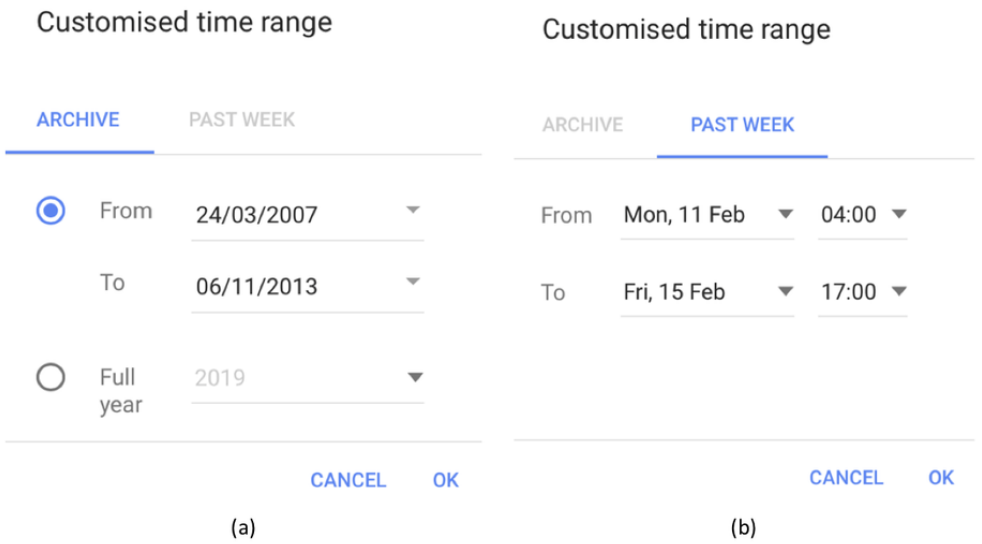
The time frame can be customized at will; for example, March 24, 2007, to November 6, 2013 (Figure 10 a). Furthermore, there is an option to select the exact hours for which data are retrieved, but only over the past week; for example, from February 11, 4 am, to February 15, 5 pm (Figure 10 b).

Finally, an important detail in the selection of the time frame is when the data retrieval changes from monthly to weekly and weekly to daily. For example, from April 28, 2013, to June 30, 2018, the data are retrieved in weekly intervals, while from April 27, 2013, to June 30, 2018, the data are retrieved in monthly intervals. Hence, the data from monthly to weekly changes in (roughly) 5 years and 2 months. For daily data, we observe that, for example, from October 4, 2017, to June 30, 2018, the data are retrieved in daily intervals, while from October 3, 2017, to June 30, 2018, the data are retrieved in weekly intervals; as such, the data interval changes from daily to weekly in (roughly) 10 months.

Table 3. Data intervals and number of observations for the default options in period selection.

Selected period	Data intervals	Number of observations
2004 to present	Monthly	>187
Past 5 years	Weekly	260
Full year (eg, 2004 or 2008)	Weekly	52
Past 12 months	Weekly	52
Past 90 days	Daily	90
Past 30 days	Daily	30
Past 7 days	Hourly	168
Past day	8 min	180
Past 4 hours	1 min	240
Past hour	1 min	60

Figure 10. Customized time range (a) from archive and (b) over the past week.



Search Categories

When exploring the online interest, the selected term can be analyzed based on a selected category. This feature is important to eliminate noisy data, especially in cases where the same word is used or can be attributed to different meanings or events. For example, the terms "yes" and "no" are very commonly searched for, so, when aiming at predicting the results of a referendum race, the search must be limited to the category "Politics" or "Campaign and elections" in order to retrieve the data that are attributed to the event. However, selecting a category is not required when the keyword searched is specific and not related to other words, meanings, and events.

The available categories are listed in Table A1 of Multimedia Appendix 1. Note that most of these categories have subcategories, which, in turn, have other subcategories, allowing the available categories to be as broad or as narrow as required.

In this paper, we focus on the category of "Health" (first level of categorization). The main available subcategories (second level of categorization) of "Health" along with all available subcategories (third and fourth levels) are presented in Table A2 of Multimedia Appendix 1.

Finally, another feature is the type of search conducted when entering a keyword, which consists of the options of "Web Search," "Image Search," "News Search,""Google Shopping," and "YouTube Search." Apart from very specific cases, the "Web Search," which is also the default option, should be selected.

Discussion

Over the past decade, Web-based data are used extensively in digital epidemiology, with online sources playing a central role in health informatics [1,2,52]. Digital disease detection [53] consists of detecting, analyzing, and predicting disease occurrence and spread, and several types of online sources are used, including mainly digital platforms [54,55]. When addressing infodemiology topics, a concept first introduced by Eysenbach [1], Google Trends is an important tool, and research on the subject is constantly expanding [56]. Most studies on Google Trends research are in health and medicine, focusing mainly on the surveillance and analysis of health topics and the forecasting of diseases, outbreaks, and epidemics. As Google Trends is open and user friendly, it is accessed and used by several researchers, even those who are not strictly related to the field of big data, but use it as a means of exploring behavioral variations toward selected topics. The latter has resulted in differences in methodologies followed, which, at times, involve mistakes.

Despite the large number of studies in this line of research, there was a lack of a methodology framework that should be followed. This has produced differences in presentation, and, more importantly, in crucial mistakes that compromise the validity of the results. In this article, we provided a concise overview of the how the tool works and proposed a step-by-step methodology (ie, the four steps of selecting the correct/appropriate keyword, region, period, and category) to ensure the validity of the results in Google Trends research. We also included research examples to provide guidance not only to the experienced eye, but also to new researchers.

As is evident by the findings of this study, there are several limitations to the use of Google Trends data. First, despite the evident potential that Google data have to offer in epidemiology and disease surveillance, there have been some issues in the past, where online search traffic data at some point failed to accurately predict disease spreading, as in the case of Google Flu Trends [57], a Google tool for the surveillance of influenza-like illness (the flu) that is no longer available. Regardless, Google Flu Trends has been accurate in the past in predicting the spread of flu, as suggested by several studies and reports [58-60].

The latter could be partly attributed to the fact that, when researching with Google Trends, the sample is unknown and it cannot be shown to be representative. Despite this and considering the increasing internet penetration, previous studies have suggested that Web-based data have been empirically shown to provide valuable and valid results in exploring and predicting behavior and are correlated with actual data [61-66]. However, recent research has suggested that online queries do not provide valid results in regions with low internet penetration or low scorings in freedom of speech [67].

Furthermore, the data that are retrieved are normalized over the selected period; thus, the exact volumes of queries are not known, limiting the way that the data can be processed and analysis can be performed. Therefore, the data should be analyzed in the appropriate way, and the results should be carefully interpreted.

In addition, the selection of keyword(s) plays a very important role in ensuring the validity of the results. In some cases, the noisy data (ie, queries not attributed to the examined term) must be excluded, which are not always trivial. This can be partly overcome by selecting a specific category, which always bares the risk of excluding results that are needed for analysis.

The analysis of Google Trends data has several other limitations, as examining Web data can bear threats to validity. Careful analysis should be performed to ensure that news reporting and sudden events do not compromise the validity of the results. In addition, as the sample is unknown, several other demographic factors such as age and sex cannot be included in the analysis.

Finally, as this field of research is relatively new, there is no standard way of reporting, resulting in the same meaning of different terms, different meanings of the same term, and different abbreviations. For example, Google Trends data are referred to as relative search volumes, search volumes, online queries, online search traffic data, normalized hits, and other terms. Thus, future research should focus on developing specific coding for Google Trends research, so that a unified way of reporting is followed by all researchers in the field.

In the era of big data, the analysis of Google queries has become a valuable tool for researchers to explore and predict human behavior, as it has been suggested that online data are correlated with actual health data. The methodology framework proposed in this article for researching with Google Trends is much needed to provide guidance for using Google Trends data in

health assessment, and, more importantly, to help researchers and health officials and organizations avoid common mistakes that compromise the validity of the results. As research on the subject is expanding, future work should include the coding in Google Trends research and extend this framework along with changes in the tool and the analysis methods.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Google Trends categories.

[PDF File (Adobe PDF File), 51KB - publichealth_v5i2e13439_app1.pdf]

References

1. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
2. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med 2011 May;40(5 Suppl 2):S154-S158. [doi: 10.1016/j.amepre.2011.02.006] [Medline: 21521589]
3. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of google trends in health care research: a systematic review. PLoS One 2014;9(10):e109583 [FREE Full text] [doi: 10.1371/journal.pone.0109583] [Medline: 25337815]
4. Mavragani A, Ochoa G, Tsagarakis KP. Assessing the Methods, Tools, and Statistical Approaches in Google Trends Research: Systematic Review. J Med Internet Res 2018 Nov 06;20(11):e270 [FREE Full text] [doi: 10.2196/jmir.9366] [Medline: 30401664]
5. Gamma A, Schleifer R, Weinmann W, Buadze A, Liebrenz M. Could Google Trends Be Used to Predict Methamphetamine-Related Crime? An Analysis of Search Volume Data in Switzerland, Germany, and Austria. PLoS One 2016;11(11):e0166566 [FREE Full text] [doi: 10.1371/journal.pone.0166566] [Medline: 27902717]
6. Mavragani A, Ochoa G. Infoveillance of infectious diseases in USA: STDs, tuberculosis, and hepatitis. J Big Data 2018 Sep 6;5(1). [doi: 10.1186/s40537-018-0140-9]
7. Mavragani A, Sampri A, Sypsa K, Tsagarakis KP. Integrating Smart Health in the US Health Care System: Infodemiology Study of Asthma Monitoring in the Google Era. JMIR Public Health Surveill 2018 Mar 12;4(1):e24 [FREE Full text] [doi: 10.2196/publichealth.8726] [Medline: 29530839]
8. Simpson SS, Adams N, Brugman CM, Conners TJ. Detecting Novel and Emerging Drug Terms Using Natural Language Processing: A Social Media Corpus Study. JMIR Public Health Surveill 2018 Jan 08;4(1):e2 [FREE Full text] [doi: 10.2196/publichealth.7726] [Medline: 29311050]
9. Wongkoblap A, Vadillo MA, Curcin V. Researching Mental Health Disorders in the Era of Social Media: Systematic Review. J Med Internet Res 2017 Dec 29;19(6):e228 [FREE Full text] [doi: 10.2196/jmir.7215] [Medline: 28663166]
10. Park SH, Hong SH. Identification of Primary Medication Concerns Regarding Thyroid Hormone Replacement Therapy From Online Patient Medication Reviews: Text Mining of Social Network Data. J Med Internet Res 2018 Oct 24;20(10):e11085 [FREE Full text] [doi: 10.2196/11085] [Medline: 30355555]
11. Ricard BJ, Marsch LA, Crosier B, Hassanpour S. Exploring the Utility of Community-Generated Social Media Content for Detecting Depression: An Analytical Study on Instagram. J Med Internet Res 2018 Dec 06;20(12):e11817 [FREE Full text] [doi: 10.2196/11817] [Medline: 30522991]
12. Huesch M, Chetlen A, Segel J, Schetter S. Frequencies of Private Mentions and Sharing of Mammography and Breast Cancer Terms on Facebook: A Pilot Study. J Med Internet Res 2017 Jun 09;19(6):e201 [FREE Full text] [doi: 10.2196/jmir.7508] [Medline: 28600279]
13. Chen T, Dredze M. Vaccine Images on Twitter: Analysis of What Images are Shared. J Med Internet Res 2018 Apr 03;20(4):e130 [FREE Full text] [doi: 10.2196/jmir.8221] [Medline: 29615386]
14. Farhadloo M, Winneg K, Chan MS, Hall Jamieson K, Albarracin D. Associations of Topics of Discussion on Twitter With Survey Measures of Attitudes, Knowledge, and Behaviors Related to Zika: Probabilistic Study in the United States. JMIR Public Health Surveill 2018 Feb 09;4(1):e16 [FREE Full text] [doi: 10.2196/publichealth.8186] [Medline: 29426815]
15. van Lent LG, Sungur H, Kunneman FA, van de Velde B, Das E. Too Far to Care? Measuring Public Attention and Fear for Ebola Using Twitter. J Med Internet Res 2017 Dec 13;19(6):e193 [FREE Full text] [doi: 10.2196/jmir.7219] [Medline: 28611015]
16. Du J, Tang L, Xiang Y, Zhi D, Xu J, Song HY, et al. Public Perception Analysis of Tweets During the 2015 Measles Outbreak: Comparative Study Using Convolutional Neural Network Models. J Med Internet Res 2018 Jul 09;20(7):e236 [FREE Full text] [doi: 10.2196/jmir.9413] [Medline: 29986843]

17. Sewalk KC, Tuli G, Hswen Y, Brownstein JS, Hawkins JB. Using Twitter to Examine Web-Based Patient Experience Sentiments in the United States: Longitudinal Study. J Med Internet Res 2018 Oct 12;20(10):e10043 [FREE Full text] [doi: 10.2196/10043] [Medline: 30314959]
18. Chen S, Xu Q, Buchenberger J, Bagavathi A, Fair G, Shaikh S, et al. Dynamics of Health Agency Response and Public Engagement in Public Health Emergency: A Case Study of CDC Tweeting Patterns During the 2016 Zika Epidemic. JMIR Public Health Surveill 2018 Nov 22;4(4):e10827 [FREE Full text] [doi: 10.2196/10827] [Medline: 30467106]
19. Alvarez-Mon MA, Asunsolo Del Barco A, Lahera G, Quintero J, Ferre F, Pereira-Sanchez V, et al. Increasing Interest of Mass Communication Media and the General Public in the Distribution of Tweets About Mental Disorders: Observational Study. J Med Internet Res 2018 May 28;20(5):e205 [FREE Full text] [doi: 10.2196/jmir.9582] [Medline: 29807880]
20. Roccetti M, Marfia G, Salomoni P, Prandi C, Zagari RM, Gningaye KFL, et al. Attitudes of Crohn's Disease Patients: Infodemiology Case Study and Sentiment Analysis of Facebook and Twitter Posts. JMIR Public Health Surveill 2017 Aug 09;3(3):e51 [FREE Full text] [doi: 10.2196/publichealth.7004] [Medline: 28793981]
21. Lu FS, Hou S, Baltrusaitis K, Shah M, Leskovec J, Sosic R, et al. Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis. JMIR Public Health Surveill 2018 Jan 09;4(1):e4 [FREE Full text] [doi: 10.2196/publichealth.8950] [Medline: 29317382]
22. Google Trends. URL: https://trends.google.com/trends/?geo=US [accessed 2019-01-10] [WebCite Cache ID 75LXQbBW2]
23. Mavragani A, Tsagarakis KP. YES or NO: Predicting the 2015 GReferendum results using Google Trends. Technological Forecasting and Social Change 2016 Aug;109:1-5. [doi: 10.1016/j.techfore.2016.04.028]
24. Mavragani A, Ochoa G. Forecasting AIDS prevalence in the United States using online search traffic data. J Big Data 2018 May 19;5(1). [doi: 10.1186/s40537-018-0126-7]
25. Tana JC, Kettunen J, Eirola E, Paakkonen H. Diurnal Variations of Depression-Related Health Information Seeking: Case Study in Finland Using Google Trends Data. JMIR Ment Health 2018 May 23;5(2):e43 [FREE Full text] [doi: 10.2196/mental.9152] [Medline: 29792291]
26. Solano P, Ustulin M, Pizzorno E, Vichi M, Pompili M, Serafini G, et al. A Google-based approach for monitoring suicide risk. Psychiatry Res 2016 Dec 30;246:581-586. [doi: 10.1016/j.psychres.2016.10.030] [Medline: 27837725]
27. Arora VS, Stuckler D, McKee M. Tracking search engine queries for suicide in the United Kingdom, 2004-2013. Public Health 2016 Aug;137:147-153. [doi: 10.1016/j.puhe.2015.10.015] [Medline: 26976489]
28. Zheluk A, Quinn C, Meylakhs P. Internet search and krokodil in the Russian Federation: an infoveillance study. J Med Internet Res 2014 Sep 18;16(9):e212 [FREE Full text] [doi: 10.2196/jmir.3203] [Medline: 25236385]
29. Zhang Z, Zheng X, Zeng DD, Leischow SJ. Tracking Dabbing Using Search Query Surveillance: A Case Study in the United States. J Med Internet Res 2016 Sep 16;18(9):e252 [FREE Full text] [doi: 10.2196/jmir.5802] [Medline: 27637361]
30. Abdellaoui R, Foulquié P, Texier N, Faviez C, Burgun A, Schück S. Detection of Cases of Noncompliance to Drug Treatment in Patient Forum Posts: Topic Model Approach. J Med Internet Res 2018 Mar 14;20(3):e85 [FREE Full text] [doi: 10.2196/jmir.9222] [Medline: 29540337]
31. Allem J, Dharmapuri L, Leventhal AM, Unger JB, Boley Cruz T. Hookah-Related Posts to Twitter From 2017 to 2018: Thematic Analysis. J Med Internet Res 2018 Nov 19;20(11):e11669 [FREE Full text] [doi: 10.2196/11669] [Medline: 30455162]
32. Berlinberg EJ, Deiner MS, Porco TC, Acharya NR. Monitoring Interest in Herpes Zoster Vaccination: Analysis of Google Search Data. JMIR Public Health Surveill 2018 May 02;4(2):e10180 [FREE Full text] [doi: 10.2196/10180] [Medline: 29720364]
33. Bragazzi NL, Mahroum N. Google Trends Predicts Present and Future Plague Cases During the Plague Outbreak in Madagascar: Infodemiological Study. JMIR Public Health Surveill 2019 Mar 08;5(1):e13142 [FREE Full text] [doi: 10.2196/13142] [Medline: 30763255]
34. Foroughi F, Lam AK, Lim MSC, Saremi N, Ahmadvand A. "Googling" for Cancer: An Infodemiological Assessment of Online Search Interests in Australia, Canada, New Zealand, the United Kingdom, and the United States. JMIR Cancer 2016 May 04;2(1):e5 [FREE Full text] [doi: 10.2196/cancer.5212] [Medline: 28410185]
35. Gianfredi V, Bragazzi NL, Mahamid M, Bisharat B, Mahroum N, Amital H, et al. Monitoring public interest toward pertussis outbreaks: an extensive Google Trends-based analysis. Public Health 2018 Dec;165:9-15. [doi: 10.1016/j.puhe.2018.09.001] [Medline: 30342281]
36. Hswen Y, Gopaluni A, Brownstein JS, Hawkins JB. Using Twitter to Detect Psychological Characteristics of Self-Identified Persons With Autism Spectrum Disorder: A Feasibility Study. JMIR Mhealth Uhealth 2019 Feb 12;7(2):e12264 [FREE Full text] [doi: 10.2196/12264] [Medline: 30747718]
37. Jones J, Pradhan M, Hosseini M, Kulanthaivel A, Hosseini M. Novel Approach to Cluster Patient-Generated Data Into Actionable Topics: Case Study of a Web-Based Breast Cancer Forum. JMIR Med Inform 2018 Nov 29;6(4):e45 [FREE Full text] [doi: 10.2196/medinform.9162] [Medline: 30497991]
38. Kandula S, Hsu D, Shaman J. Subregional Nowcasts of Seasonal Influenza Using Search Trends. J Med Internet Res 2017 Dec 06;19(11):e370 [FREE Full text] [doi: 10.2196/jmir.7486] [Medline: 29109069]

39. Keller MS, Mosadeghi S, Cohen ER, Kwan J, Spiegel BMR. Reproductive Health and Medication Concerns for Patients With Inflammatory Bowel Disease: Thematic and Quantitative Analysis Using Social Listening. J Med Internet Res 2018 Jun 11;20(6):e206 [FREE Full text] [doi: 10.2196/jmir.9870] [Medline: 29891471]
40. Mejova Y, Weber I, Fernandez-Luque L. Online Health Monitoring using Facebook Advertisement Audience Estimates in the United States: Evaluation Study. JMIR Public Health Surveill 2018 Mar 28;4(1):e30 [FREE Full text] [doi: 10.2196/publichealth.7217] [Medline: 29592849]
41. Odlum M, Yoon S, Broadwell P, Brewer R, Kuang D. How Twitter Can Support the HIV/AIDS Response to Achieve the 2030 Eradication Goal: In-Depth Thematic Analysis of World AIDS Day Tweets. JMIR Public Health Surveill 2018 Nov 22;4(4):e10262 [FREE Full text] [doi: 10.2196/10262] [Medline: 30467102]
42. Phillips CA, Barz LA, Li Y, Schapira MM, Bailey LC, Merchant RM. Relationship Between State-Level Google Online Search Volume and Cancer Incidence in the United States: Retrospective Study. J Med Internet Res 2018 Jan 08;20(1):e6 [FREE Full text] [doi: 10.2196/jmir.8870] [Medline: 29311051]
43. Poirier C, Lavenu A, Bertaud V, Campillo-Gimenez B, Chazard E, Cuggia M, et al. Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine Learning Methods: Comparison Study. JMIR Public Health Surveill 2018 Dec 21;4(4):e11361 [FREE Full text] [doi: 10.2196/11361] [Medline: 30578212]
44. Radin M, Sciascia S. Infodemiology of systemic lupus erythematous using Google Trends. Lupus 2017 Jul;26(8):886-889. [doi: 10.1177/0961203317691372] [Medline: 28162030]
45. Vasconcellos-Silva PR, Carvalho DBF, Trajano V, de La Rocque LR, Sawada ACMB, Juvanhol LL. Using Google Trends Data to Study Public Interest in Breast Cancer Screening in Brazil: Why Not a Pink February? JMIR Public Health Surveill 2017 Apr 06;3(2):e17 [FREE Full text] [doi: 10.2196/publichealth.7015] [Medline: 28385679]
46. Wakamiya S, Kawai Y, Aramaki E. Twitter-Based Influenza Detection After Flu Peak via Tweets With Indirect Information: Text Mining Study. JMIR Public Health Surveill 2018 Sep 25;4(3):e65 [FREE Full text] [doi: 10.2196/publichealth.8627] [Medline: 30274968]
47. Wang H, Chen D. Economic Recession and Obesity-Related Internet Search Behavior in Taiwan: Analysis of Google Trends Data. JMIR Public Health Surveill 2018 Apr 06;4(2):e37 [FREE Full text] [doi: 10.2196/publichealth.7314] [Medline: 29625958]
48. Watad A, Watad S, Mahroum N, Sharif K, Amital H, Bragazzi NL, et al. Forecasting the West Nile Virus in the United States: An Extensive Novel Data Streams-Based Time Series Analysis and Structural Equation Modeling of Related Digital Searching Behavior. JMIR Public Health Surveill 2019 Feb 28;5(1):e9176 [FREE Full text] [doi: 10.2196/publichealth.9176] [Medline: 30601755]
49. Xu C, Wang Y, Yang H, Hou J, Sun L, Zhang X, et al. Association Between Cancer Incidence and Mortality in Web-Based Data in China: Infodemiology Study. J Med Internet Res 2019 Jan 29;21(1):e10677 [FREE Full text] [doi: 10.2196/10677] [Medline: 30694203]
50. Google Trends. How Trends data is adjusted URL: https://support.google.com/trends/answer/4365533?hl=en-GB&ref_topic=6248052 [accessed 2019-01-10] [WebCite Cache ID 75LXW9Qoh]
51. Scharkow M, Vogelgesang J. Measuring the Public Agenda using Search Engine Queries. International Journal of Public Opinion Research 2011 Mar 01;23(1):104-113. [doi: 10.1093/ijpor/edq048]
52. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital epidemiology. PLoS Comput Biol 2012 Jul;8(7):e1002616 [FREE Full text] [doi: 10.1371/journal.pcbi.1002616] [Medline: 22844241]
53. Brownstein J, Freifeld CC, Madoff LC. Digital disease detection--harnessing the Web for public health surveillance. N Engl J Med 2009 May 21;360(21):2153--5,2157 [FREE Full text] [doi: 10.1056/NEJMp0900702] [Medline: 19423867]
54. Chunara R, Goldstein E, Patterson-Lomba O, Brownstein JS. Estimating influenza attack rates in the United States using a participatory cohort. Sci Rep 2015 Apr 2;5(1). [doi: 10.1038/srep09540] [Medline: 25835538]
55. Leal-Neto OB, Dimech GS, Libel M, Oliveira W, Ferreira JP. Digital disease detection and participatory surveillance: overview and perspectives for Brazil. Rev Saude Publica 2016;50:17 [FREE Full text] [doi: 10.1590/S1518-8787.2016050006201] [Medline: 27191153]
56. Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. J Med Internet Res 2013;15(7):e147 [FREE Full text] [doi: 10.2196/jmir.2740] [Medline: 23896182]
57. Google Flu Trends Data. URL: https://www.google.org/flutrends/about/ [accessed 2019-02-09] [WebCite Cache ID 763FITMeo]
58. Carneiro H, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 2009 Nov 15;49(10):1557-1564. [doi: 10.1086/630200] [Medline: 19845471]
59. Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, et al. Influenza forecasting with Google Flu Trends. PLoS One 2013 Feb;8(2):e56176 [FREE Full text] [doi: 10.1371/journal.pone.0056176] [Medline: 23457520]
60. Eurosurveillance editorial team. Google Flu Trends includes 14 European countries. Euro Surveill 2009;14(40) [FREE Full text] [doi: 10.2807/ese.14.40.19352-en]
61. Mavragani A, Ochoa G. The Internet and the Anti-Vaccine Movement: Tracking the 2017 EU Measles Outbreak. BDCC 2018 Jan 16;2(1):2. [doi: 10.3390/bdcc2010002]

62. Preis T, Moat HS, Stanley HE, Bishop SR. Quantifying the advantage of looking forward. Sci Rep 2012;2:350 [FREE Full text] [doi: 10.1038/srep00350] [Medline: 22482034]
63. Preis T, Moat HS, Stanley HE. Quantifying trading behavior in financial markets using Google Trends. Sci Rep 2013;3:1684 [FREE Full text] [doi: 10.1038/srep01684] [Medline: 23619126]
64. Mavragani A, Sypsa K, Sampri A, Tsagarakis K. Quantifying the UK Online Interest in Substances of the EU Watchlist for Water Monitoring: Diclofenac, Estradiol, and the Macrolide Antibiotics. Water 2016 Nov 18;8(11):542. [doi: 10.3390/w8110542]
65. Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA Annu Symp Proc 2006:244-248 [FREE Full text] [Medline: 17238340]
66. Jun S, Yoo HS, Choi S. Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. Technological Forecasting and Social Change 2018 May;130:69-87. [doi: 10.1016/j.techfore.2017.11.009]
67. Mavragani A, Tsagarakis KP. Predicting referendum results in the Big Data Era. J Big Data 2019 Jan 14;6:3. [doi: 10.1186/s40537-018-0166-z]

Edited by G Eysenbach; submitted 18.01.19; peer-reviewed by SH Hong, A Zheluk, O Leal Neto, C Geoghegan, MS Aslam; comments to author 07.02.19; revised version received 17.02.19; accepted 23.03.19; published 11.05.19

Please cite as:
Mavragani A, Ochoa G
Google Trends in Infodemiology and Infoveillance: Methodology Framework
JMIR Public Health Surveill 2019;5(2):e13439
URL: http://publichealth.jmir.org/2019/2/e13439/
doi:10.2196/13439
PMID:

Original Paper

Integrating Smart Health in the US Health Care System: Infodemiology Study of Asthma Monitoring in the Google Era

Amaryllis Mavragani[1], BSc, MSc; Alexia Sampri[1], DipEng, MSc; Karla Sypsa[2], MPharm; Konstantinos P Tsagarakis[3], DipEng, PhD

[1]Department of Computing Science and Mathematics, Faculty of Natural Sciences, University of Stirling, Stirling, United Kingdom

[2]Department of Pharmacy and Forensic Science, King's College London, University of London, London, United Kingdom

[3]Business and Environmental Technology Economics Lab, Department of Environmental Engineering, Democritus University of Thrace, Xanthi, Greece

Corresponding Author:
Amaryllis Mavragani, BSc, MSc
Department of Computing Science and Mathematics
Faculty of Natural Sciences
University of Stirling
University Campus
Stirling, FK9 4LA
United Kingdom
Phone: 44 752 378 2711
Email: amaryllis.mavragani1@stir.ac.uk

Abstract

Background: With the internet's penetration and use constantly expanding, this vast amount of information can be employed in order to better assess issues in the US health care system. Google Trends, a popular tool in big data analytics, has been widely used in the past to examine interest in various medical and health-related topics and has shown great potential in forecastings, predictions, and nowcastings. As empirical relationships between online queries and human behavior have been shown to exist, a new opportunity to explore the behavior toward asthma—a common respiratory disease—is present.

Objective: This study aimed at forecasting the online behavior toward asthma and examined the correlations between queries and reported cases in order to explore the possibility of nowcasting asthma prevalence in the United States using online search traffic data.

Methods: Applying Holt-Winters exponential smoothing to Google Trends time series from 2004 to 2015 for the term "asthma," forecasts for online queries at state and national levels are estimated from 2016 to 2020 and validated against available Google query data from January 2016 to June 2017. Correlations among yearly Google queries and between Google queries and reported asthma cases are examined.

Results: Our analysis shows that search queries exhibit seasonality within each year and the relationships between each 2 years' queries are statistically significant ($P<.05$). Estimated forecasting models for a 5-year period (2016 through 2020) for Google queries are robust and validated against available data from January 2016 to June 2017. Significant correlations were found between (1) online queries and National Health Interview Survey lifetime asthma ($r=-.82$, $P=.001$) and current asthma ($r=-.77$, $P=.004$) rates from 2004 to 2015 and (2) between online queries and Behavioral Risk Factor Surveillance System lifetime ($r=-.78$, $P=.003$) and current asthma ($r=-.79$, $P=.002$) rates from 2004 to 2014. The correlations are negative, but lag analysis to identify the period of response cannot be employed until short-interval data on asthma prevalence are made available.

Conclusions: Online behavior toward asthma can be accurately predicted, and significant correlations between online queries and reported cases exist. This method of forecasting Google queries can be used by health care officials to nowcast asthma prevalence by city, state, or nationally, subject to future availability of daily, weekly, or monthly data on reported cases. This method could therefore be used for improved monitoring and assessment of the needs surrounding the current population of patients with asthma.

(JMIR Public Health Surveill 2018;4(1):e24) doi:10.2196/publichealth.8726

KEYWORDS

asthma; big data; forecasting; Google trends; health care; health informatics; internet behavior; nowcasting; online behavior; smart health

Introduction

Health informatics is the field where information technology, computer science, social sciences, and health care meet [1]. Recently, with the use of big data (ie, large data volumes characterized by high speed and wide dataset variety [2-4]) being all the more applied in research in general, health informatics provides fertile ground for big data applications.

According to Gu et al [5], big data health care research consists of 3 research stages: disease, life and health, and nursing. Focus is being given to various aspects of diseases, technology, and health care services in areas such as epidemics, data mining, machine learning, and customized service [5]. Big data is being increasingly integrated in health care informatics [5-6] and has been used in the past in smart city management.

Over the last few years during the integration of the health pillar in smart cities, where big data is being continuously gathered and analyzed [7], the concept of smart health has been rising [8-10]. Smart health as a concept is derived from the intersection of medical informatics, public health, and business, where large volumes of social media data, payer-provider big data, genomic-driven big data, and biomedical data are being used for the monitoring and evaluation of patients' conditions [10]. As life expectancy increases, so does the cost of health care, and thus innovative methods are required to achieve improved cost-effective quality services. The use of big data in smart health can assist in P4 medicine (preventive, participatory, predictive, and personalized) [8], in the detection, prediction, and prevention of diseases [5], and in the health industry in general [10] while also taking into account the cost, data sources and quality, and population [4].

What has been of notable popularity in big data analytics is the analysis of online search queries [11-12], mainly using Google Trends [13], a popular open tool that has been widely integrated in scientific research over the course of the past decade, mainly focused on health-related topics [6]. Examples include analysis of online interest in multiple sclerosis [14], epilepsy [15-16], silicosis [17], dementia [18], urinary tract infection [19], Ebola [20], the flu [21-23], tobacco and lung cancer [24], epidemics [25-26], and even in illegal drugs such as dabbing [27], krokodil [28], and methamphetamine [29]. This use of big data has formed the cornerstone of a new concept, the science of infodemiology, which uses the vast variety of data available on the internet such as online queries, publications, or posts on blogs and websites for real-time data analysis with the aim of informing public health and public policy, thus providing a viable alternative to the time-consuming traditional methods of gathering health care data such as population surveys and registries. The use of infodemiology data for surveillance purposes is called infoveillance and could potentially allow for more timely and targeted health care interventions [30].

In this study, online queries for the term "asthma" in the United States were analyzed in order to explore the possibility of nowcasting (ie, forecasting the present) asthma prevalence using Google Trends. Asthma was selected because it is a common chronic respiratory disease characterized by exacerbations, also known as asthma attacks; therefore, the reported cases are bound to show seasonality as well as constant interest.

Asthma is a chronic condition characterized by airway inflammation and hyper-responsiveness that causes airways to constrict in response to exercise, infection, exposure to allergens, and occupational exposures [31]. In 2014, it was estimated that approximately 7.4% of the adult US population and 8.6% of US children lived with asthma [32]. During childhood, asthma is more prevalent in males, whereas in adulthood prevalence shifts toward females. Black and multirace people also have a higher prevalence than white people [33-34].

Asthma presents with coughing, wheezing, and chest tightness that seem to be worse during the night and early mornings. These symptoms, along with a family history of asthma or atopic dermatitis, can prompt investigations to confirm an asthma diagnosis. Exacerbation of normal asthma symptoms is more common in patients with uncontrolled asthma or in high-risk patients [35]. Certain types of asthma exacerbations are linked to particular seasons of the year with those caused by pollen and mold being truly seasonal [36]. It has been shown that pediatric patients experience a peak of asthma exacerbations during the fall and spring months [37], whereas adult patients experience a peak of asthma exacerbations at year end [38].

The management of asthma usually involves the use of several inhalers, leading to a rather complicated treatment regime that presents difficulties in terms of patient compliance because it interferes with their daily living activities. Poor compliance can lead to increased morbidity as well as increased cost of treatment [39]. Apart from treatment compliance, another important factor that weighs in the success of the treatment is inhaler technique, as improper inhaler use is linked to poor asthma control. Studies have shown that 33% to 94% of patients do not receive any training regarding proper inhaler technique, which leads to a great number of patients using inhalers incorrectly [40]. Asthma self-management education and personalized advice can improve a patient's asthma control and quality of life, along with reducing asthma exacerbations and hospital admissions [41].

Asthma has several social complications such as limiting patients' activity levels [42], which has an economic impact on the country's health care system. It was estimated that in 2007, medical expenses, missed work and school days, and early deaths due to asthma cost the United States $56 billion [43].

Google Trends data have been previously shown to be valid by many studies [44], and work on the subject has shown the tool's contribution to forecasting [45-46] and analysis of online behavior, provided careful selection of the examined terms [47]. The aim of this paper is to examine if nowcasting asthma prevalence in the United States is possible using online search traffic data.

Methods

Monthly time series from Google Trends for the keyword "asthma" from 2004 to 2015 in the United States and by individual state were used. The data were normalized by Google and downloaded in .csv format on July 7, 2017, between 12:47 and 13:02 for the United States and on July 18 between 14:03 and 14:33 for each of the 50 states and the District of Columbia. The data adjustment procedure is reported by Google as follows [48]: "Search results are proportionate to the time and location of a query: Each data point is divided by the total searches of the geography and time range it represents, to compare relative popularity. Otherwise places with the most search volume would always be ranked highest. The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics. Different regions that show the same number of searches for a term will not always have the same total search volumes."

The seasonality of asthma queries was explored followed by the estimation of the forecasts for the online interest in the term from 2016 through 2020 for the country as well as for each state. The additive method for the Holt-Winters exponential smoothing (using the statistical programming language R) is employed. The Holt-Winters equations [49] can be seen in Figure 1.

In order to further elaborate on the seasonality, the Pearson correlations for Google Trends data for the term "asthma" between each 2 years from 2004 to 2015 in the United States were calculated. Finally, the Pearson correlations between Google queries and the National Health Interview Survey (NHIS) prevalence data [50] from 2004 to 2015 and Behavioral Risk Factor Surveillance System (BRFSS) prevalence data [51] from 2004 to 2014 were examined.

Asthma is not included in the list of diseases with a Centers for Disease Control and Prevention (CDC) surveillance case definition, defined as "a set of uniform criteria used to define a disease for public health surveillance. Surveillance case definitions enable public health officials to classify and count cases consistently across reporting jurisdictions. They provide uniform criteria of national notifiable infectious and non-infectious conditions for reporting purposes" [52]. Thus, nationwide surveys are used to gather information regarding asthma prevalence, including additional information on asthma control, medications, and hospitalizations [53]. The BRFSS is a "state-based, random-digit–dialed telephone survey designed to monitor the prevalence of the major behavioral risks among adults associated with premature morbidity and mortality," and the NHIS is a "multistage probability sample survey designed to solicit health and demographic information about the population, conducted annually with face-to-face interviews in a nationally representative sample of households" [54].

In 2011, the BRFSS changed its weighting methodology in addition to also including mobile phone respondents. Therefore, any comparisons between years before and after 2011 should be carefully interpreted. In this study, no such comparisons are made, as each year's online queries are compared with the respective year's asthma reported cases, thus including no cross-year comparisons. For this study, we used the CDC definition of asthma prevalence, based on affirmative responses to the following NHIS questions: (adults) "Have you ever been told by a doctor or other health professional that you had asthma?" and "Do you still have asthma?" and (children) "Has a doctor or other professional ever told you that [sample child] had asthma?" and "Does [sample child] still have asthma?" [55].

Figure 1. Equations for Holt-Winters exponential smoothing, where y_x and $\hat{y}_x$ denote the initial series and the forecasts, respectively. The l_x, b_x, and s_x denote the level, the trend, and seasonal estimates for month x, respectively, with m denoting the period of the seasonality (ie, 12 in this case), and $h^+_m=[(h-1) \bmod m]+1$. The level, trend, and seasonal change smoothing factors are denoted by constants α, β^*, and γ, respectively. The estimated values for the coefficients for the level and trend are denoted by a and b, respectively, while the seasonal coefficients are denoted by $s_1,\ldots,s_{12}$, for month 1,...,12, respectively.

$$(1) \quad \hat{y}_{x+h|x} = l_x + hb_x + s_{x-m+h^+_m}$$

$$(2) \quad l_x = \alpha(y_x - s_{x-m}) + (1-\alpha)(l_{x-1} + b_{x-1})$$

$$(3) \quad b_x = \beta^*(l_x - l_{x-1}) + (1-\beta^*)b_{x-1}$$

$$(4) \quad s_x = \gamma(y_x - l_{x-1} - b_{x-1}) + (1-\gamma)s_{x-m}$$

Results

Online Interest in the United States

Figure 2 shows a heat map of the United States classified into 5 groups of interest in the term "asthma" from 2004 to 2015 (ie, 0 to 20, 21 to 40, 41 to 60, 61 to 80, and 81 to 100; light blue to darker blue).

Out of the 50 states and District of Columbia, 29 fall into the 81 to 100 group, 21 in the 61 to 80 group, only 1 (Oregon) in the 41 to 60 group, and none in the 21 to 40 and 0 to 20 groups. This classification indicates that the examined term is of high interest to the population of the United States. The detailed data for Figure 2 are available in Multimedia Appendix 1, Table A1.

Figures 3 and 4 depict the changes in online interest in the term "asthma" for the period 2004 to 2015 and the seasonal changes

for each year from 2004 to 2015, respectively. As is evident, the data follow a seasonal trend. All years' data, as presented in Figure 4, follow a similar pattern during a full year, supporting our hypothesis that the seasonality of asthma prevalence in the United States is depicted in online searches.

Figure 5 consists of the changes by state in online interest in the term "asthma" by year from 2004 to 2015. All data are available in Multimedia Appendix 1, Table A2.

There has been a significant increase in searches for the term "asthma" in the states from 2004 to 2015, with the lowest count of states in the 81 to 100 group being in 2007 and the highest in 2012. The top asthma-related queries in the United States from January 2004 to December 2015 include "allergy asthma" (100), "asthma symptoms" (45), "asthma attack" (35), "what is asthma" (25), "asthma inhaler" (20), "asthma children" (15), "exercise asthma" (15), "asthma medications" (10), and "allergy and asthma center" (10).

As is evident, online behavioral changes toward the term "asthma" depict behavior toward said disease. The next steps are to examine if forecasting online interest in the United States is possible and identify existing relationships between online search traffic data and reported asthma cases.

Figure 2. Online interest by state in the term "asthma" from 2004 to 2015.

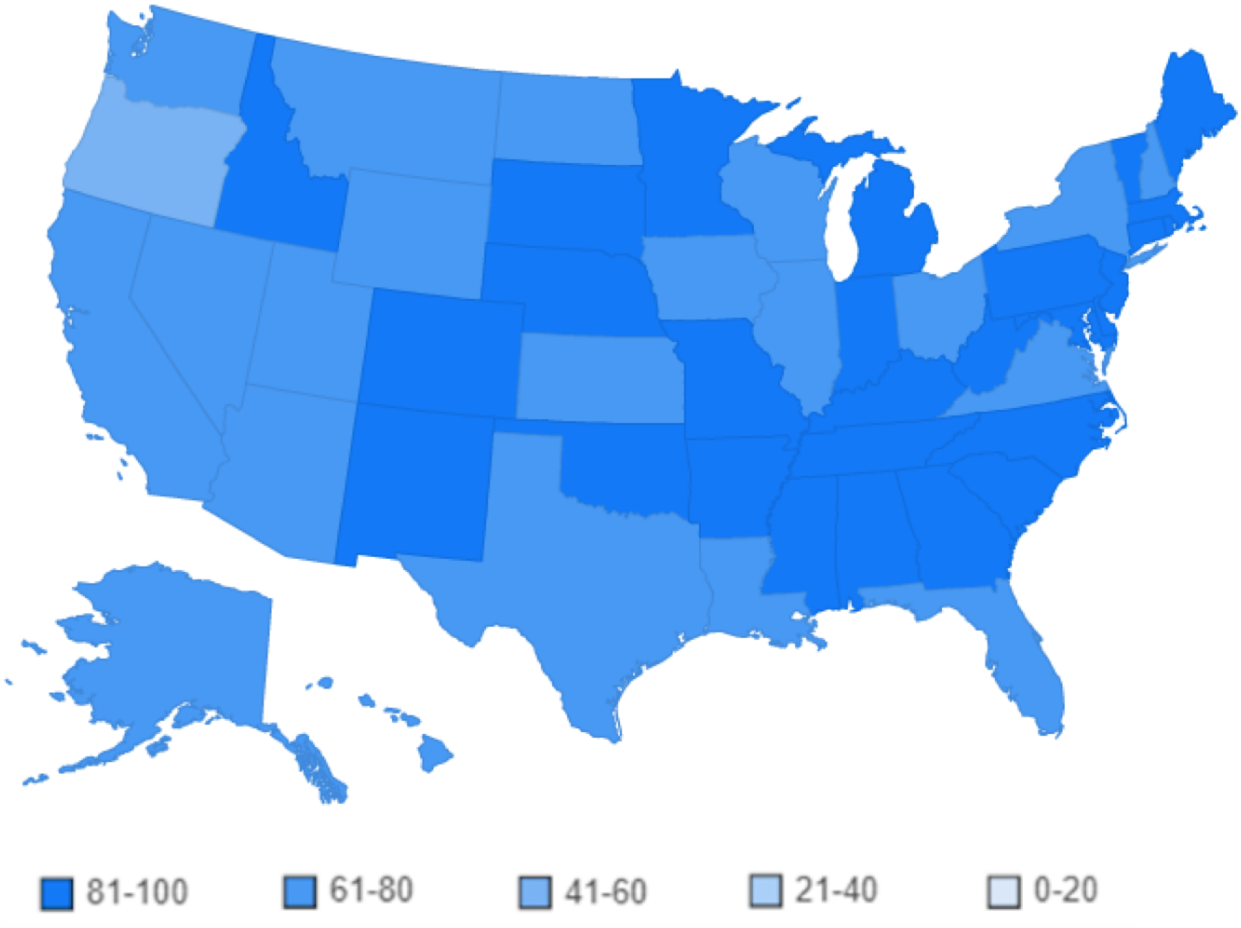


Figure 3. Monthly changes in online interest in the term "asthma" from 2004 to 2015.

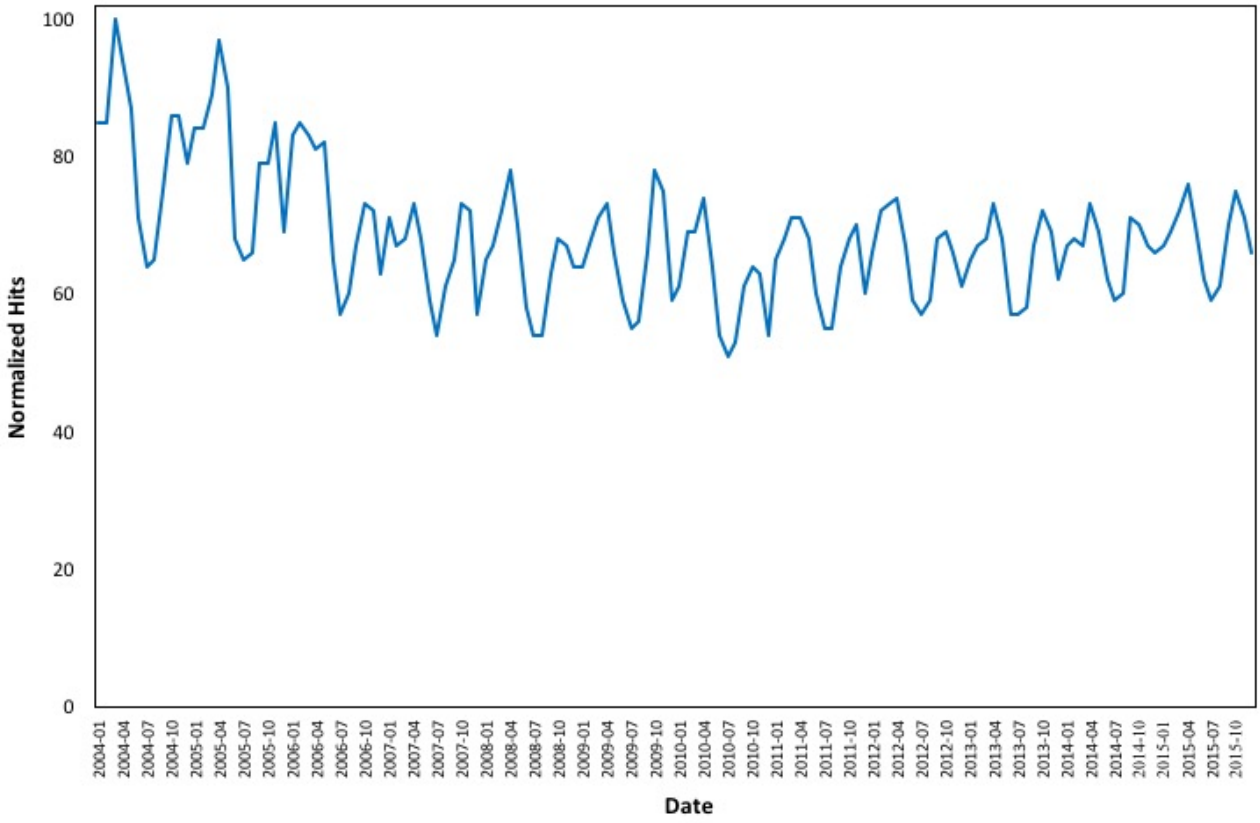


Figure 4. Weekly changes in online interest in the term "asthma" for each year from 2004 to 2015.

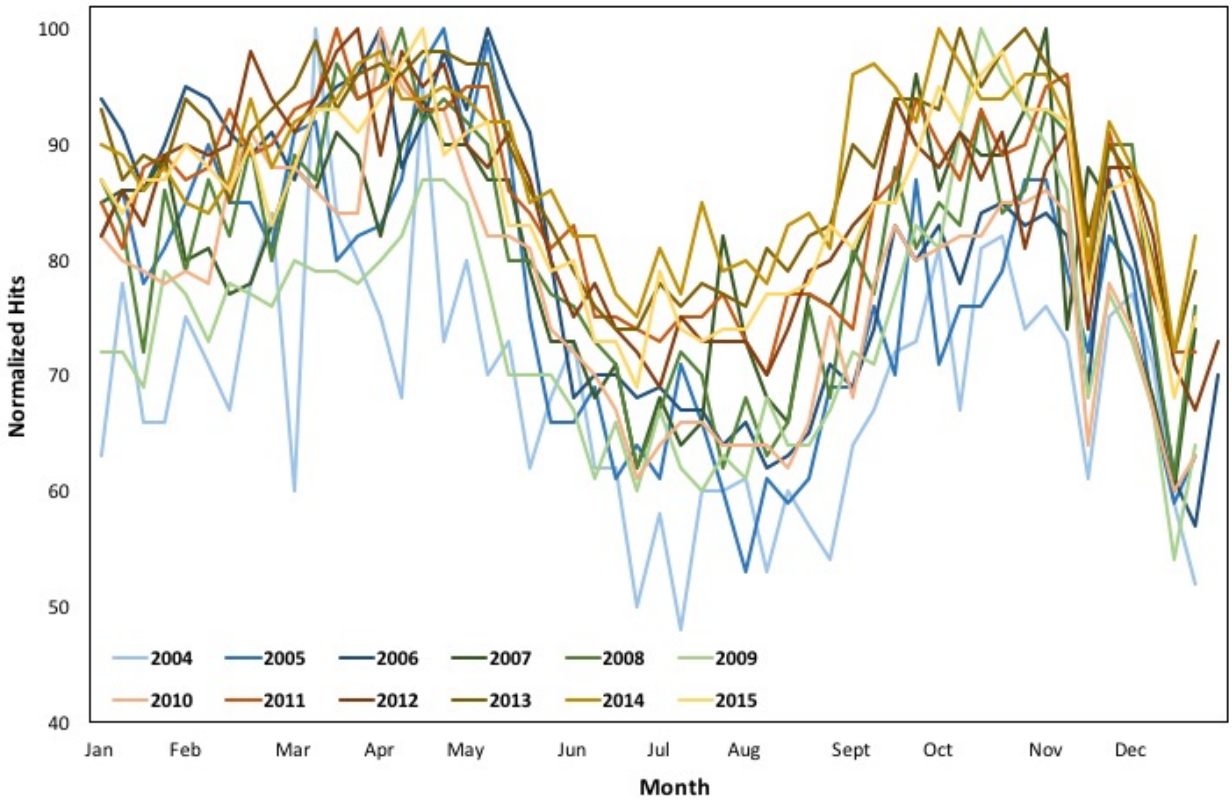
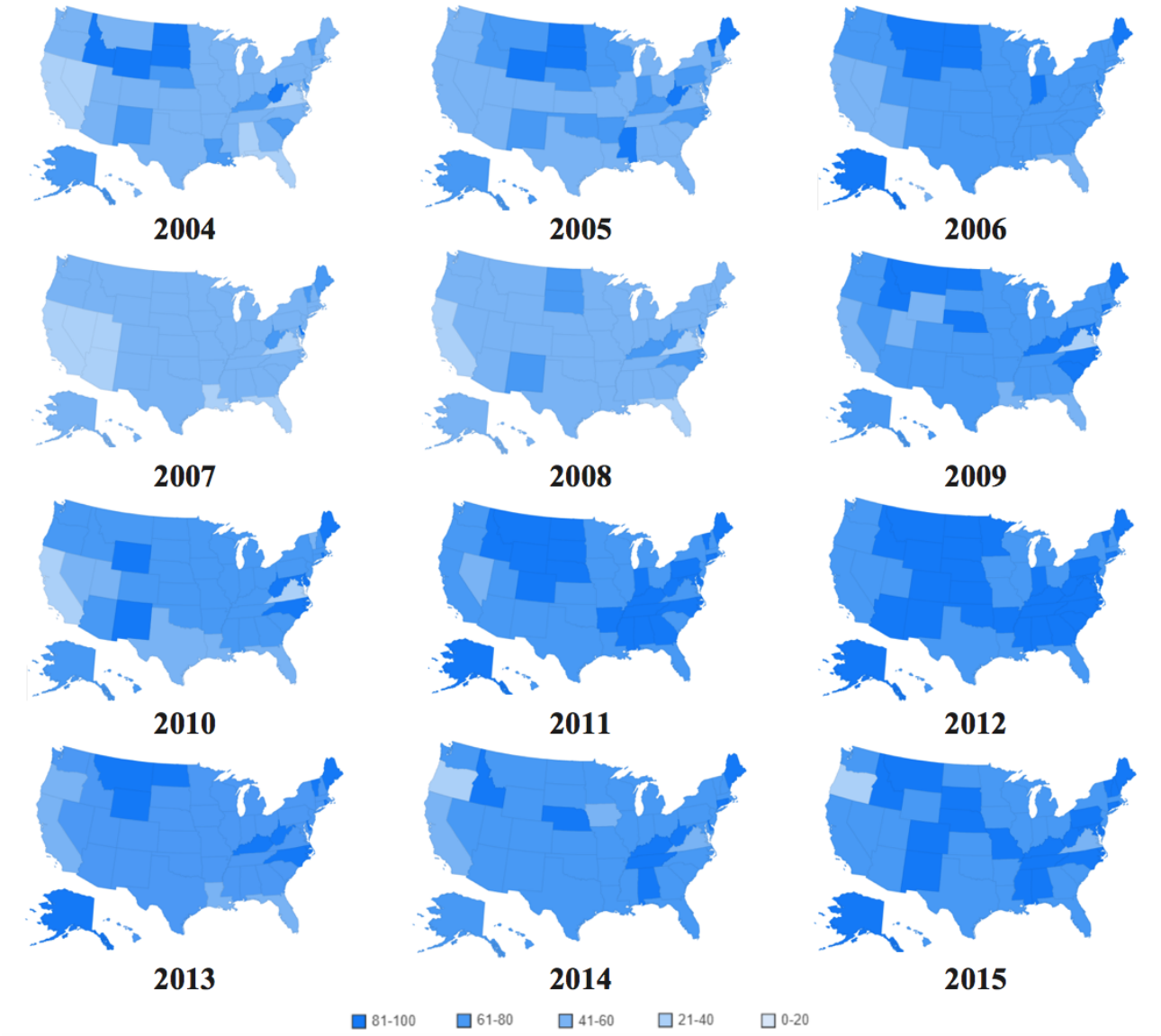


Figure 5. Online interest by state in the term "asthma" per year from 2004 to 2015.



Forecasting Online Interest in the United States

Figure 6 depicts changes in online interest over the period 2004 to 2015 and estimated forecasts from 2005 to 2020. The estimated model closely approximates the actual Google queries for the term “asthma” in the United States over the examined period.

The smoothing parameters for the additive Holt-Winters exponential smoothing with trend and additive seasonal component are α=.33, β^*=0, and γ=.65. The estimated values for the coefficients for the level, trend and season are as follows: a=69.54, b=–.07, s_1=–.94, s_2=1.44, s_3=3.37, s_4=7.84, s_5=2.51, s_6=–5.68, s_7=–8.51, s_8=–7.20, s_9=1.89, s_{10}=4.67, s_{11}=1.11, and s_{12}=–3.53.

In order to elaborate on the robustness of the forecasting model, the estimated values are validated against the available Google queries for the term “asthma” from January 2016 to June 2017, as is shown in Figure 7. It is evident that the forecasts follow the same curve and well approximate the actual Google Trends data for the aforementioned period.

It is therefore suggested that the online behavior exhibits seasonality and can be predicted. The last step in exploring if nowcasting of asthma prevalence in the United States is possible using Google Trends is to examine the correlations between Google Trends data and reported lifetime and current asthma.

Google Trends Versus Reported Asthma

As shown in Figure 4, each examined year’s online interest seems to follow a similar seasonal trend from January to December. To elaborate on the seasonal trend, the Pearson correlations between each 2 years’ queries are calculated (Table 1). The monthly Google Trends data between each 2 years from 2004 to 2015 exhibit high correlations, while all comparisons are statistically significant, with P<.05.

Figure 6. Google Trends (2004 to 2015) versus forecasts (2005 to 2020) in the United States.

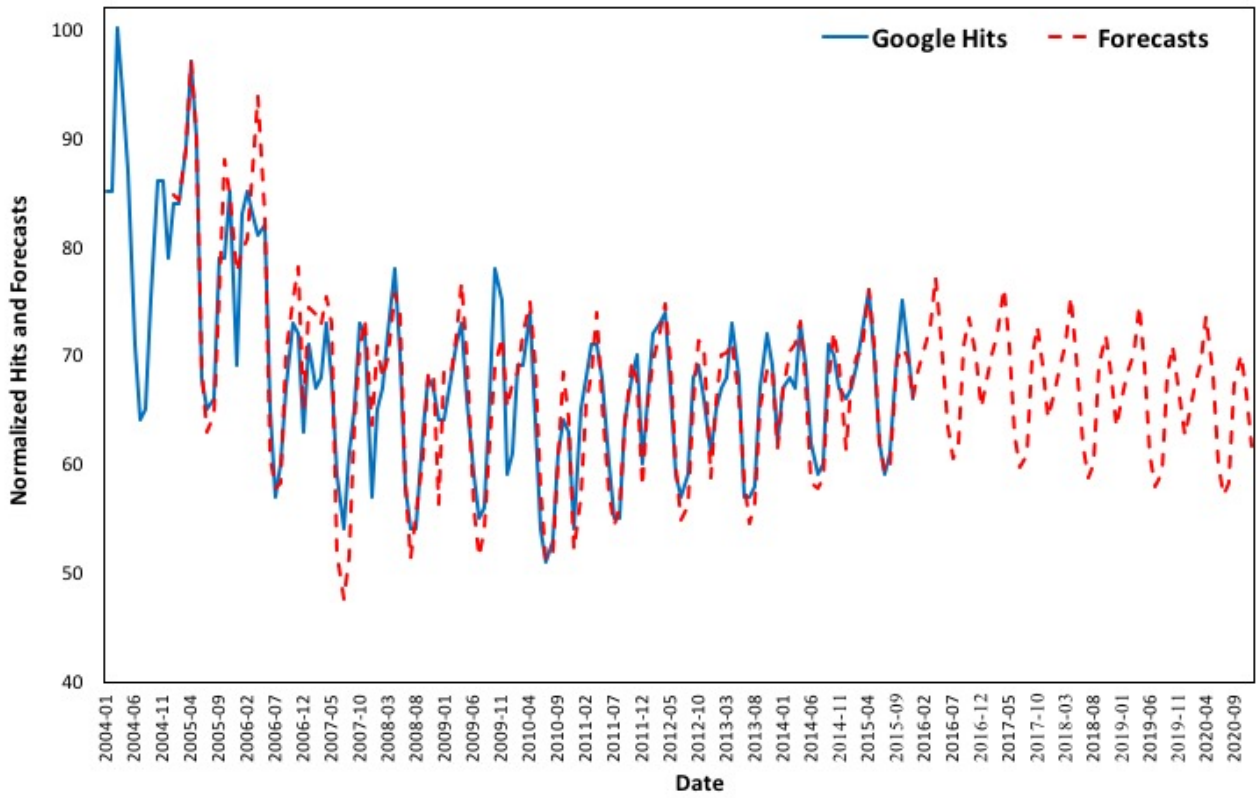


Figure 7. Google Trends (2004 to 2015) versus forecasts (January 2016 to June 2017) in the United States.

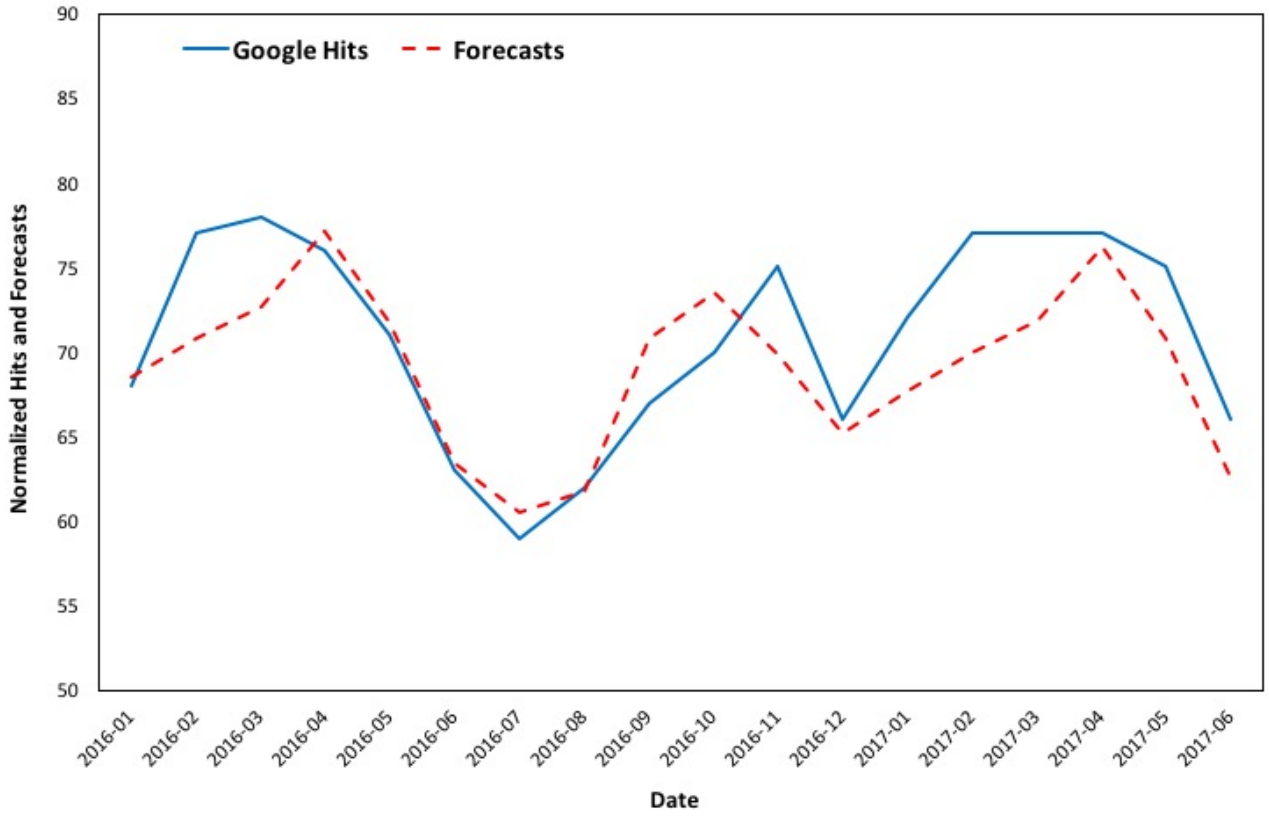


Table 1. Pearson correlations between each 2 years' normalized Google asthma queries in the United States from 2004 to 2015.

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
2005	.89	—	—	—	—	—	—	—	—	—	—
2006	.86	.89	—	—	—	—	—	—	—	—	—
2007	.77	.85	.77	—	—	—	—	—	—	—	—
2008	.94	.93	.81	.78	—	—	—	—	—	—	—
2009	.79	.76	.64	.89	.80	—	—	—	—	—	—
2010	.88	.94	.87	.82	.92	.81	—	—	—	—	—
2011	.94	.93	.85	.87	.93	.91	.93	—	—	—	—
2012	.88	.90	.85	.81	.90	.82	.98	.91	—	—	—
2013	.84	.87	.72	.89	.90	.93	.89	.92	.90	—	—
2014	.75	.82	.68	.77	.87	.78	.82	.83	.86	.92	—
2015	.86	.85	.69	.86	.92	.93	.88	.92	.90	.98	.93

Table 2. Total lifetime and current asthma National Health Interview Survey (2004 to 2015) and Behavioral Risk Factor Surveillance System (2004 to 2014) prevalence data.

Year	NHIS[a]			BRFSS[b]		
	Lifetime asthma	Current asthma	Asthma hits[c]	Lifetime asthma	Current asthma	Asthma hits[c]
2004	30,189	20,545	81.41	33,084,183	20,422,385	83.17
2005	32,621	22,227	79.58	30,661,476	19,453,974	80.33
2006	34,132	22,876	72.58	35,107,599	22,853,570	73.92
2007	34,008	22,879	65.66	36,832,798	23,556,048	68.17
2008	38,450	23,333	65.00	38,050,505	24,521,005	66.92
2009	39,930	24,567	65.83	38,033,371	24,051,245	67.92
2010	39,191	25,710	61.41	39,005,338	25,069,373	62.83
2011	39,504	25,943	64.58	34,759,106	22,605,961	66.42
2012	39,982	25,553	65.91	39,085,744	25,954,771	67.67
2013	37,328	22,648	65.25	41,030,777	26,227,484	67.00
2014	40,461	24,009	66.58	40,706,401	26,957,918	68.75
2015	40,153	24,633	68.16	—	—	—

[a]NHIS: National Health Interview Survey.

[b]BRFSS: Behavioral Risk Factor Surveillance System.

[c]Values slightly vary due to the different time frame: 2004 to 2015 for NHIS and 2004 to 2014 for BRFSS.

To further explore the relationships between online searches and asthma prevalence in the United States, data on the yearly cases of lifetime and current asthma for all ages from the NHIS prevalence data from 2004 to 2015 [50] and the BRFSS prevalence data [51] from 2004 to 2014 (Table 2) are used.

The Pearson correlations of the annual NHIS prevalence data with the annual averages of the normalized Google Trends data from 2004 to 2015 show high correlations between lifetime asthma (r=–.82, P=.001) and current asthma (r=–.77, P=.004). BRFSS prevalence data also exhibit high correlations with Google Trends data for lifetime (r=–.78, P=.003) and current asthma (r=–.79, P=.002). The Spearman correlations for the aforementioned pairs of variables all exhibit the same negative relationship, although not all are statistically significant.

Although statistically significant, all Pearson correlations are negative, and lag analysis should be employed to identify the time interval of response between asthma online interest and case reporting or vice versa. Although Google Trends data for the term "asthma" in the United States over the examined period are monthly, the data on lifetime and current asthma are yearly; until weekly or monthly data are available, further analysis cannot by done.

Forecasting Online Interest by State

In order to show that the method of nowcasting asthma prevalence in the United States using Google queries is possible, this methodology is applied in each of the 50 states and the District of Columbia and exhibits good forecasting results. Figures 8 to 11 depict the changes in online interest in the term

"asthma" from 2004 to 2015 and forecasts from 2016 to 2020 for the 4 most populated states (ie, California, Texas, Florida, and New York), and the graphs for all states can be found in Multimedia Appendix 2, Figures B1-B51. The values of the smoothing parameters α, β^*, and γ and the coefficients for each state's forecasts can be found in Multimedia Appendix 1, Tables A3 and A4, respectively. As online behavioral changes can be predicted and data on asthma cases are correlated with online queries, nowcasting of asthma could be possible provided short-interval data (eg, monthly, weekly, or even daily) are available.

According to the results, online interest in Alaska, Nebraska, New Hampshire, Oklahoma, and Tennessee exhibits increasing forecast trends from 2016 to 2020. On the contrary, online interest in Delaware, Kansas, Oregon, and Virginia exhibits decreasing forecast trends from 2016 to 2020. Overall, the states of Arizona, California, Connecticut, Florida, Georgia, Illinois, Indiana, Maryland, Michigan, Missouri, New Jersey, New York, North Carolina, Pennsylvania, Texas, and Washington show high interest in the term "asthma" throughout the examined period, while in Hawaii and Wyoming, interest is low. Virginia is the only state where online interest exhibits very significant variations from 2004 to 2016.

Our study indicates that analysis of online behavior toward asthma by state can assist with nowcasting asthma prevalence. Since search queries and reporting of asthma are shown to correlate in the United States, if short-interval data (eg, weekly or monthly) were made available, a robust nowcasting model could be developed.

Figure 8. Google Trends (2004 to 2015) versus forecasts (2005 to 2020) in California.

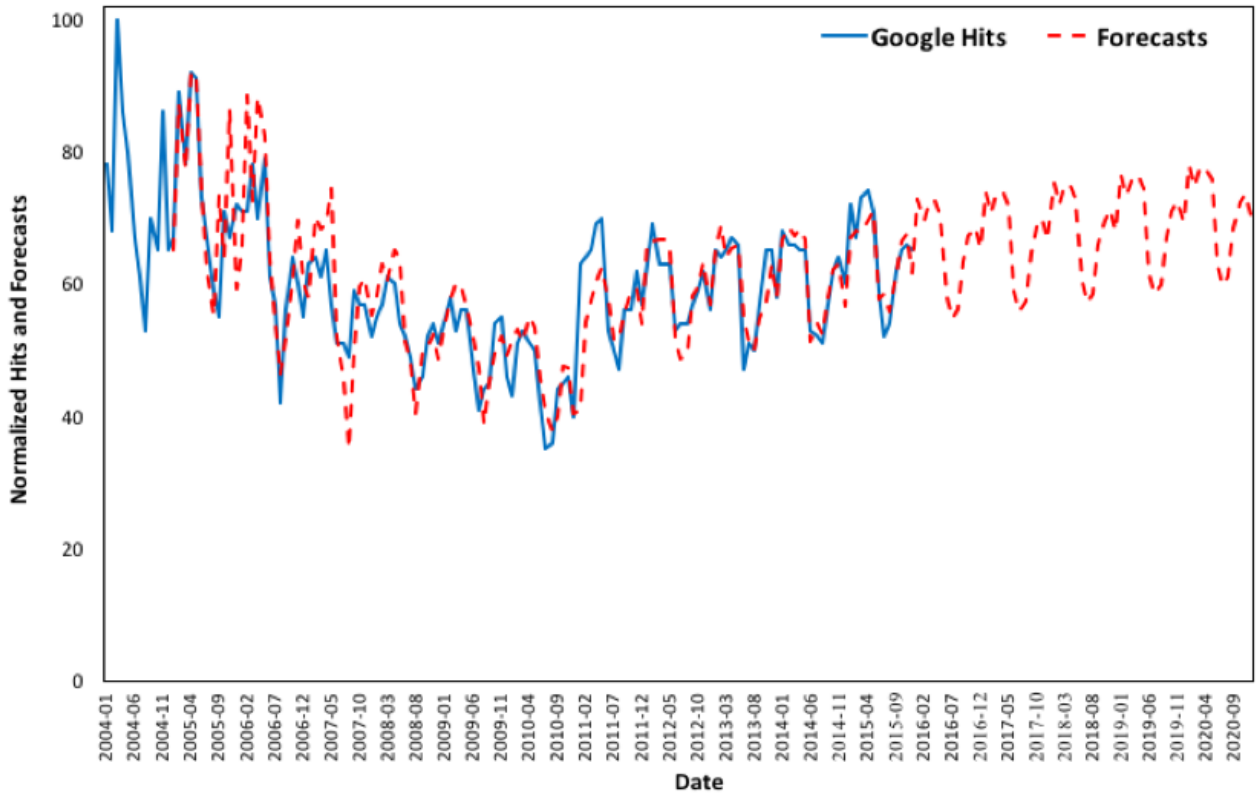


Figure 9. Google Trends (2004 to 2015) versus forecasts (2005 to 2020) in Texas.

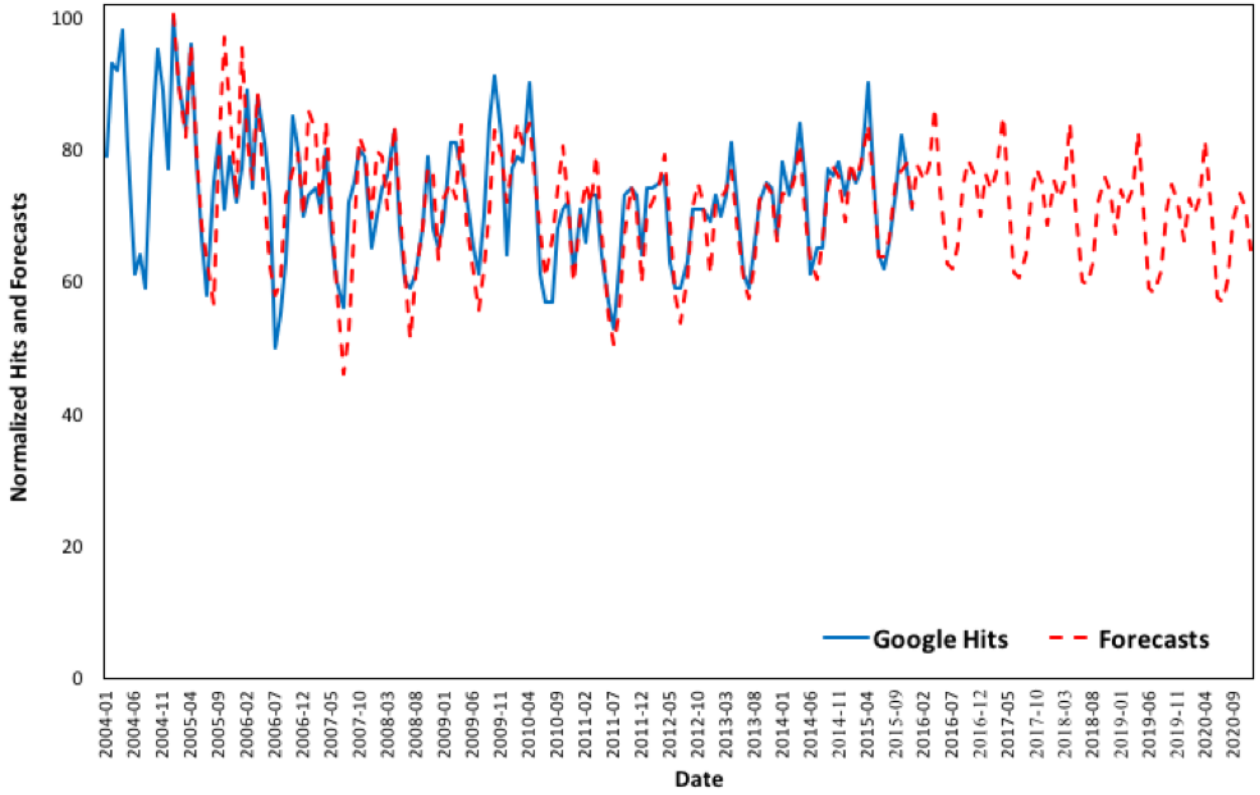


Figure 10. Google Trends (2004 to 2015) versus forecasts (2005 to 2020) in Florida.

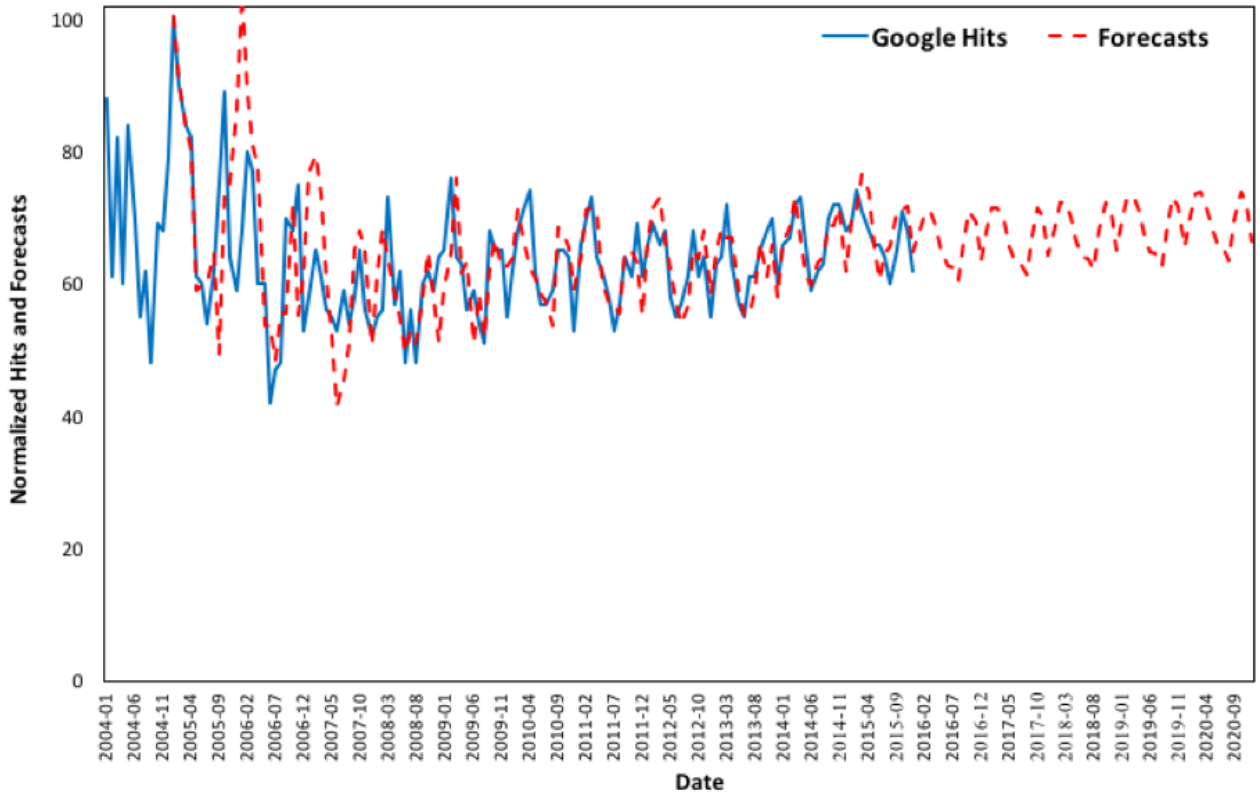
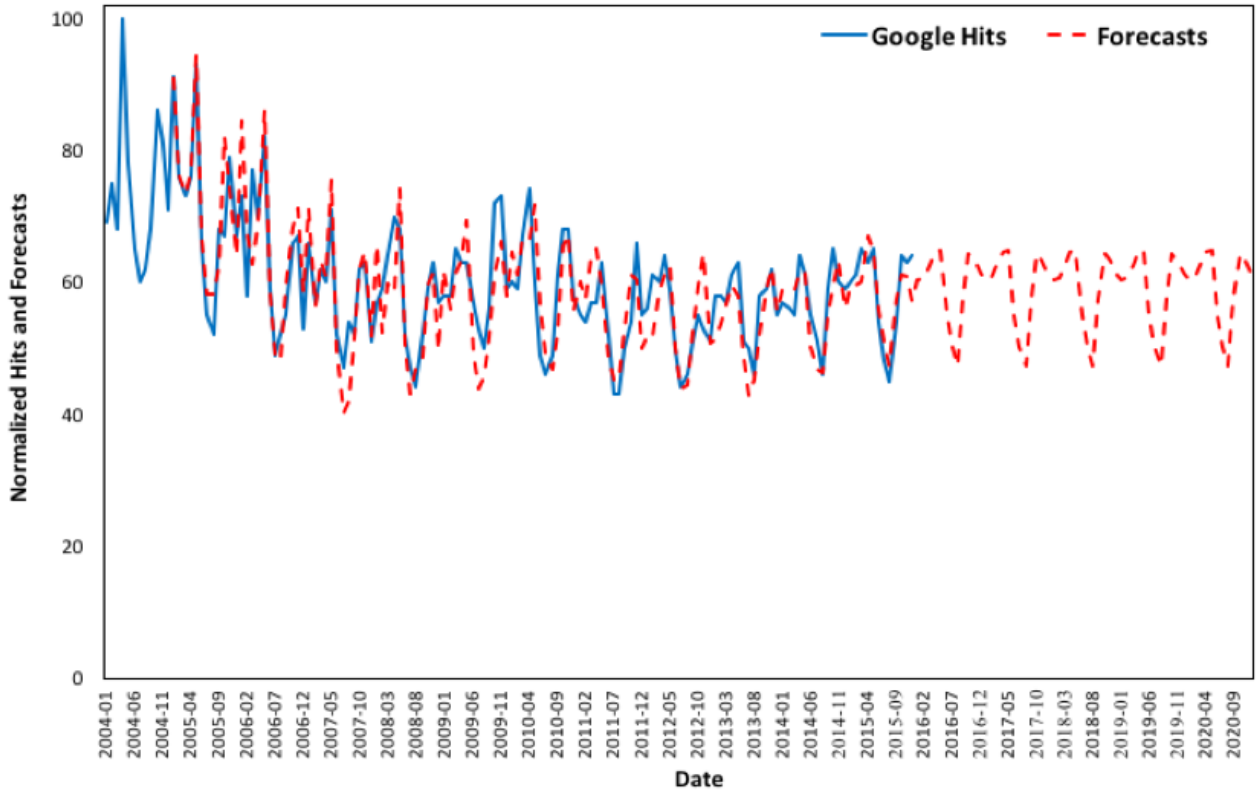


Figure 11. Google Trends (2004 to 2015) versus forecasts (2005 to 2020) in New York.



Discussion

Principal Findings

In addressing integration of smart health into smart city management, monitoring of search traffic data could be useful in predictions and nowcastings, as has also been suggested by previous work on the subject. This study shows that online interest can be predicted nationally and by state. Therefore, governments, policy makers, and health care officials have the ability to use these data to better address the responsiveness of the US health care system at national, regional, state, or even city level in order to nowcast asthma prevalence. Google Trends also provides detailed regional US data, and this method can be applied in other countries as well.

Empirical relationships between Google Trends and human behavior have been suggested, therefore nowcasting asthma prevalence in the United States is possible using online search traffic data, subject to availability of daily, weekly, or monthly data. In this study, it was shown that online search traffic data are highly correlated between each 2 years during the examined period and that Google Trends data are correlated with reported cases of lifetime and current asthma in the United States from 2004 to 2015.

After analyzing changes in online interest in the United States over the examined period, the next step was to identify any seasonal similarities between each 2 years' (monthly) search queries. As the hits between each 2 years from 2004 to 2015 on the term "asthma" were highly correlated, the seasonal effect was evident; using Holt-Winters exponential smoothing, 5-year forecasts for online interest in the term from 2016 to 2020 nationally and in each state were estimated. Validated against available data from January 2016 to June 2017, the forecasts were well fitted and accurately approximated the actual Google Trends data for the same period, suggesting seasonal behavioral changes over the course of a year can be accurately predicted using the proposed method. Google Trends data are correlated with reported cases of lifetime and current asthma, and thus nowcasting asthma prevalence in the United States is suggested to be possible using online search traffic data. As the calculated correlations are negative at this point and there is a lag between internet queries and asthma reporting and vice versa, short-interval data (eg, monthly, weekly, and daily—not available at this point) are required in order to identify said lag.

Limitations

This study has limitations. It cannot be assumed that each hit corresponds to an asthma case and vice versa because hits could be also attributed to academic or research reasons or general interest on the subject, and they could be influenced by news reports or social media. Queries related to asthma could be also influenced by factors such as changes of health insurance and weather or environmental conditions that trigger similar symptoms. This is a general limitation when examining online queries, despite the empirical relationships that have been shown to exist between Google Trends and health data.

The sample is not representative, although as internet penetration increases, so does the possibility of higher volumes of online queries being related to asthma cases. Additionally, nowcasting asthma prevalence using online search queries is not possible at this point because the available data on reported lifetime and

current asthma are yearly. If monthly, weekly, or daily data on past asthma prevalence were available and the correlations between search traffic data and reported asthma are validated, the possibility of nowcasting asthma could be further explored.

This study has not accounted for state-by-state confounders that could influence search patterns, such as the socioeconomic status and demographics of different states that might be relevant to asthma prevalence, as this exceeds the scope of this paper. The latter, along with the impact of socioeconomic and cultural differences on asthma reporting and online search patterns, are of interest for further investigation. In addition, more search terms related to asthma symptoms such as "breathlessness" and "wheezing" could be included in future research on asthma monitoring in the United States.

Conclusion

The findings of this study support previous work on the subject and highlight the value of online data in health and medical informatics. Google Trends data have been shown to be useful and valuable in the monitoring, surveillance, or prediction of epidemics and outbreaks [20,25-26,56], as have been various other internet sources such as Twitter [57], medical portals [58], and Baidu [59]. Google queries provide us with the revealed and not the stated user interest contrary to traditional survey methods [60], and the use of Web data will benefit the exploration of behavior in medical issues [61]. Data from traditional sources and big data should be combined in order to take full advantage of all available information [62]. When daily, weekly, or monthly data on reported asthma cases are made available, data from online sources like Google Trends could be used centrally and then applied by state or used by each city or state individually, assisting with the integration of the smart health concept in smart city management.

Internet behavior can be measured by infodemiology metrics as information patterns and population health are related [30]. Surveillance of asthma is mainly assessed through nationwide surveys and interviews, and data on asthma prevalence are only available long after the cases of asthma are reported. Nowcasting Google queries on selected terms related to asthma could assist health officials at both national and state levels to detect any behavioral variations toward the disease, providing time-effective allocation of resources and a more cost-effective approach to asthma assessment. This study suggests a relationship between asthma prevalence and Google Trends data. In the future, analysis of online queries could be valuable in the monitoring and evaluation of the responsiveness of the US health care system to asthma patient admissions and prescription drug needs, as well as assisting with the implementation of targeted health interventions and campaigns during periods when increased asthma admissions are predicted.

Conflicts of Interest

None declared.

Multimedia Appendix 1

State data tables.

[PDF File (Adobe PDF File), 52KB - publichealth_v4i1e24_app1.pdf]

Multimedia Appendix 2

Google Trends (2004 to 2015) versus forecasts (2005 to 2020) by state.

[PDF File (Adobe PDF File), 3MB - publichealth_v4i1e24_app2.pdf]

References

1. Vignesh RP, Sivasankar E, Pitchiah R. Framework for smart health: toward connected data from big data. 2015 Presented at: Intelligent Computing and Applications Proceedings of the International Conference; December 22-24, 2014; New Delhi p. 423-433. [doi: 10.1007/978-81-322-2268-2]
2. Hilbert M, López P. The world's technological capacity to store, communicate, and compute information. Science 2011 Apr 01;332(6025):60-65 [FREE Full text] [doi: 10.1126/science.1200970] [Medline: 21310967]
3. Chen CP, Zhang C. Data-intensive applications, challenges, techniques and technologies: a survey on big data. Information Sciences 2014 Aug;275:314-347. [doi: 10.1016/j.ins.2014.01.015]
4. Al Nuaimi E, Al Neyadi H, Mohamed N, Al-Jaroodi J. Applications of big data to smart cities. J Internet Serv Appl 2015 Dec 1;6(1). [doi: 10.1186/s13174-015-0041-5]
5. Gu D, Li J, Li X, Liang C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. Int J Med Inform 2017 Feb;98:22-32. [doi: 10.1016/j.ijmedinf.2016.11.006] [Medline: 28034409]
6. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of google trends in health care research: a systematic review. PLoS One 2014;9(10):e109583 [FREE Full text] [doi: 10.1371/journal.pone.0109583] [Medline: 25337815]
7. Khan J, Anjum A, Soomro K, Tahir M. Towards cloud based big data analytics for smart future cities. J Cloud Comp 2015 Feb 18;4(1). [doi: 10.1186/s13677-015-0026-8]

8. Holzinger A, Röcker C, Ziefle M. From smart health to smart hospitals. Lect Notes Comput Sc 2015;8700:1-20. [doi: 10.1007/978-3-319-16226-3_1]
9. Jung H, Chung K. Sequential pattern profiling based bio-detection for smart health service. Cluster Comput 2014 Apr 3;18(1):209-219. [doi: 10.1007/s10586-014-0370-3]
10. Pramanik I, Lau R, Demirkan H, Azad A. Smart health: big data enabled health paradigm within smart cities. Expert Syst Appl 2017 Nov;87:370-383 [FREE Full text] [doi: 10.1016/j.eswa.2017.06.027]
11. Preis T, Moat HS, Stanley HE, Bishop SR. Quantifying the advantage of looking forward. Sci Rep 2012;2:350 [FREE Full text] [doi: 10.1038/srep00350] [Medline: 22482034]
12. Preis T, Moat HS, Stanley HE. Quantifying trading behavior in financial markets using Google Trends. Sci Rep 2013;3:1684 [FREE Full text] [doi: 10.1038/srep01684] [Medline: 23619126]
13. Google Trends. URL: https://trends.google.com/trends/explore [accessed 2017-08-11] [WebCite Cache ID 6sdKCrWKt]
14. Brigo F, Lochner P, Tezzon F, Nardone R. Web search behavior for multiple sclerosis: an infodemiological study. Mult Scler Relat Disord 2014 Jul;3(4):440-443. [doi: 10.1016/j.msard.2014.02.005] [Medline: 25877054]
15. Bragazzi NL, Bacigaluppi S, Robba C, Nardone R, Trinka E, Brigo F. Infodemiology of status epilepticus: a systematic validation of the Google Trends-based search queries. Epilepsy Behav 2016 Feb;55:120-123. [doi: 10.1016/j.yebeh.2015.12.017] [Medline: 26773681]
16. Brigo F, Igwe SC, Ausserer H, Nardone R, Tezzon F, Bongiovanni LG, et al. Why do people Google epilepsy? An infodemiological study of online behavior for epilepsy-related search terms. Epilepsy Behav 2014 Feb;31:67-70. [doi: 10.1016/j.yebeh.2013.11.020] [Medline: 24361764]
17. Bragazzi NL, Dini G, Toletone A, Brigo F, Durando P. Leveraging big data for exploring occupational diseases-related interest at the level of scientific community, media coverage and novel data streams: the example of silicosis as a pilot study. PLoS One 2016;11(11):e0166051 [FREE Full text] [doi: 10.1371/journal.pone.0166051] [Medline: 27806115]
18. Wang H, Chen D, Yu H, Chen Y. Forecasting the incidence of dementia and dementia-related outpatient visits with Google Trends: evidence From Taiwan. J Med Internet Res 2015;17(11):e264 [FREE Full text] [doi: 10.2196/jmir.4516] [Medline: 26586281]
19. Rossignol L, Pelat C, Lambert B, Flahault A, Chartier-Kastler E, Hanslik T. A method to assess seasonality of urinary tract infections based on medication sales and Google Trends. PLoS One 2013;8(10):e76020 [FREE Full text] [doi: 10.1371/journal.pone.0076020] [Medline: 24204587]
20. Alicino C, Bragazzi NL, Faccio V, Amicizia D, Panatto D, Gasparini R, et al. Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes. Infect Dis Poverty 2015 Dec 10;4:54 [FREE Full text] [doi: 10.1186/s40249-015-0090-9] [Medline: 26654247]
21. Kang M, Zhong H, He J, Rutherford S, Yang F. Using Google Trends for influenza surveillance in South China. PLoS One 2013;8(1):e55205 [FREE Full text] [doi: 10.1371/journal.pone.0055205] [Medline: 23372837]
22. Davidson MW, Haim DA, Radin JM. Using networks to combine big data and traditional surveillance to improve influenza predictions. Sci Rep 2015 Jan 29;5:8154 [FREE Full text] [doi: 10.1038/srep08154] [Medline: 25634021]
23. Dukic V, Lopes H, Polson N. Tracking epidemics with Google Flu Trends data and a state-space SEIR model. J Am Stat Assoc 2012 Aug 14;107(500):1410-1426. [doi: 10.1080/01621459.2012.713876]
24. Zhang Z, Zheng X, Zeng DD, Leischow SJ. Information seeking regarding tobacco and lung cancer: effects of seasonality. PLoS One 2015;10(3):e0117938 [FREE Full text] [doi: 10.1371/journal.pone.0117938] [Medline: 25781020]
25. Sentana-Lledo D, Barbu CM, Ngo MN, Wu Y, Sethuraman K, Levy MZ. Seasons, searches, and intentions: what the Internet can tell us about the bed bug (Hemiptera: Cimicidae) epidemic. J Med Entomol 2016 Jan;53(1):116-121. [doi: 10.1093/jme/tjv158] [Medline: 26474879]
26. Fenichel EP, Kuminoff NV, Chowell G. Skip the trip: air travelers' behavioral responses to pandemic influenza. PLoS One 2013;8(3):e58249 [FREE Full text] [doi: 10.1371/journal.pone.0058249] [Medline: 23526970]
27. Zhang Z, Zheng X, Zeng DD, Leischow SJ. Tracking dabbing using search query surveillance: a case study in the United States. J Med Internet Res 2016 Sep 16;18(9):e252 [FREE Full text] [doi: 10.2196/jmir.5802] [Medline: 27637361]
28. Zheluk A, Quinn C, Meylakhs P. Internet search and krokodil in the Russian Federation: an infoveillance study. J Med Internet Res 2014 Sep 18;16(9):e212 [FREE Full text] [doi: 10.2196/jmir.3203] [Medline: 25236385]
29. Gamma A, Schleifer R, Weinmann W, Buadze A, Liebrenz M. Could Google Trends be used to predict methamphetamine-related crime? An analysis of search volume data in Switzerland, Germany, and Austria. PLoS One 2016;11(11):e0166566 [FREE Full text] [doi: 10.1371/journal.pone.0166566] [Medline: 27902717]
30. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
31. Akinbami LJ, Moorman JE, Bailey C, Zahran HS, King M, Johnson CA, et al. Trends in asthma prevalence, health care use, and mortality in the United States, 2001-2010. NCHS Data Brief 2012 May(94):1-8 [FREE Full text] [Medline: 22617340]
32. FastStats: Asthma. National Center for Health Statistics. URL: http://www.cdc.gov/nchs/fastats/asthma.htm [accessed 2018-02-15] [WebCite Cache ID 6sdJKRPmf]

33. Centers for Disease Control and Prevention. Most recent asthma data URL: http://www.cdc.gov/asthma/most_recent_data.htm [WebCite Cache ID 6sdJSHspd]
34. Asthma and African Americans.: US Department of Health and Human Services URL: https://minorityhealth.hhs.gov/omh/browse.aspx?lvl=4&lvlid=15 [accessed 2018-02-21] [WebCite Cache ID 6xP2qadGK]
35. Global Initiative for Asthma. 2017 Pocket guide for asthma management and prevention URL: http://ginasthma.org/download/520/ [WebCite Cache ID 6xFIlR1Ry]
36. Johnston NW, Sears MR. Asthma exacerbations: epidemiology. Thorax 2006 Aug;61(8):722-728 [FREE Full text] [doi: 10.1136/thx.2005.045161] [Medline: 16877691]
37. Larsen K, Zhu J, Feldman LY, Simatovic J, Dell S, Gershon AS, et al. The annual September peak in asthma exacerbation rates. Still a reality? Ann Am Thorac Soc 2016 Feb;13(2):231-239. [doi: 10.1513/AnnalsATS.201508-545OC] [Medline: 26636481]
38. Gerhardsson DVM, Gustafson P, McCrae C, Edsbäcker S, Johnston N. Seasonal and geographic variations in the incidence of asthma exacerbations in the United States. J Asthma 2017 Oct;54(8):818-824. [doi: 10.1080/02770903.2016.1277538] [Medline: 28102717]
39. Gillisen A. Patient's adherence in asthma. J Physiol Pharmacol 2007 Nov;58 Suppl 5(Pt 1):205-222 [FREE Full text] [Medline: 18204131]
40. Axtell S, Haines S, Fairclough J. Effectiveness of various methods of teaching proper inhaler technique. J Pharm Pract 2017 Apr;30(2):195-201. [doi: 10.1177/0897190016628961] [Medline: 26912531]
41. Pinnock H. Supported self-management for asthma. Breathe (Sheff) 2015 Jun;11(2):98-109 [FREE Full text] [doi: 10.1183/20734735.015614] [Medline: 26306110]
42. Winer RA, Qin X, Harrington T, Moorman J, Zahran H. Asthma incidence among children and adults: findings from the Behavioral Risk Factor Surveillance system asthma call-back survey—United States, 2006-2008. J Asthma 2012 Feb;49(1):16-22. [doi: 10.3109/02770903.2011.637594] [Medline: 22236442]
43. Centers for Disease Control and Prevention. Vital Signs: Asthma in the US URL: http://www.cdc.gov/vitalsigns/asthma/ [WebCite Cache ID 6sdJfBsTX]
44. McCallum M, Bury G. Public interest in the environment is falling: a response to Ficetola (2013). Biodivers Conserv 2014 Feb 14;23(4):1057-1062. [doi: 10.1007/s10531-014-0640-7]
45. Jun S, Park D. Consumer information search behavior and purchasing decisions: empirical evidence from Korea. Technol Forecast Soc Change 2016 Jun;107:97-111. [doi: 10.1016/j.techfore.2016.03.021]
46. Han S, Chung H, Kang B. It is time to prepare for the future: forecasting social trends. In: Computer Applications for Database, Education, and Ubiquitous Computing. Berlin: Springer; 2012:325-331.
47. Scharkow M, Vogelgesang J. Measuring the public agenda using search engine queries. Int J Public Opin Res 2011 Mar 01;23(1):104-113. [doi: 10.1093/ijpor/edq048]
48. How Trends data is adjusted. URL: https://support.google.com/trends/answer/4365533?hl=en [accessed 2017-08-11] [WebCite Cache ID 6sdJp7avA]
49. Hyndman R, Athanasopoulos G. Forecasting Principles and Practice. 2014. URL: https://www.otexts.org/fpp [accessed 2018-02-15] [WebCite Cache ID 6xFJlXCQI]
50. Centers for Disease Control and Prevention. 2015. National Health Interview Survey (NHIS) Data: 2015 lifetime asthma, current asthma, asthma attacks among those with current asthma URL: https://www.cdc.gov/asthma/nhis/2015/data.htm [WebCite Cache ID 6sdJhJx5r]
51. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System (BRFSS) prevalence data: asthma URL: https://www.cdc.gov/asthma/brfss/default.htm [WebCite Cache ID 6sdJl5fpw]
52. Centers for Disease Control and Prevention. Surveillance Case Definitions URL: https://wwwn.cdc.gov/nndss/conditions/ [WebCite Cache ID 6wPl4S6NQ]
53. Centers for Disease Control and Prevention. Breathing easier URL: https://www.cdc.gov/asthma/pdfs/breathing_easier_brochure.pdf [WebCite Cache ID 6u57QQ7tx]
54. Centers for Disease Control and Prevention. Data and Surveillance: asthma URL: https://www.cdc.gov/asthma/tables_graphs.htm [WebCite Cache ID 6wPlB6D40]
55. Centers for Disease Control and Prevention. Trends in asthma prevalence, health care use, and mortality in the United States, 2001-2010 URL: https://www.cdc.gov/nchs/data/databriefs/db94.pdf [WebCite Cache ID 6xFKYZQio]
56. Samaras L, García-Barriocanal E, Sicilia M. Syndromic surveillance models using Web data: the case of influenza in Greece and Italy using Google Trends. JMIR Public Health Surveill 2017 Nov 20;3(4):e90 [FREE Full text] [doi: 10.2196/publichealth.8015] [Medline: 29158208]
57. Kagashe I, Yan Z, Suheryani I. Enhancing seasonal influenza surveillance: topic analysis of widely used medicinal drugs using Twitter data. J Med Internet Res 2017 Sep 12;19(9):e315 [FREE Full text] [doi: 10.2196/jmir.7393] [Medline: 28899847]
58. Pesälä S, Virtanen MJ, Sane J, Mustonen P, Kaila M, Helve O. Health information-seeking patterns of the general public and indications for disease surveillance: register-based study using Lyme disease. JMIR Public Health Surveill 2017 Nov 06;3(4):e86 [FREE Full text] [doi: 10.2196/publichealth.8306] [Medline: 29109071]

59. Liu K, Huang S, Miao Z, Chen B, Jiang T, Cai G, et al. Identifying potential norovirus epidemics in China via Internet surveillance. J Med Internet Res 2017 Aug 08;19(8):e282 [FREE Full text] [doi: 10.2196/jmir.7855] [Medline: 28790023]
60. Mavragani A, Tsagarakis K. YES or NO: predicting the 2015 GReferendum results using Google Trends. Technol Forecast Soc Change 2016 Aug;109:1-5. [doi: 10.1016/j.techfore.2016.04.028]
61. Ayers JW, Althouse BM, Dredze M. Could behavioral medicine lead the web data revolution? JAMA 2014 Apr 9;311(14):1399-1400. [doi: 10.1001/jama.2014.1505] [Medline: 24577162]
62. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science 2014 Mar 14;343(6176):1203-1205. [doi: 10.1126/science.1248506] [Medline: 24626916]

Abbreviations

BRFSS: Behavioral Risk Factor Surveillance System
CDC: Centers for Disease Control and Prevention
NHIS: National Health Interview Survey

Edited by G Eysenbach; submitted 11.08.17; peer-reviewed by N Bragazzi, Z Zhang, A Zheluk; comments to author 21.09.17; revised version received 15.10.17; accepted 13.01.18; published 12.03.18

Please cite as:
Mavragani A, Sampri A, Sypsa K, Tsagarakis KP
Integrating Smart Health in the US Health Care System: Infodemiology Study of Asthma Monitoring in the Google Era
JMIR Public Health Surveill 2018;4(1):e24
URL: http://publichealth.jmir.org/2018/1/e24/
doi:10.2196/publichealth.8726
PMID:

big data and cognitive computing

MDPI

Article

The Internet and the Anti-Vaccine Movement: Tracking the 2017 EU Measles Outbreak

Amaryllis Mavragani * and Gabriela Ochoa

Department of Computing Science and Mathematics, Faculty of Natural Sciences, University of Stirling, Stirling, Scotland FK9 4LA, UK; gabriela.ochoa@cs.stir.ac.uk

* Correspondence: amaryllis.mavragani1@stir.ac.uk; Tel.: +44-(0)-7523-782711

Received: 26 November 2017; Accepted: 13 January 2018; Published: 16 January 2018

Abstract: In the Internet Era of information overload, how does the individual filter and process available knowledge? In addressing this question, this paper examines the behavioral changes in the online interest in terms related to Measles and the Anti-Vaccine Movement from 2004 to 2017, in order to identify any relationships between the decrease in immunization percentages, the Anti-Vaccine Movement, and the increased reported Measles cases. The results show that statistically significant positive correlations exist between monthly Measles cases and Google queries in the respective translated terms in most EU28 countries from January 2011 to August 2017. Furthermore, a strong negative correlation ($p < 0.01$) exists between the online interest in the term 'Anti Vaccine' and the Worldwide immunization percentages from 2004 to 2016. The latter could be supportive of previous work suggesting that conspiracist ideation is related to the rejection of scientific propositions. As Measles require the highest immunization percentage out of the vaccine preventable diseases, the 2017 EU outbreak could be the first of several other diseases' outbreaks or epidemics in the near future should the immunization percentages continue to decrease. Big Data Analytics in general and the analysis of Google queries in specific have been shown to be valuable in addressing health related topics up to this point. Therefore, analyzing the variations and patterns of available online information could assist health officials with the assessment of reported cases, as well as taking the required preventive actions.

Keywords: anti-vaccine; anti-vaccine movement; Google Trends; Internet; measles; MMR; online behavior; vaccination

1. Introduction

It was in 1998 when Wakefield et al. [1] published a paper in the Scientific Journal 'The Lancet', suggesting that they identified "*a chronic enterocolitis in children that may be related to neuropsychiatric dysfunction. In most cases, onset of symptoms was after measles, mumps, and rubella immunisation*". This study was conducted on a sample of 12 children, with the overall interpretation of the results being—in simple words—that autism is associated with the Measles-Mumps-Rubella (MMR) vaccine [2]. The claims of this study have since been proven to be false, with over 20 epidemiologic studies showing that no causality or relationship exists between vaccination and autism [3]. Said studies were methodologically solid, i.e., conducted in several countries and by different researchers, while employing epidemiologic and statistical methods for large population sizes. Furthermore, a meta-analysis of more than 40 studies showed that no links between vaccination and autism exist [4].

After the panel hearing where Andrew Wakefield lost his medical license in January 2010 [5], 'The Lancet' retracted the paper, stating that: "*Following the judgment of the UK General Medical Council's Fitness to Practise Panel on Jan 28, 2010, it has become clear that several elements of the 1998 paper by Wakefield et al. are incorrect, contrary to the findings of an earlier investigation. In particular, the claims in the*

original paper that children were "consecutively referred" and that investigations were "approved" by the local ethics committee have been proven to be false. Therefore we fully retract this paper from the published record" [6].

It was at that point when the Anti-Vaccine Movement started to become publicly known, while the retraction of the paper and Wakefield losing his license was the beginning of one of the most well-known conspiracy theories, i.e., that the MMR vaccine causes autism and thus vaccination should be avoided. Said Anti-Vaccine skepticism does not only refer to the MMR, but has been widened to include vaccines in general. This reaction comes not as a surprise, as it has been shown in the past that the rejection of scientific propositions and conspiracist ideation are related [7].

Was the rise of the Anti-Vaccine Movement a result of the public's attraction to conspiracy theories? Was it a result of the past years' increased Internet penetration? Was it a combination of the two? In any case, what can now be observed is a decrease of the immunization coverages in most of the EU countries, resulting in the recent EU Measles outbreak. Specifically, despite that the reported Measles cases decreased in 2009, they experienced an increase by "*a factor of four between 2010 and 2011*" [8]. Out of the vaccine preventable diseases, Measles require the highest immunization percentage coverage [9]. If the EU28 immunization percentages continue to drop, how long will it be before we are talking about an epidemic?

Almost 20 years have passed since Wakefield's [1] study was published, but we are only now able to clearly see the effects of the Anti-Vaccine Movement on public health. Though over the past decades we as a society managed through vaccination to significantly decrease death rates caused by the respective diseases, the spreading of such bogus arguments has resulted in the reappearance of several vaccine preventable diseases, as is the case of Measles. Before the age of the Internet, news channels, newspapers, and other forms of official information sources would not so easily and with such high speed reproduce studies and claims that were not proven to be correct. This is unfortunately the case today in blogs, forums, and social media, constituting a perfect example of how a great life-changing discovery like the World Wide Web could be used to negatively affect public health.

In order to investigate the behavior towards Measles and the Anti-Vaccine Movement, we use data from Google Trends [10], a popular open tool for examining online behavior in Big Data Analytics [11,12]. Subject to careful selection of the examined terms for robust results [13], online queries have been suggested to be beneficial in analyzing behavioral changes [14], while the value and validity of Google Trends' data have been highlighted by previous work on the subject [15,16]. Over the past decade, data from Google Trends have been used to examine the behavior towards several health related topics [17]. As Google Trends' data provide information on the revealed and not the stated users' preferences, they have been shown to assist with the assessment of human behavior in health issues, and that empirical relationships between online search traffic data and official health data exist. For example, Google queries on the respective selected terms have been shown to correlate with suicide rates [18,19], prescription drugs issuing [20,21] and revenues [22], and influenza [23,24]. In addition, online queries have also been shown to be valuable in predicting, detecting, and assessing epidemics and outbreaks [25–27].

Google Trends' data have been effectively employed up to this point in the fields of health and medicine in assessing behavioral changes and in examining relationships that exist between online behavior and human behavior. Towards contributing to the discussion of how online search traffic data can be used in order to analyze and predict human behavior, we first examine the online behavioral variations towards Measles and the Anti-Vaccine Movement from January 2004 to August 2017, Worldwide and in the EU28. Furthermore, we identify the relationships between Google queries and immunization percentages, and investigate the Internet's role in the 2017 EU Measles outbreak—caused by decreased immunization—in relation to the overall Anti-Vaccine skepticism.

The rest of the paper is structured as follows: Section 2 covers the methodology and the procedure of the data collection, the results are presented in Section 3 and discussed in Section 4, while Section 5 consists of the overall concluding remarks.

2. Data and Methods

Data from Google Trends are downloaded online in '.csv' format and are normalized over the selected period. Google reports the adjustment procedure as follows: "*Search results are proportionate to the time and location of a query: Each data point is divided by the total searches of the geography and time range it represents, to compare relative popularity. Otherwise places with the most search volume would always be ranked highest. The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics. Different regions that show the same number of searches for a term will not always have the same total search volumes*" [28]. Depending on the retrieval time, data may slightly vary.

In this study, the examined period is from 1 January 2004 to 31 August 2017. The retrieved normalized data are monthly, with the number of observations being N = 164 (months), i.e., 13 years × 12 months + 8 months for each examined dataset. The datasets are 59 in total, i.e., 28 countries (English term) + 26 countries (translated terms excluding Ireland and the UK) + 5 Worldwide. The examined keywords are the English term 'Measles' and the respective translated terms, retrieved separately (independent searches, not comparisons) for each of the examined terms and for each of the examined EU countries. For the Worldwide assessment, the keywords 'Measles', 'Mumps', 'Rubella', 'MMR', and 'Anti Vaccine' were used.

The selected countries are the EU28 (translation of the term 'Measles' in the respective language in the parenthesis, obtained via Google Translate [29]): Austria (masern), Belgium (rougeole/mazelen), Bulgaria (дребна шарка), Croatia (ospice), Cyprus (ιλαρά), Czech Republic (spalničky), Denmark (mæslinger), Estonia (leetrid), Finland (rougeole), France (rougeole), Germany (Masern), Greece (ιλαρά), Hungary (kanyaró), Ireland (measles), Italy (morbillo), Latvia (masalas), Lithuania (tymai), Luxembourg (rougeole/mëllech), Malta (ħosba), Netherlands (mazelen), Poland (odra), Portugal (sarampo), Romania (pojar), Slovakia (osýpky), Slovenia (ošpice), Spain (sarampión), Sweden (mässling), and the UK (measles).

Google Trends is not case-sensitive, but it does take into account accents. Thus, for each country that is not English speaking and the respective term for Measles contains accents or any letter variations, relevant differentiation of terms were compared, and the term with the most search volumes was selected. Only Spain, Greece, and Cyprus exhibited highest interest in the respective translated term without accents, i.e., sarampion, ιλαρα, and ιλαρα, respectively; thus, the latter were used for the analysis. For the rest of the countries, the terms are used exactly as written above. Furthermore, the English term's online interest in each of the EU28 countries was examined, as the term 'Measles' exhibited high search volumes in many countries, and could, therefore, not be excluded from the analysis.

In order to analyze the interest in the Anti-Vaccine Movement, the following terms were compared: 'anti vaxx', 'anti-vaxx', 'anti vacc', 'anti-vacc', 'anti vax', 'anti-vax', 'anti vaccine', and 'anti-vaccine'. For the analysis to be robust, the term should be carefully selected; thus, as the term with significantly higher search volumes was 'Anti Vaccine', it is the one used in order to assess the Worldwide interest in the Anti-Vaccine Movement. For this study, the analysis consists of the following steps:

(a) Assessment of the Worldwide changes in the online interest in the the terms 'MMR' and the repsective diseases, i.e., 'Measles', 'Mumps', and 'Rubella',
(b) Analysis of the online interest in the English and the respective translated terms for 'Measles' for the selected period in all EU28 countries,
(c) Concise presentation and analysis of the 1st and 2nd MMR doses immunization percentages in Europe and the EU28,
(d) Examining of the relationships between online activity, vaccine population coverage (obtained through the WHO website [30]), and reported cases of Measles in each of the EU28 countries, by calculating the Pearson Correlation coefficients.

Data on country immunization percentages and reported Measles cases are obtained by the World Health Organization (WHO) [30], and defined by WHO as "*laboratory confirmed, epidemiologically linked, and clinical cases as reported to the World Health Organization*" [31].

3. Results

This section consists of the analysis of the Worldwide and the EU28 countries' online interest in terms related to Measles and the Anti-Vaccine Movement, followed by the examining of the relationships between Google Trends' data, vaccine population coverage, and reported cases.

3.1. Worldwide Online Interest

Figure 1 depicts the online interest in the terms 'Measles', 'Mumps', 'Rubella', and 'MMR' from 1 January 2004 to 31 August 2017. Note that for the term 'MMR', the results may be increased at points due to the same abbreviation shared with online gamers, though this does not affect the results, as the peaks and overall interest variations of the 'MMR' term are similar to the aforementioned diseases.

The related queries from 2004 to 2017 for the term 'Measles' include 'measles symptoms' (100), 'measles vaccine' (75), 'measles rash' (75), 'measles outbreak' (45), 'symptoms of measles' (30), 'mmr' (25), 'measles vaccination' (20), 'vaccination' (20), and 'measles treatment' (20). For the term 'Mumps', the related queries for the same period include 'symptoms mumps' (100), 'measles mumps' (65), 'mumps vaccine' (35), 'mumps adults' (25), 'mumps treatment' (20), 'mmr' (20), 'mumps disease' (15), and 'mumps outbreak' (15). For Rubella, the related queries include 'measles rubella' (100), 'rubella vaccine' (85), 'rubella virus' (70), 'rubella pregnancy' (50), 'what is rubella' (40), 'rubella rash' (35), 'rubella symptoms' (35), 'rubella test' (30), 'mmr' (25), 'measles rubella vaccine' (20), 'congenital rubella' (20), 'rubella syndrome' (20), and 'rubella in pregnancy' (20). For the MMR vaccine, the related queries from 2004 to 2017 include 'mmr vaccine' (100), 'autism' (20), 'what is mmr' (20), 'autism mmr' (20), 'mmr side effects' (15), 'mmr vaccination' (10), 'vaccination' (10), 'check mmr' (10), 'vaccines' (10), 'mmr vaccines' (10), 'mmr vaccine autism' (10), and 'mmr shot' (10).

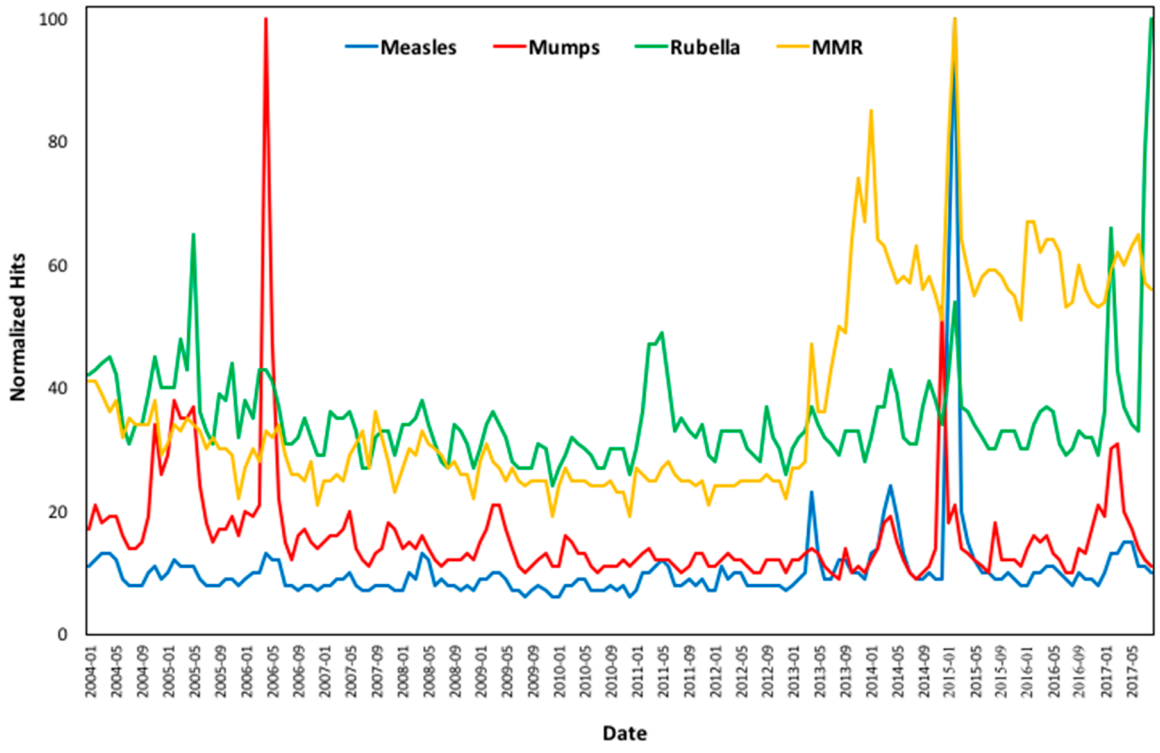


Figure 1. Worldwide Interest in 'Measles', 'Mumps', 'Rubella', and 'MMR' from 2004 to 2017.

All four examined terms exhibit similar behavior during the examined period, i.e., from January 2004 to August 2017. The interest peaks at several points during this time, while increased interest is evident in January 2015. This can be attributed to the 2015 Measles outbreak in Disneyland with 32 confirmed cases, most of them regarding unvaccinated people [32,33].

Figures 2–5 depict the Worldwide interest by country over the examined period for the terms Measles, Mumps, Rubella, and MMR, respectively. Note that the gray color indicates no significant results in search volumes, i.e., the score is zero.

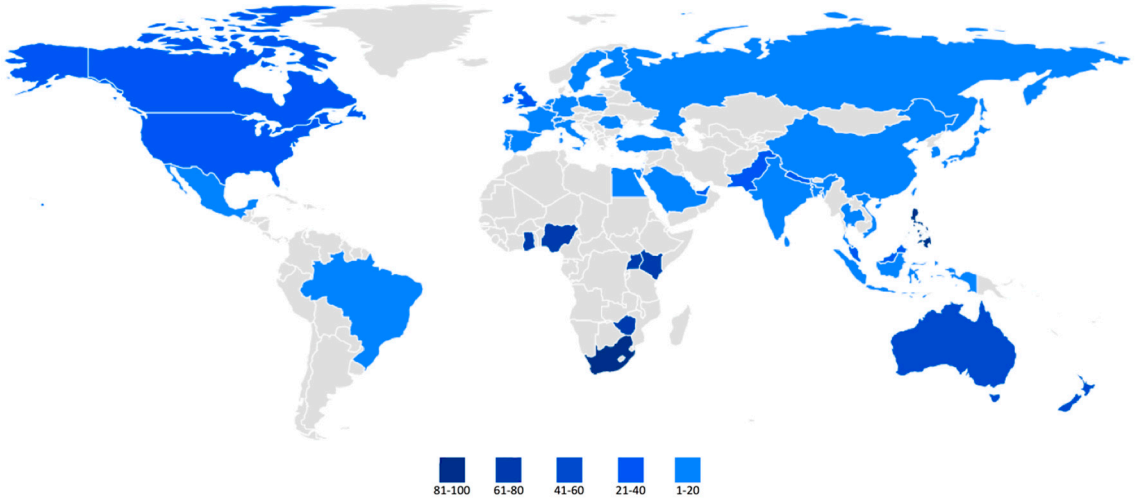


Figure 2. Worldwide Interest by Country in Measles from 2004 to 2017 (gray indicates zero scoring).

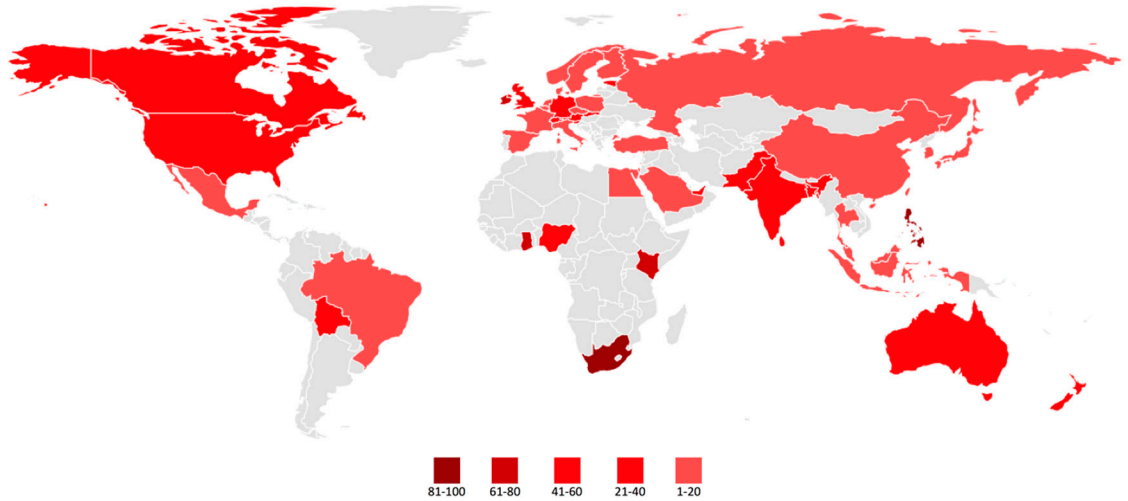


Figure 3. Worldwide Interest by Country in Mumps from 2004 to 2017 (gray indicates zero scoring).

In all terms, moderate interest is exhibited in Europe, Asia, and Northern America. However, the highest interest is observed in some countries in Africa, despite the fact that most countries have zero scoring, i.e., the search volumes are not high enough to be examined. The following African countries show constant interest in the three diseases: South Africa scores 89 in Measles, 100 in Mumps, 19 in Rubella, and 12 in MMR. Kenya exhibits high interest in Measles (67) and Mumps (72), and lower in Rubella (21); Ghana also exhibits higher interest in Measles (70) and Mumps (70), and lower in Rubella (22); Nigeria scores 67 in Measles, 53 in Mumps, and 11 in Rubella. Note that South Africa is the only country in the continent to have search volumes for the term 'MMR' high enough to be analyzed.

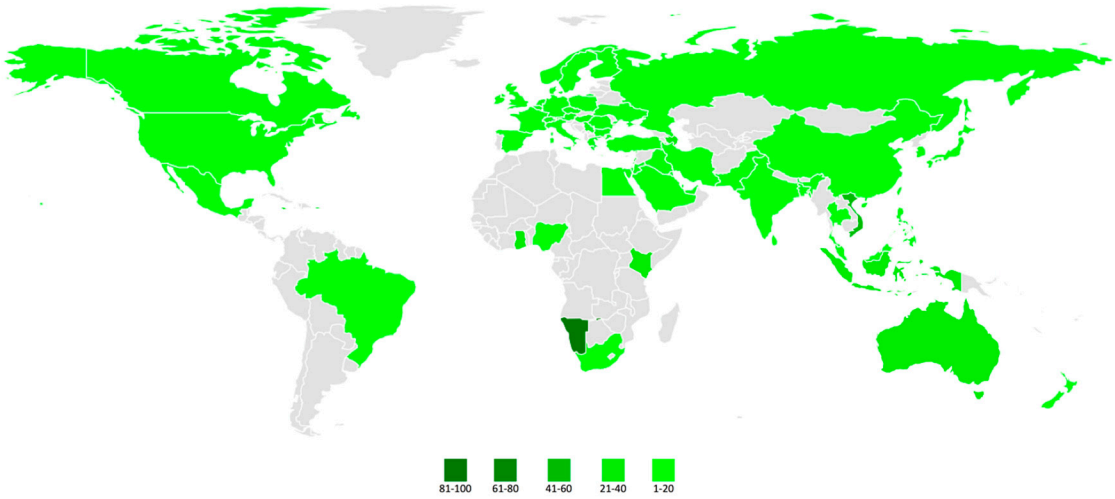


Figure 4. Worldwide Interest by Country in Rubella from 2004 to 2017 (gray indicates zero scoring).

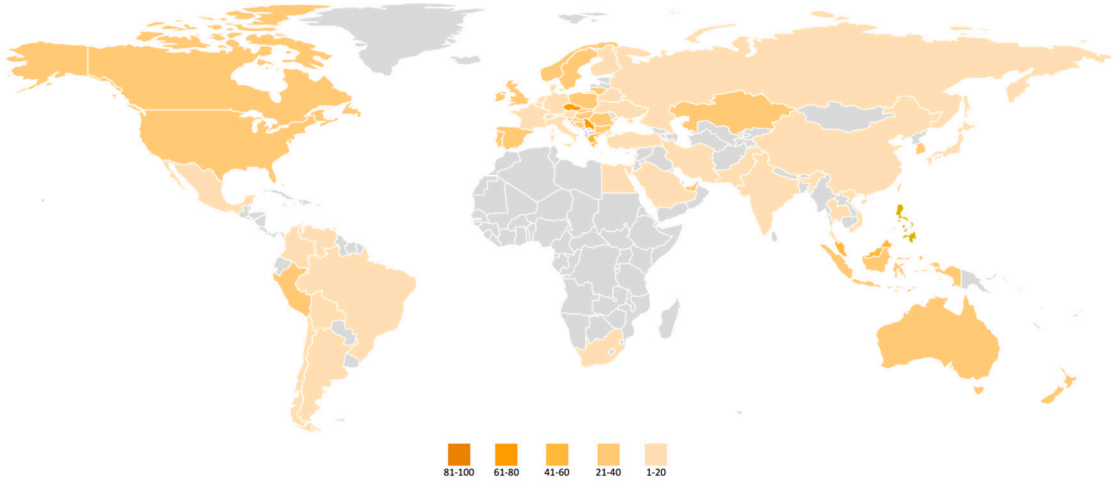


Figure 5. Worldwide Interest by Country in MMR from 2004 to 2017 (gray indicates zero scoring).

Figure 6 depicts the changes in the online interest in the term 'Anti Vaccine' from 1 January 2004 to 31 August 2017. The country with the highest search volumes is Canada (100), followed by Australia (96) and USA (93). The related queries include 'anti vaccine movement' (100), 'anti vaccination' (30), 'why anti vaccine' (20), 'measles' (15), 'anti vaccine doctors' (15), 'anti vaccine arguments' (15), 'anti vaccine celebrities' (10), 'anti vaccine websites' (5), 'andrew wakefield' (5), and 'measles outbreak' (5).

The online interest in the Anti-Vaccine Movement is rising, as all the more Internet users look for online information about anti-vaccination. What should be noted at this point is that the public searched for the terms 'anti vaccine arguments' and 'anti vaccine celebrities' in large volumes, which could be a worrying statement about how people choose to inform themselves in such crucial matters for public health.

A peak is observed in January 2010, which coincides with Wakefield losing his license [5], while the peak over the whole examined period, i.e., from 2004 to 2017, is observed in 2015, which could be attributed to the Measles outbreak in Disneyland [33]. This peak is during the same time that the online interest for the terms 'Measles', 'Mumps', 'Rubella', and 'MMR' also peaks (Figure 1). Overall, as depicted in Figure 6, the online interest for the term 'Anti Vaccine' has significantly increased over the past 13 years, with the average interest in 2017 being more than 10 times higher than what it was in 2004.

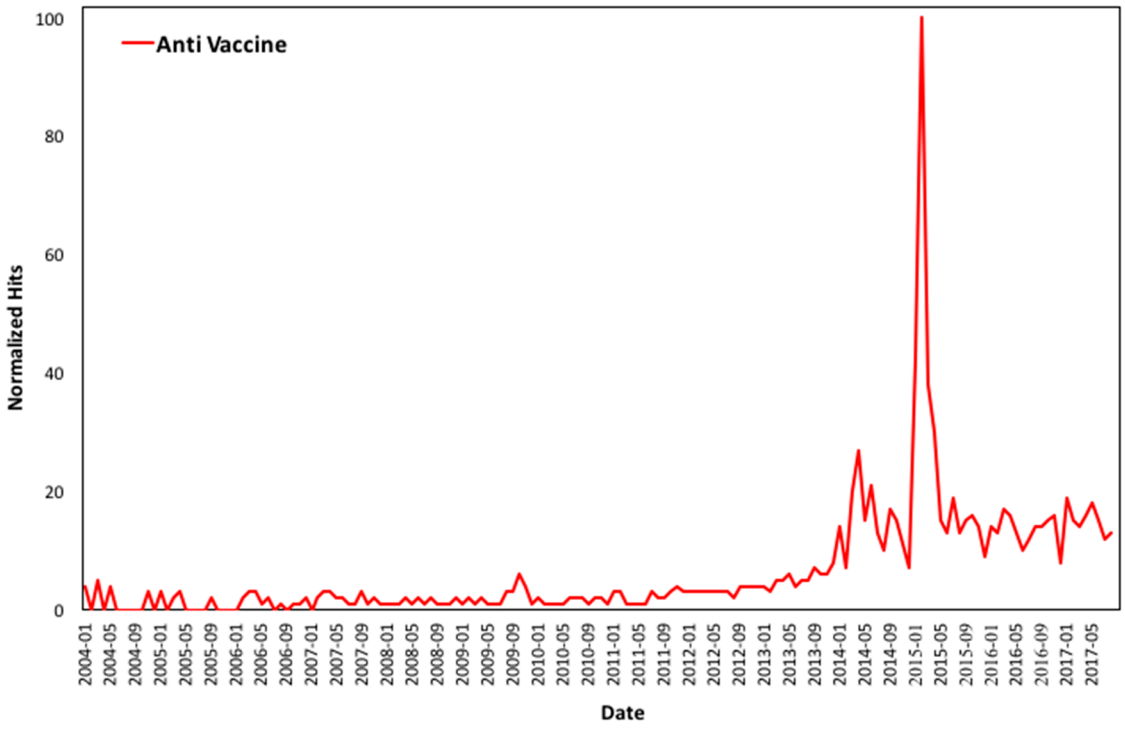


Figure 6. Worldwide Online Interest in the term 'Anti Vaccine' from January 2004 to August 2017.

In order to explore the links between the Anti-Vaccine Movement and the immunization percentages Worldwide, the Pearson Correlation coefficients are calculated. The yearly averages for the normalized Google queries for the term 'Anti Vaccine' (Worldwide) and the global population coverage of the 2nd dose of the vaccine for Measles from 1 January 2004 to 31 December 2016 exhibit a high negative correlation ($r = -0.7627$, $p < 0.01$).

For the years 2004–2015 and 2004–2014, Google queries and the 2nd dose population coverage are also highly (negatively) correlated ($r = -0.71$ with $p < 0.01$, and $r = -0.7076$ with $p < 0.05$, respectively). This indicates that the immunization coverage decreases as the online interest in the term 'Anti Vaccine' increases. Though statistically significant differences are not observed for the periods from 2004 to 2013 and from 2004 to 2012, the relationship is still negative. Correlations between the population coverage for the 1st dose of the Measles vaccine and the Worldwide online interest in the term 'Anti Vaccine' were not observed, which could be attributed to the time gap for the suggested age between the 1st and 2nd dose of the Measles vaccine.

3.2. EU28 Online Interest

Figure 7 depicts the monthly normalized (measured in a scale from 0 to 100) Google Trends' data in the English term 'Measles' (blue) and its respective translations (red) in all EU28 countries from January 2004 to August 2017 (independent searches, not comparisons).

Most countries exhibit increased interest in early 2015, thus supporting the argument that the Disneyland Measles outbreak in 2015 [33] affected Google searches for said disease. The countries that have increased and shown consistent interest throughout the examined period, i.e., that do not have many zeros, for either the English or the translated term, include Austria, Belgium, Denmark, Finland, France, Germany, Ireland, Italy, the Netherlands, Poland, Romania, Spain, Sweden, and the UK.

Figure 7. *Cont.*

Figure 7. *Cont.*

Figure 7. *Cont.*

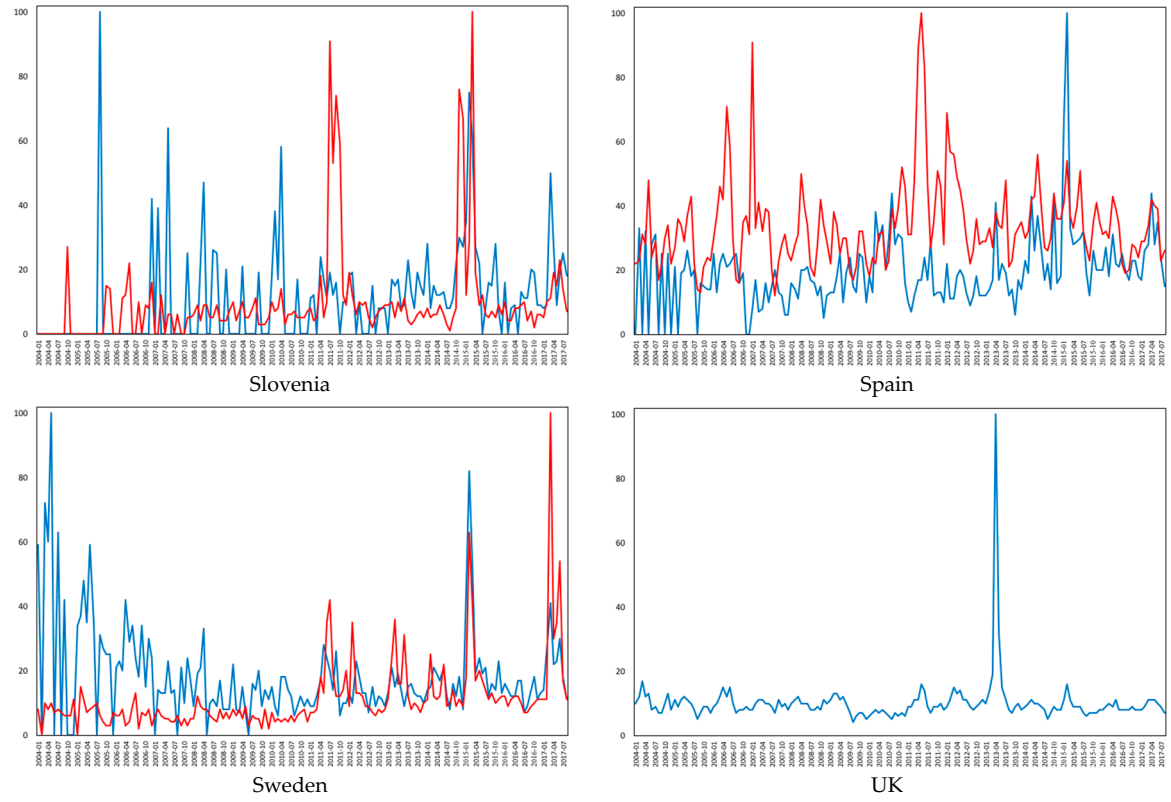


Figure 7. EU28 Online Interest in the English (blue) and Translated (red) Terms for 'Measles' from January 2004 to August 2017.

3.3. Measles EU Immunization Coverage

Figure 8 depicts the average of the population coverage percentages in the EU28 from 1980 to 2016 for the 1st dose of the Measles vaccine, and the average of the population coverage percentages for the 2nd dose of the Measles vaccine from 2000 to 2016. For the 1st dose, data for all EU28 countries from 1994 to 2017 are available, while datasets are not complete from 1980 to 1993. For the 2nd dose, data are only partly available for Belgium, Cyprus, Finland, France, Greece Italy, and Luxembourg, and no data are available for Ireland, while for the remaining EU28 countries full datasets are available.

The average percentage coverage has significantly increased since 1980, though it has experienced a drop in the past few years. The overall peaks in population coverage in Europe since 1980 are in 2013 and 2012 for the 1st and 2nd dose, respectively. In the EU28, the respective peaks are in 2013 and 2012. The percentages of the 2nd dose of the Measles vaccine in the EU28 are decreasing, even dropping below 90% in 2016. In order to be fully immunized, both doses of the Measles vaccine are required. As only 88.96% and 89.48% of the population in Europe and the EU28, respectively, are immunized, the current Measles outbreak can be explained, while the fear of an epidemic is justified.

Table 1 consists of the average percentage coverages of Europe and the EU28 for the 1st and the 2nd dose of the Measles vaccine from 2000 to 2016 [30].

Figures 9 and 10 map the EU28 population coverage for the Measles vaccine for the 1st and 2nd dose in 2016, respectively. Note that no data are available for Ireland for the 2nd dose.

For the 1st dose, Italy and Romania exhibit very low population coverage percentages, i.e., below 90%, while only 12 out of the 28 EU countries are above the 95% safety threshold. For the 2nd dose, the countries with the lowest immunization percentages in the Measles vaccine, i.e., less than 80%, are France and Romania, with Greece and Italy closely following similar attitude towards said vaccine, with immunization percentages of 83% in both countries. In the EU28, only Croatia, Hungary, and Slovakia are above the 95% safety threshold recommended by WHO.

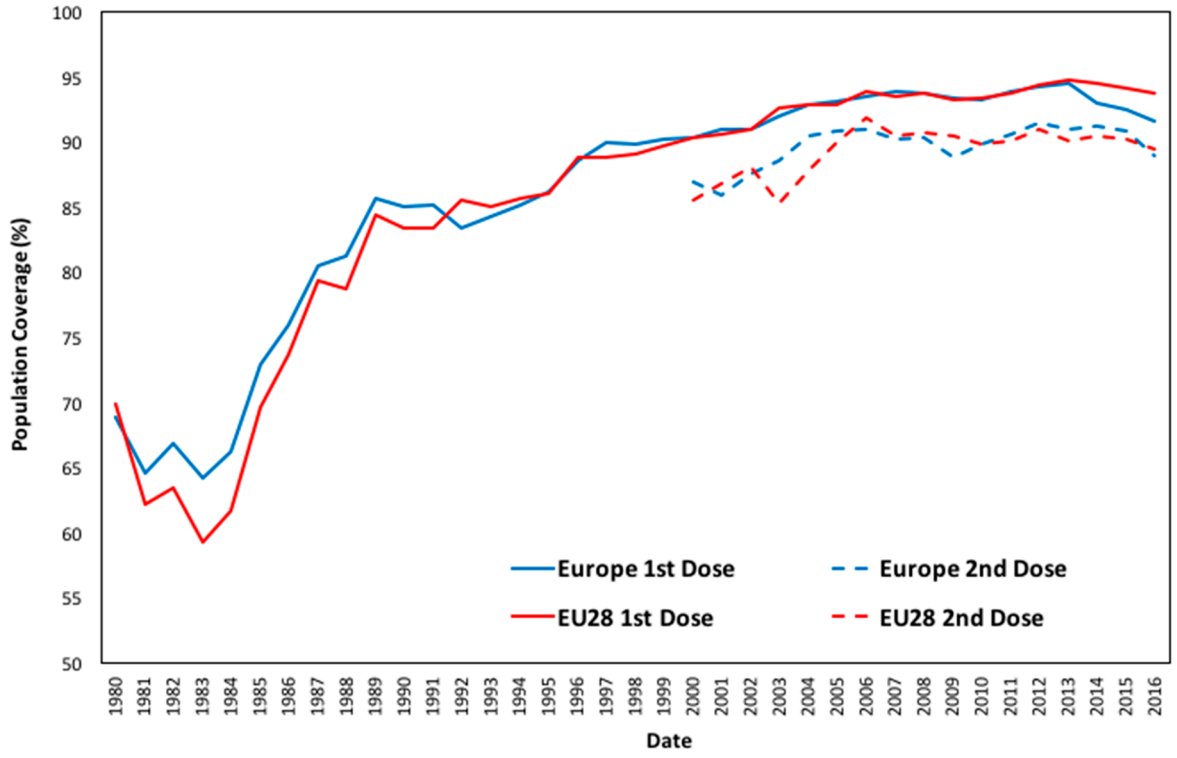


Figure 8. EU28 Population Coverage (%) of the 1st and 2nd Dose of the Vaccine for Measles from 1980 to 2016.

Table 1. Europe and EU28 Average Population Coverage Percentages (%) from 2000 to 2016.

Year	Europe 1st Dose	Europe 2nd Dose	EU 1st Dose	EU 2nd Dose
2000	90.40	87.00	90.43	85.57
2001	91.02	85.91	90.68	86.88
2002	91.04	87.62	91.07	88.11
2003	92.06	88.68	92.68	85.37
2004	92.87	90.51	92.89	87.85
2005	93.12	90.95	92.89	90.00
2006	93.51	91.02	93.93	91.86
2007	93.98	90.30	93.57	90.48
2008	93.77	90.45	93.75	90.74
2009	93.38	88.94	93.36	90.57
2010	93.28	89.83	93.39	89.88
2011	93.94	90.63	93.82	90.13
2012	94.34	91.48	94.50	91.08
2013	94.51	91.08	94.86	90.15
2014	93.06	91.30	94.61	90.46
2015	92.58	90.96	94.18	90.27
2016	91.64	88.96	93.86	89.48

France and Greece exhibit the highest rates of vaccine skepticism in the EU28, with Romania and Italy being in the 5th and 7th place, respectively [34]. As is depicted in Figure 10, these countries are the four countries with the lowest 2nd dose immunization percentages. Note that Romania poses a special case, where there are large populations of Romani people not being vaccinated, which was, however, the case in the past as well.

The EU countries are at an eminent risk of a Measles epidemic given the low immunization percentages and the high number of reported Measles cases, as has been supported by actions taken by several EU countries. For example, the Italian government has issued new legislation

making vaccination mandatory, with parents of unvaccinated children facing fines [35]. In Germany, kindergarten administrators have to report parents that refuse to be advised by doctors about vaccination [36], while Romania is preparing to issue a similar mandatory vaccination law [37].

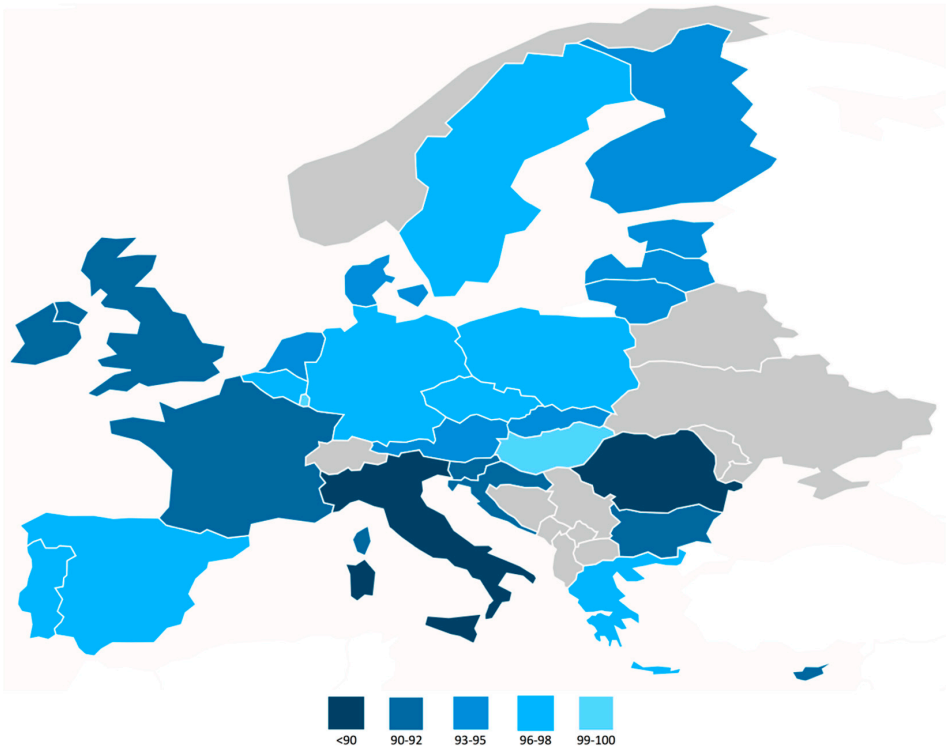


Figure 9. EU28 Population Coverage (%) for the 1st Dose in 2016.

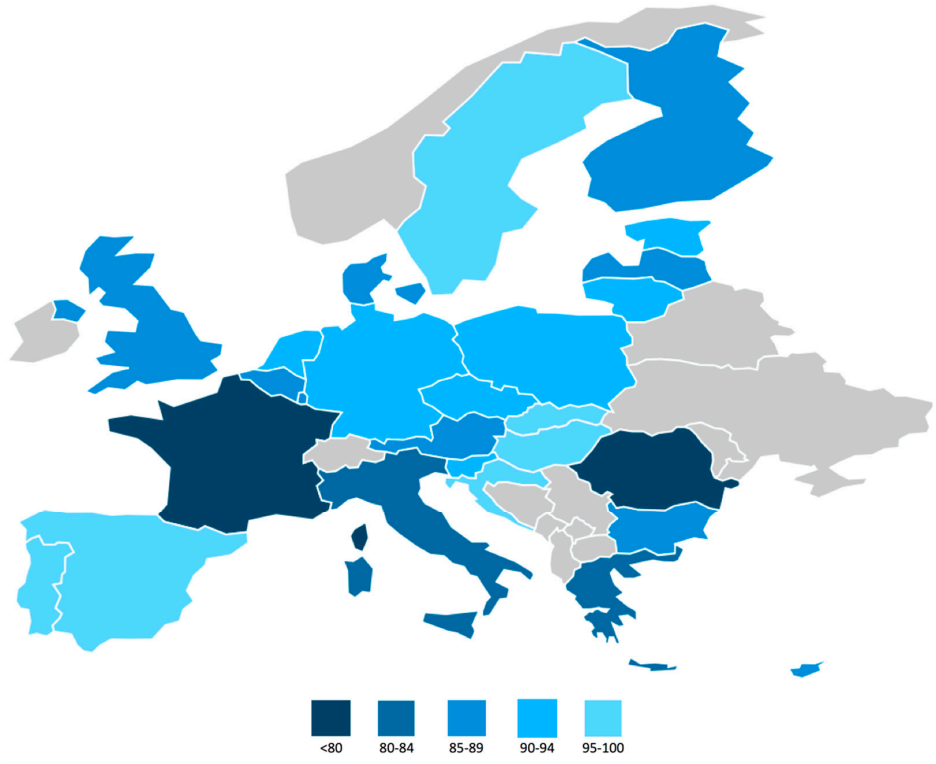


Figure 10. EU28 Population Coverage (%) for the 2nd Dose in 2016.

3.4. Confirmed EU Measles Cases

Table 2 consists of the total confirmed Measles cases from 2011 to (September) 2017 in the EU28 (by country code). In 2017, the countries with the most cases of Measles are Italy with 4204, followed by Romania with 3117, Germany with 796, Belgium with 359, and France with 352. Portugal, Hungary, Greece, Slovakia, Cyprus, and Luxembourg all had zero (0) cases in 2016, while they had several reported cases of Measles in 2017. Bulgaria had a high increase in reported cases (from 1 to 166), while Belgium and Czechia experienced significantly increased reporting of Measles cases, from 5 to 359 and from 7 to 134, respectively. Ireland, Lithuania, Poland, and the UK had significantly less reported Measles cases in 2017 compared to 2016, while Latvia and Malta had no cases in both 2016 and 2017.

Table 2. Total Confirmed Measles Cases in the EU28 from 2011 to 2017 [30].

Code	2011	2012	2013	2014	2015	2016	2017	Code	2011	2012	2013	2014	2015	2016	2017
AUT	219	35	79	119	300	28	80	ITA	5189	607	2256	3286	265	863	4204
BEL	1163	101	46	75	126	5	359	LVA	0	3	0	36	0	0	0
BGR	155	1	14	0	0	1	166	LTU	7	2	35	11	51	22	2
HRV	7	3	1	17	218	4	7	LUX	6	2	0	1	0	0	3
CYP	0	1	0	10	0	0	3	MLT	4	0	1	0	1	0	0
CZE	18	21	15	223	8	7	134	NLD	95	10	2640	144	7	6	11
DNK	84	2	17	29	9	3	1	POL	38	73	83	109	48	138	28
EST	7	4	2	0	4	2	1	PRT	2	7	1	0	0	0	34
FIN	29	4	2	2	1	5	5	ROU	4170	6164	1158	59	8	2432	3117
FRA	15,214	859	272	267	373	79	352	SVK	2	1	0	0	0	0	4
DEU	1600	167	1781	525	2383	328	796	SVN	22	2	1	52	19	1	6
GRC	40	3	2	1	1	0	11	ESP	3508	1210	131	153	55	38	141
HUN	0	1	1	0	0	0	25	SWE	26	31	51	26	22	3	24
IRL	193	104	51	35	13	43	12	GBR	1083	1903	1900	137	92	571	112

The 25% of the EU28 countries (7 in total) that score lowest in vaccination trust are France, Greece, Belgium, Romania, Slovenia, Bulgaria, and Italy [34]. All seven countries have reported increased cases of Measles in 2017 compared to 2016. At this point, it is interesting to note that all four countries in the EU28 that had a 2nd dose vaccination coverage in 2016 less than 85% (France, Romania, Italy, and Greece) are included in the list with the seven most vaccine-skeptical EU countries. The rest of the countries included in the list, namely Slovenia, Bulgaria, and Belgium, have a 2016 2nd dose population coverage of 93%, 88%, and 85%, respectively, all lower than the 95% safety threshold.

In order to explore the relationship between reported Measles cases and Google queries, the Pearson correlations for monthly data from January 2011 to August 2017 (N = 78) are calculated. For the respective translated terms, statistically significant positive correlations are observed for most of the EU28 countries, i.e., in Austria ($r = 0.4783$, $p < 0.01$), Belgium ($r = 0.5604$, $p < 0.01$), Croatia ($r = 0.65$, $p < 0.01$), Czechia ($r = 0.7410$, $p < 0.01$), Finland ($r = 0.7332$, $p < 0.01$), France ($r = 0.8908$, $p < 0.01$), Germany ($r = 0.5730$, $p < 0.01$), Italy ($r = 0.5555$, $p < 0.01$), Latvia ($r = 0.6253$, $p < 0.01$), Lithuania ($r = 0.6429$, $p < 0.01$), Netherlands ($r = 0.8725$, $p < 0.01$), Portugal ($r = 0.8508$, $p < 0.01$), Romania ($r = 0.4884$, $p < 0.01$), Slovakia ($r = 0.5997$, $p < 0.01$), Slovenia ($r = 0.5890$, $p < 0.01$), Spain ($r = 0.7734$, $p < 0.01$), Sweden ($r = 0.3459$, $p < 0.01$), Denmark ($r = 0.2418$, $p < 0.05$), Estonia ($r = 0.2433$, $p < 0.05$), Luxembourg ($r = 0.2553$, $p < 0.05$), and Hungary ($r = 0.2213$, $p < 0.10$).

Statistically significant positive correlations were also observed between the online interest in the English term and monthly reported Measels cases from January 2011 to August 2017 in several EU28 countries, namely in Austria ($r = 0.2594$, $p < 0.05$), Croatia ($r = 0.5825$, $p < 0.01$), Czechia ($r = 0.3112$, $p < 0.01$), Germany ($r = 0.5041$, $p < 0.01$), Ireland ($r = 0.3580$, $p < 0.01$), Italy ($r = 0.2368$, 0.05), Portugal ($r = 0.6211$, $p < 0.01$), Slovakia ($r = 0.2215$, $p < 0.10$), Slovenia ($r = 0.3348$, $p < 0.01$), and the UK ($r = 0.6307$, $p < 0.01$).

4. Discussion

This study's first aim was to track the 2017 EU Measles outbreak using online search traffic data from Google Trends. Given the rise of the Anti-Vaccine Movement over the past years that could be

attributed to false evidence that the MMR vaccine is associated with autism, the relations between Google queries related to anti vaccination and the recent outbreak in Measles are explored.

The results of this study suggest that there is a relation between the online interest in the Anti-Vaccine Movement and the decrease in vaccination percentages, and that the online queries for the term 'Measles' are potively correlated with Measles reported cases in most EU28 countries. Previous work has also suggested similar relationships between online data and reported cases of disease epidemics or outbreaks. Though this study considered data up to fall 2017, the serious issue of the EU Measles outbreak continues to exist and is showing increasing trends. In October 2017, the European Centre for Disease Prevention and Control stated that Measles could be further spread in Europe [38]. Greece faces a serious issue of Measles outbreak, with a total of 215 cases reported from 17 May 2017 to 1 October 2017 [38], while two children were admitted in the ICU with Measles-related complications [39]. Since the beginning of 2017, Italy has reported 4617 cases, four of which resulted to death, while in Romania, 34 deaths have been reported since January 2017 [38].

Measles require the highest immunization percentage compared to the other vaccine preventable diseases [9], and are one of the main causes of infant mortality that could have been vaccine-prevented [40]. The EU countries are facing an outbreak, which could result in an epidemic if the immunization percentages do not increase. Implications will also be evident from an economic point of view, as the treatment for Measles is higher than the cost of vaccination. Finally, other 'forgotten' diseases could soon resurface due to the Anti-Vaccine Movement, despite the scientists' and health officials' 'cry' for vaccination. As one to two out of 1000 diagnosed Measles cases in children result to death [41], in the case of an epidemic the casualties will be numerous.

As indicated by the results, online search traffic data could be proven a valid and valuable data source for governments and health officials for the monitoring of the behavior towards Measles and the Anti-Vaccine Movement. An interesting factor to be examined would be the degree of association of the Anti-Vaccine Movement with the overall recent political, social, and economic changes occurring in the EU at the moment. It has been suggested that advocacy and communication play a significant role in increasing Measles vaccination [8], while the measure of mandatory vaccinations is also considered or already enforced in several European countries [42]. However, governmental populism can negatively affect measures that should be taken in order to prevent disease spreading [43], while websites with available information on '*vaccine myths*' and anti-vaccination are more than the ones discussing the benefits of vaccination [44]. All the above add to the important factor influencing individuals to dismiss information about the positive effects of vaccines is that the communication of scientific issues is negatively affected by conspiracist ideation [7].

This study has some limitations. At first, the Measles outbreak practically occurred in 2017, with an increasing trend after September, which is out of this study's examined time-frame. Furthermore, the sample is not representative, as not all queries for the term 'Measles' correspond to reported cases and vice versa, and not all Anti-Vaccine searches correspond to not vaccinating and vice versa. Despite that, empirical correlations between Google Trends and official health data in various topics have been previously shown to exist. The 'Anti Vaccine' term is highly correlated ($p < 0.01$) with the Worldwide population coverage of the 2nd dose of the Measles vaccine; thus, future research on the subject could explore and further elaborate on the relationship between the Anti-Vaccine Movement and the decrease of the immunization percentages in Measles and other vaccine preventable diseases. Towards this direction, the relationships between the online interest in an extended list of English and translated terms related to the Anti-Vaccine Movement, the countries' immunization percentages, and reported cases of Measles, as well as other vaccine preventable diseases for individual countries, could be investigated.

The decrease in immunization percentages for Measles is a serious issue that negatively affects public health, while the impact could have been foreseen with the monitoring of the online behavior towards the Anti-Vaccine Movement over the last years. Exploring patterns of available information —that are related to population health as suggested by the science of infodemiology [45]—is increasingly

employed to effectively deal with topics related to public health. Epidemics and outbreaks may not exhibit seasonality, cycles, or long-term patterns, and thus the commonly used statistical tools and methods of analyzing seasonal diseases, such as asthma of the flu, may not be appropriate. Therefore, it is essential to effectively visualize these large amounts of data in order to explore and detect the trends and variations in interest over time, identify underlying patterns, and relate peaks to real-life events. This is highlighted by the results of this study that suggest that monitoring the changes in the online interest in selected terms could provide valuable information in behavioral variations and patterns. Said patterns could assist with the analysis of human behavior in public health issues, all the while providing health officials with valuable information to assess these issues and take preventive measures.

5. Conclusions

In the era of online information overload, can the use of the Internet affect public health? To address this question, this study aimed at tracking the 2017 EU Measles outbreak, by analyzing the online behavioral variations in terms related to Measles and the Anti-Vaccine Movement. The results suggest that statistically significant positive correlations exist between the monthly reported cases of Measles and the online interest in the respective translated term in most of the EU28 countries from January 2011 to August 2017. Furthermore, the term 'Anti Vaccine' is highly negatively correlated with the Worldwide immunization percentages from 2004 to 2016, i.e., as the online interest in the term 'Anti Vaccine' increases, the immunization percentages decrease.

This finding is supportive of previous research suggesting that conspiracist ideation is related to the rejection of science, as the negative relationship between online interest in the term 'Anti Vaccine' and the immunization percentages could be indicative of the role that the Internet plays in the spreading of false information, consequently affecting public health. In the case of Measles, the results are now starting to show, with reported Measles cases taking a sudden upturn over the past year in the EU, as immunization percentages (2nd dose) have significantly dropped since 2012—in most countries below the 95% safety threshold.

During the past few years, Big Data Analytics in general and the analysis of Internet behavior in specific have been shown to be effective at assessing various public health topics, as it has been suggested that patterns of available information are related to population health. Measles could be just the first of many to follow to exhibit such increase in reported cases, given that Measles require the highest immunization percentage out of the vaccine preventable diseases. Therefore, continuous monitoring is required for nowcasting the new cases that occur daily in relation to the variations in online interest, in order for the respective countries' Health Care Systems to be prepared, and for health officials to deal with reported cases in a timely manner and take the appropriate preventive measures, especially in countries and regions of high risk.

Author Contributions: Amaryllis Mavragani collected the data, performed the analysis, and wrote the paper. Gabriela Ochoa had the overall supervision.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wakefield, A.J.; Murch, S.H.; Anthony, A.; Linnell, J.; Casson, D.M.; Malik, M.; Berelowitz, M.; Dhillon, A.P.; Thomson, M.A.; Harvey, P.; et al. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *Lancet* **1998**, *351*, 637–641. [CrossRef]
2. Andrew Wakefield Fights Back. Available online: sciencebasedmedicine.org/wakefield-fights-back/ (accessed on 9 November 2017).
3. Plotkin, S.; Gerber, J.S.; Offit, P.A. Vaccines and Autism: A Tale of Shifting Hypotheses. *Clin. Infect. Dis.* **2009**, *48*, 456–461.

4. Taylor, L.E.; Swerdfeger, A.L.; Eslick, G.D. Vaccines are not associated with autism: An evidence-basedmeta-analysis of case-control and cohort studies. *Vaccines* **2014**, *32*, 3623–3629. [CrossRef] [PubMed]
5. Fitness to Practise Panel Hearing. 28 January 2010. Available online: www.casewatch.org/foreign/wakefield/gmc_findings.pdf (accessed on 9 November 2017).
6. Lancet Retracts Wakefield Paper. Available online: https://www.autism-watch.org/news/lancet.shtml (accessed on 9 November 2017).
7. Lewandowsky, S.; Gignac, G.E.; Oberauer, K. The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science. *PLoS ONE* **2013**, *8*, e75637. [CrossRef] [PubMed]
8. Carrillo-Santisteve, P.; Lopalco, P.L. Measles still spreads in Europe: Who is responsible for the failure to vaccinate? *Clin. Microbiol. Infect.* **2012**, *18*, 50–56. [CrossRef] [PubMed]
9. Durrheim, D.N.; Crowcroft, N.S.; Strebel, P.M. Measles—The epidemiology of elimination. *Vaccine* **2014**, *32*, 6880–6883. [CrossRef] [PubMed]
10. Google Trends. Available online: trends.google.com/trends/explore (accessed on 9 November 2017).
11. Preis, T.; Moat, H.S.; Stanley, H.E.; Bishop, S.R. Quantifying the advantage of looking forward. *Sci. Rep.* **2012**, *2*, 350. [CrossRef] [PubMed]
12. Preis, T.; Moat, H.S.; Stanley, H.E. Quantifying trading behavior in financial markets using Google Trends. *Sci. Rep.* **2013**, *3*, 1684. [CrossRef] [PubMed]
13. Scharkow, M.; Vogelgesang, J. Measuring the public agenda using search engine queries. *Int. J. Public Opin. Res.* **2011**, *23*, 104–113. [CrossRef]
14. Burnap, P.; Rana, O.F.; Avis, N.; Williams, M.; Housley, W.; Edwards, A.; Morgan, J.; Sloan, L. Detecting tension in online communities with computational Twitter analysis. *Technol. Forecast. Soc. Chang.* **2015**, *95*, 96–108. [CrossRef]
15. McCallum, M.L.; Bury, G.W. Public interest in the environment is falling: A response to Ficetola (2013). *Biodivers. Conserv.* **2014**, *23*, 1057–1062. [CrossRef]
16. Jun, S.P.; Park, D.H.; Yeom, J. The possibility of using search traffic information to explore consumer product attitudes and forecast consumer preference. *Technol. Forecast. Soc. Chang.* **2014**, *86*, 237–253. [CrossRef]
17. Nuti, S.V.; Wayda, B.; Ranasinghei, I.; Wang, S.; Dreyer, R.P.; Chen, S.I.; Murugiah, K. The Use of Google Trends in Health Care Research: A Systematic Review. *PLoS ONE* **2014**, *9*, e109583.
18. Solano, P.; Ustulin, M.; Pizzorno, E.; Vichi, M.; Pompili, M.; Serafini, G.; Amore, M. A Google-based approach for monitoring suicide risk. *Psychiatry Res.* **2016**, *246*, 581–586. [CrossRef] [PubMed]
19. Arora, V.S.; Stuckler, D.; McKee, M. Tracking search engine queries for suicide in the United Kingdom, 2004–2013. *Public Health* **2016**, *137*, 147–153. [CrossRef] [PubMed]
20. Mavragani, A.; Sypsa, K.; Sampri, A.; Tsagarakis, KP. Quantifying the UK Online Interest in substances of the EU Watch List for Water Monitoring: Diclofenac, Estradiol, and the Macrolide Antibiotics. *Water* **2016**, *8*, 542. [CrossRef]
21. Gahr, M.; Uzelac, Z.; Zeiss, R.; Connemann, B.J.; Lang, D.; Schönfeldt-Lecuona, C. Linking annual prescription volume of antidepressants to corresponding web search query data: A possible proxy for medical prescription behavior? *J. Clin. Psychopharmacol.* **2015**, *35*, 681–685. [CrossRef] [PubMed]
22. Schuster, N.M.; Rogers, M.A.; McMahon, L.F., Jr. Using search engine query data to track pharmaceutical utilization: A study of statins. *Am. J. Manag. Care* **2010**, *16*, e215–e219. [PubMed]
23. Cho, S.; Sohn, C.H.; Jo, M.W.; Shin, S.Y.; Lee, J.H.; Ryoo, S.M.; Kim, W.Y.; Seo, D.W. Correlation between national influenza surveillance data and Google Trends in South Korea. *PLoS ONE* **2013**, *8*, e81422. [CrossRef] [PubMed]
24. Domnich, A.; Panatto, D.; Signori, A.; Lai, P.L.; Gasparini, R.; Amicizia, D. Age-related differences in the accuracy of web query-based predictions of Influenza-Like Illness. *PLoS ONE* **2015**, *10*, 0127754. [CrossRef] [PubMed]
25. Zhou, X.; Ye, J.; Feng, Y. Tuberculosis surveillance by analyzing google trends. *IEEE Trans. Biomed. Eng.* **2011**, *58*, 2247–2254. [CrossRef] [PubMed]
26. Alicino, C.; Bragazzi, N.L.; Faccio, V.; Amicizia, D.; Panatto, D.; Gasparini, R.; Icardi, G.; Orsi, A. Assessing Ebola-related web search behaviour: Insights and implications from an analytical study of Google Trends-based query volumes. *Infect. Dis. Poverty* **2015**, *4*, 54. [CrossRef] [PubMed]
27. Hossain, L.; Kam, D.; Kong, F.; Wigand, R.T.; Bossomaier, T. Social media in Ebola outbreak. *Epidemiol. Infect.* **2016**, *144*, 2136–2143. [CrossRef] [PubMed]

28. How Trends Data Is Adjusted. Google Trends. Available online: support.google.com/trends/answer/4365533?hl=en (accessed on 4 November 2017).
29. Google Translate. Available online: translate.google.com (accessed on 21 September 2017).
30. Immunization, Vaccines and Biologicals: Data, Statistics and Graphics. Available online: http://www.who.int/immunization/monitoring_surveillance/data/en/ (accessed on 9 November 2017).
31. Measles and Rubella Surveillance Data. World Health Organization. Available online: http://www.who.int/immunization/monitoring_surveillance/burden/vpd/surveillance_type/active/measles_monthlydata/en/ (accessed on 9 November 2017).
32. Measles Outbreak Worsens in US after Unvaccinated Woman Visits Disneyland. Available online: www.theguardian.com/us-news/2015/jan/14/measles-outbreak-spreads-unvaccinated-woman-disneyland (accessed on 9 November 2017).
33. Too Rich to Get Sick? Disneyland Measles Outbreak Reflects Anti-Vaccination Trend. Available online: www.theguardian.com/us-news/2015/jan/17/too-rich-sick-disneyland-measles-outbreak-reflects-anti-vaccination-trend (accessed on 9 November 2017).
34. Larson, H.J.; de Figueiredo, A.; Xiahong, Z.; Schulz, W.S.; Verger, P.; Johnston, I.G.; Cook, A.R.; Jones, N.S. The State of Vaccine Confidence 2016: Global Insights through a 67-Country Survey. *Ebiomedicine* **2016**, *12*, 295–301. [CrossRef] [PubMed]
35. Italy Makes Vaccination Mandatory for Children. Available online: http://www.dw.com/en/italy-makes-vaccination-mandatory-for-children/a-38911682 (accessed on 9 November 2017).
36. Germany Moves to Improve Child Vaccination Rate. Available online: http://www.dw.com/en/germany-moves-to-improve-child-vaccination-rate/a-39004792 (accessed on 9 November 2017).
37. Romania Measles Outbreak: Is Mandatory Vaccination The Answer? Available online: www.vaccinestoday.eu/stories/romania-measles-outbreak-mandatory-vaccination-answer/ (accessed on 9 November 2017).
38. Measles Outbreaks in Europe (Update 9). Available online: http://www.fitfortravel.nhs.uk/news/newsdetail.aspx?Id=22403 (accessed on 9 November 2017).
39. Two Children in the ICU Due to Measles. (In Greek). Available online: http://news.in.gr/greece/article/?aid=1500164587 (accessed on 9 November 2017).
40. Measles Data and Statistics. Centers for Disease Control and Prevention. Available online: www.cdc.gov/measles/downloads/measlesdataandstatsslideset.pdf (accessed on 9 November 2017).
41. Complications of Measles. Measles; Centers for Disease Control and Prevention. Available online: www.cdc.gov/measles/about/complications.html (accessed on 9 November 2017).
42. Compulsory Vaccination and Rates of Coverage Immunisation in Europe. Available online: http://www.asset-scienceinsociety.eu/reports/page1.html (accessed on 9 November 2017).
43. Michailidou, D.; Kennedy, J. When Populism Can Kill. Project Syndicate, 2017. Available online: www.project-syndicate.org/commentary/populism-anti-vaccine-movement-by-domna-michailidou-and-jonathan-kennedy-2017-07 (accessed on 9 November 2017).
44. Ruiz, J.B.; Bell, R.A. Understanding vaccination resistance: Vaccine search term selection bias and the valence of retrieved information. *Vaccine* **2014**, *32*, 5776–5780. [CrossRef] [PubMed]
45. Eysenbach, G. Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. *J. Med. Internet Res.* **2009**, *11*, e11. [CrossRef] [PubMed]

Mavragani and Ochoa *J Big Data* (2018) 5:17
https://doi.org/10.1186/s40537-018-0126-7

Journal of Big Data

RESEARCH Open Access

Forecasting AIDS prevalence in the United States using online search traffic data

Amaryllis Mavragani* and Gabriela Ochoa

*Correspondence:
amaryllis.mavragani1@stir.ac.uk
Department of Computing Science and Mathematics, Faculty of Natural Sciences, University of Stirling, Stirling FK9 4LA, UK

Abstract

Over the past decade and with the increasing use of the Internet, the assessment of health issues using online search traffic data has become an integral part of Health Informatics. Internet data in general and from Google Trends in particular have been shown to be valid and valuable in predictions, forecastings, and nowcastings; and in detecting, tracking, and monitoring diseases' outbreaks and epidemics. Empirical relationships have been shown to exist between Google Trends' data and official data in several health topics, with the science of infodemiology using the vast amount of information available online for the assessment of public health and policy matters. The aim of this study is to provide a method of forecasting AIDS prevalence in the US using online search traffic data from Google Trends on AIDS related terms. The results at first show that significant correlations between Google Trends' data and official health data on AIDS prevalence (2004–2015) exist in several States, while the estimated forecasting models for AIDS prevalence show that official health data and Google Trends data on AIDS follow a logarithmic relationship. Overall, the results of this study support previous work on the subject suggesting that Google data are valid and valuable for the analysis and forecasting of human behavior towards health topics, and could further assist with Health Assessment in the US and in other countries and regions with valid available official health data.

Keywords: AIDS, Big data, Forecasting, Google Trends, HIV, Internet, Online behavior

Introduction

Big Data, characterized by large volumes, high processing speed, and wide variety of datasets [1–3], have been shown to be very valuable in health care research, with Health Informatics being the field in which big data analytics have been extensively applied [4]. A popular way of addressing the challenge of Big Data is the analysis of online search traffic data [5, 6], mainly with data from Google Trends [7]. Over the past decade, this field of research, i.e., analyzing online search traffic data, has been widely used and is growing in popularity for assessing various topics, though it has mostly focused on the fields of Health and Medicine [8].

Many studies on the subject have empirically shown that Google Trends' data are related to public health data. Topics that have been explored up to this point include the analysis, assessment, and prediction of epidemics and outbreaks, as, for example, Ebola [9, 10], Measles [11], the Bed-Bug epidemic [12], and Tuberculosis [13]. A much studied

topic is that of influenza like illness (the flu), which is a seasonal disease and has shown well performing results in the past [14–17].

Recently, more topics on relating Google data with official health data have been visited, as in the case of suicide rates, where it has been show that Google queries can be used to monitor the risk of suicide [18, 19]. On a different direction, there has been shown that correlations exist between Google Trends data and prescription drugs issuing [20, 21] and revenues [22]. Apart from prescription drugs, focus has been given to illegal drugs as well, with notable examples including the tracking of dabbing in the US [23], Krokodil in Russia [24], and Methamphetamines in Central Europe [25].

According to Infodemiology [26], data available on the Internet can be used to inform public health and policy by monitoring the public's behavior towards diseases, selecting the relevant available information, as well as monitoring how the public reacts to health marketing campaigns. Though it is widely supported and evident that official health data and online search traffic data are correlated, the most important step towards health assessment using Google Trends is that of finding methods of predicting and nowcasting diseases' occurrence and outbreaks, as well as forecasting seasonal diseases' prevalence.

Though seasonality has been assessed in various cases, such as, for information on tobacco and lung cancer [27], the restless legs syndrome [28], and in sleep-disordered breathing [29], studies developing methods towards the direction of forecasting and nowcasting exhibit significantly lower numbers. Despite that, recent research has exhibited promising results in the forecasting of various diseases and outbreaks, as, for example, Tuberculosis [13], influenza like illness [17], pertussis [30], suicide risk [18], and dementia [31].

As Infodemiology data can be retrieved in real time and thus allow the nowcasting of human behavior based on Internet data, the detection, monitoring, and prediction of epidemics and outbreaks can be much assisted by the analysis of Google queries. A topic that is of high significance and interest is that of AIDS (Acquired Immune Deficiency Syndrome) and HIV (Human Immunodeficiency Virus). HIV is a virus that is mainly transmitted via sexual intercourse and needle/syringe use [32]. The treatment for HIV consists of the antiretroviral therapy, which controls the HIV virus. If the HIV remains without treatment, it affects the immune system, which worsens as time passes. The HIV infection consists of 3 stages: (1) acute HIV infection, (2) clinical latency, and (3) AIDS; the latter being the most severe stage of the HIV infection [33], which leads to an increased number of '*opportunistic infections*' [32].

People would more easily search for information online than consult a doctor in general. In the case of AIDS, as it is a sensitive subject, the anonymity provided by the Internet allows people to search for information online. Thus the monitoring of Internet data is essential in the overall assessment of AIDS prevalence in regions where Internet penetration is high, as in the case of the United States. Novel methods of assessment are needed, as data on 'AIDS Prevalence', 'AIDS Diagnoses', and 'AIDS Deaths' provided by the Centers for Disease Control and Prevention (CDC) are not available in real time, as gathering, analyzing and making these data available is a long process that takes over a year.

AIDS is categorized as an epidemic [34], and as such it needs constant assessment. The aim of this paper is to analyze the online interest in AIDS related terms and estimate

forecasting models for AIDS prevalence in the US using data from Google Trends. The rest of this paper is structured as follows: the "Research methodology" section consists of the procedure of the data collection and methodology followed to analyze and forecast AIDS prevalence, in the "Results" section the results of the analysis are presented, the "Discussion" section consists of the discussion of the analysis, while the "Conclusions" section consists of the overall conclusions and future research suggestions.

Research methodology

Data

Data from Google Trends are downloaded online in '.csv' format and are normalized over the selected time-frame as follows: "*Search results are proportionate to the time and location of a query: Each data point is divided by the total searches of the geography and time range it represents, to compare relative popularity. Otherwise places with the most search volume would always be ranked highest. The resulting numbers are then scaled on a range of 0–100 based on a topic's proportion to all searches on all topics. Different regions that show the same number of searches for a term will not always have the same total search volumes.*" [35]. Google Trends is not case-sensitive, though takes into account spelling errors and accents. In this study, this effect is minimized, as the examined term, i.e. AIDS, is universal, not translated, and difficult to misspell. Note that data may slightly vary when retrieved at different time points.

Methods

The choice of terms is crucial for the robustness of the results when using online data [36]. In Google Trends, the four options below are available when retrieving data for the examined disease. The term's online interest can be retrieved in the 'Search Term' form, i.e. include all queries that had the respective term, thereafter referred to as 'AIDS (Search Term)'. In addition, Google Trends groups related queries under other search terms as well, which in this case are 'AIDS (Illness)'. Finally, Google Trends also gives the option of including terms related to the topics of 'Management of AIDS/HIV (Topic)', and 'Diagnosis of HIV/AIDS (Topic)'.

Analysis stages

At first, an overall assessment of all four available terms and topics' variations in online interest is provided, so as to identify the option that would increase the validity of further analysis on the subject. The next step towards examining the possibility of forecasting AIDS prevalence and incidence, is to identify any existing correlations between Google data on related terms and topics and official health data for AIDS. In this study, data on 'AIDS Prevalence' (2004–2015) are retrieved by the CDC website [37]. Depending on the significance of the calculated Pearson correlations, the possibility of forecasting AIDS prevalence in the US will be assessed. Finally, forecasting models of AIDS prevalence based on Google Trends' data for the US as well as for each 50 States plus DC are estimated.

Results

At first, an overall assessment of the online interest towards AIDS in the US is performed, followed by the exploring of the correlations between AIDS prevalence and Google Trends data in the US and each US State individually. Finally, forecasting models for AIDS prevalence in the US are estimated, at both national and State level, so as to elaborate on the usefulness of the tool in health assessment in the US.

AIDS online interest in the US

Figure 1 consists of the changes in the online interest in 'AIDS (Search Terms)' and 'AIDS (Illness)' from January 2004 to December 2015, while Fig. 2 depicts the monthly normalized online interest in the 'Google Trends' topics of 'Diagnosis of HIV/AIDS' and 'Management of HIV/AIDS' from January 2004 to December 2015.

The top related queries for 'AIDS (Search Term)' include 'aids hiv' (100), 'hearing aids' (99), 'hiv' (97), 'aids symptoms' (33), 'aids and hiv' (25), 'aids day' (24), 'africa aids' (22), 'aids cure' (16), 'aids test' (11), 'aids statistics' (11), and 'aids virus' (10). For 'AIDS (Illness)', the top related queries include 'aids' (100), 'hiv' (26), 'aids hiv' (14), 'hiv/aids' (6), 'aids symptoms' (5), 'africa' (4), 'aids day' (4), 'hiv symptoms' (3), 'aids cure' (2), 'hiv infection' (2), 'hiv transmission' (2), and 'aids statistics' (2).

For the topic of 'Diagnosis of HIV/AIDS', the top related queries include 'hiv' (100), 'hiv test' (53), 'hiv testing' (50), 'free hiv testing' (13), 'test for hiv' (11), 'hiv symptoms' (9), 'hiv home test' (7), 'aids' (6), 'hiv aids' (6), 'hiv rapid test' (4), 'free hiv test' (4), 'hiv positive' (4), 'hiv test results' (4), 'positive hiv test' (3), 'rapid hiv testing' (3), 'hiv test kit' (3), and 'oraquick hiv test' (2). For the topic 'Management of HIV/AIDS', the top related queries include 'antiretroviral' (100), 'hiv' (86), 'aids' (59), 'antiretroviral therapy' (58), 'aids drugs' (38), 'antiretrovirals' (28), 'hiv treatment' (23), 'antiretroviral treatment' (22),

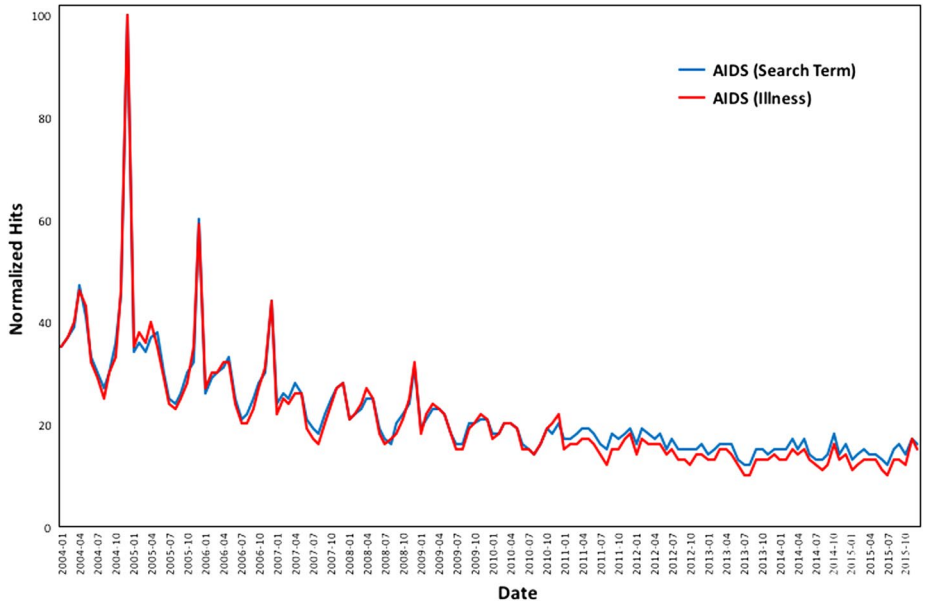


Fig. 1 Monthly Normalized Google Trends' Data for 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015

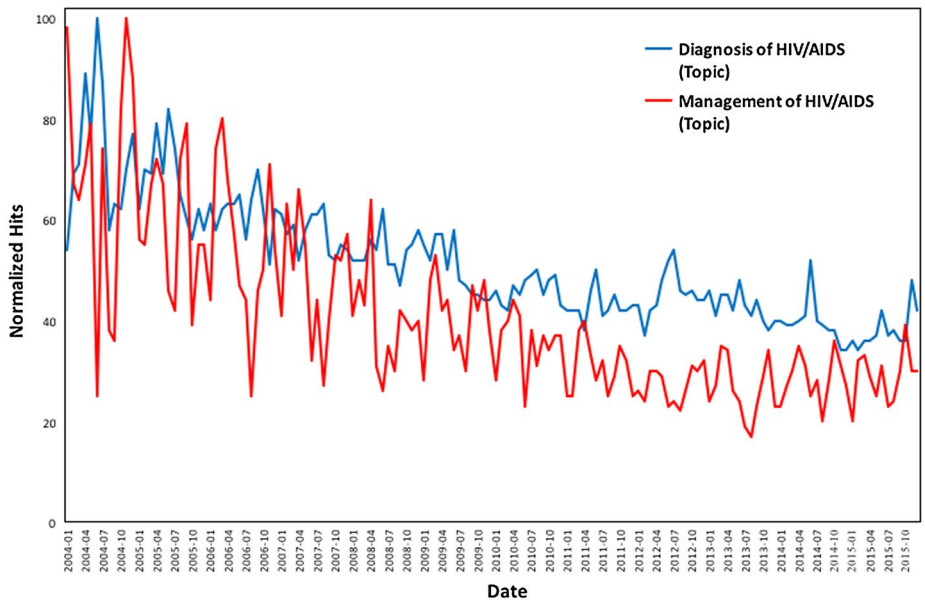


Fig. 2 Monthly normalized Google Trends' Data for 'Diagnosis of HIV/AIDS (Topic)' and 'Management of HIV/AIDS (Topic)' from January 2004 to December 2015

'hiv aids' (20), 'antiretroviral drugs' (16), 'hiv management' (12), 'highly active antiretroviral therapy' (7), and 'hiv medications' (4).

Figure 3 consists of the heat maps of the online interest by US State from January 2004 to December 2015 for 'AIDS (Search Term)', 'AIDS (Illness)', 'Diagnosis of HIV/AIDS (Topic)', and 'Management of HIV/AIDS (Topic)'.

It is evident that the terms related to AIDS exhibit high and constant interest from 2004 to 2015. The topics of 'Diagnosis of HIV/AIDS (Topic)' and 'Management of HIV/AIDS (Topic)' cover a narrow range of AIDS related terms and will thus not be included in further analysis.

AIDS prevalence vs. Google Trends

In order to examine the possibility of forecasting AIDS prevalence in the US, the relationships between online search traffic data from Google and official health data on AIDS prevalence are at first examined, by calculating the respective correlations at both national and State level. Depending on the significance of the correlations, the possibility of forecasting AIDS prevalence in the US will be examined. For the analysis of AIDS related queries, both Google Trends categories, i.e. 'AIDS (Search Term)' and 'AIDS (Illness)', are analyzed. Data for the categories 'AIDS Deaths', 'AIDS Diagnoses', and 'AIDS Prevalence' are available for 12 years, i.e. from January 2004 to December 2015.

Statistically significant correlations are observed between 'AIDS Prevalence' with both 'AIDS (Search Term)' ($r=-0.9508$, $p<0.01$) and with 'AIDS (Illness)' ($r=-0.9615$, $p<0.01$) in the US. For 'AIDS (Search Term)', statistically significant correlations are observed with 'AIDS Diagnoses' ($r=0.8743$, $p<0.01$), and with 'AIDS Deaths' ($r=0.9343$, $p<0.01$). Significant correlations are also identified for 'AIDS Diagnoses' with 'AIDS (Illness)' ($r=0.8945$, $p<0.01$), and for 'AIDS Deaths' with 'AIDS (Illness)' ($r=0.9423$,

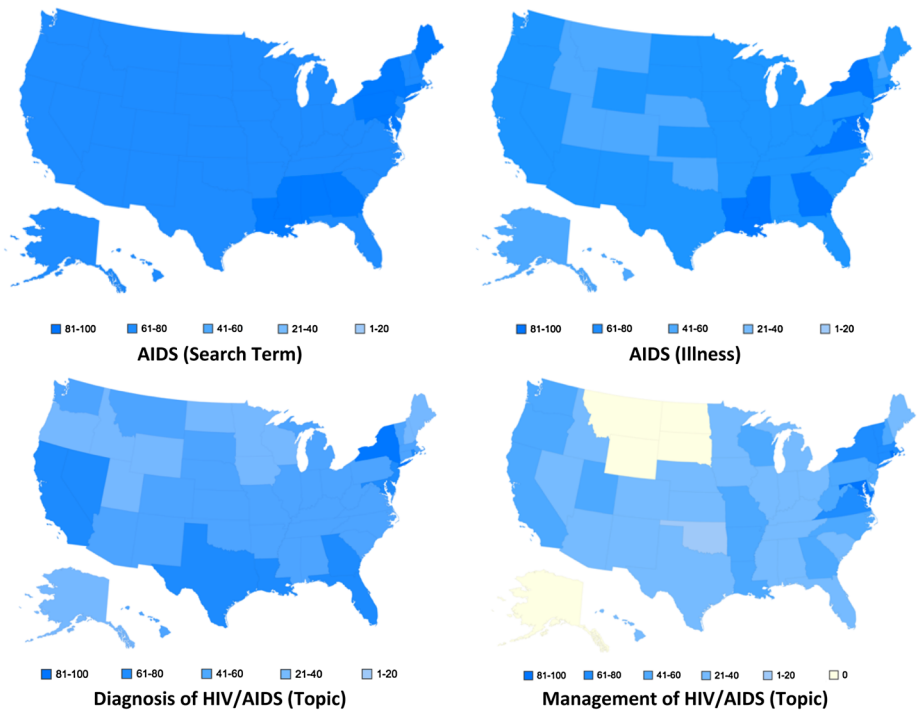


Fig. 3 Online Interest by State for 'AIDS (Search Term)', 'AIDS (Illness)', 'Diagnosis of HIV/AIDS (Topic)', and 'Management of HIV/AIDS (Topic)' from January 2004 to December 2015

$p < 0.01$). Therefore, we proceed to the next step of identifying correlations between online and health data in each US State.

Table 1 consists of the Pearson correlation coefficients (r) between 'AIDS Prevalence' and (a) 'AIDS (Search Term)' and (b) 'AIDS (Illness)' from January 2004 to December 2015, while Table 2 consists of the Pearson correlation coefficients (r) between 'AIDS Diagnoses' and a) 'AIDS (Search Term)' and b) 'AIDS (Illness)' from January 2004 to December 2015. Table 3 consists of the Pearson correlation coefficients (r) between 'AIDS Deaths', and a) 'AIDS (Search Term)' and b) 'AIDS (Illness)' from January 2004 to December 2015.

For 'AIDS Prevalence', all correlations are statistically significant. Therefore it is evident that the online behavior towards AIDS follows that of 'AIDS Prevalence'. Thus the States that exhibit statistically significant correlations are further selected for the forecasting of AIDS in the US.

For 'AIDS Diagnoses', the States with significance of correlation of $p < 0.01$ in both examined terms are Arkansas, California, Connecticut, Delaware, DC, Florida, Illinois, Indiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Missouri, Nevada, New Hampshire, New Jersey, New York, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, Washington, and West Virginia. For 'AIDS Deaths', the respective States are Arizona, California, Connecticut, Delaware, DC, Florida, Georgia, Illinois, Louisiana, Maryland, Massachusetts, Michigan, Mississippi, Missouri, New Jersey, New York, Pennsylvania, Tennessee, Texas, Utah, and Washington.

Table 1 Pearson correlation coefficients between 'AIDS Prevalence' and 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015

	AIDS (search term)	AIDS (illness)		AIDS (search term)	AIDS (illness)		AIDS (search term)	AIDS (illness)
AL	−0.8731	−0.9099	KY	−0.9521	−0.9302	ND	−0.7700	−0.8425
AK	−0.8568	−0.9003	LA	−0.8049	−0.8713	OH	−0.9231	−0.9311
AZ	−0.8319	−0.8386	ME	−0.8993	−0.9376	OK	−0.8958	−0.9137
AR	−0.9096	−0.9223	MD	−0.9554	−0.9519	OR	−0.9316	−0.9040
CA	−0.9716	−0.9710	MA	−0.9577	−0.9550	PA	−0.9713	−0.9909
CO	−0.9289	−0.9570	MI	−0.9894	−0.9936	RI	−0.9830	−0.9572
CT	−0.8418	−0.8244	MN	−0.9335	−0.9460	SC	−0.8690	−0.9142
DE	−0.9022	−0.8641	MS	−0.8308	−0.8752	SD	−0.8308	−0.8227
DC	−0.9174	−0.9164	MO	−0.9627	−0.9651	TN	−0.9034	−0.9340
FL	−0.9463	−0.9444	MT	−0.8317	−0.8975	TX	−0.9135	−0.9174
GA	−0.8951	−0.8851	NE	−0.9429	−0.8986	UT	−0.8004	−0.8384
HI	−0.8978	−0.8976	NV	−0.8408	−0.9104	VT	−0.8266	−0.8376
ID	−0.8227	−0.8233	NH	−0.9074	−0.9626	VA	−0.8710	−0.9375
IL	−0.9689	−0.9714	NJ	−0.9794	−0.9804	WA	−0.9575	−0.9530
IN	−0.9290	−0.9265	NM	−0.8858	−0.8354	WV	−0.7816	−0.8241
IA	−0.9550	−0.9519	NY	−0.9890	−0.9926	WI	−0.9298	−0.9313
KS	−0.9396	−0.9191	NC	−0.9308	−0.9402	WY	−0.9393	−0.8585

All correlations reported in this table are statistically significant with $p < 0.01$

Table 2 Pearson correlation coefficients between 'AIDS Diagnoses' and (a) AIDS (Search Term)' and (b) 'AIDS (Illness)' from January 2004 to December 2015

	AIDS (search term)	AIDS (illness)		AIDS (search term)	AIDS (illness)		AIDS (search term)	AIDS (illness)
AL	0.3785	0.3723	KY	0.4961	0.4486	ND	−0.3019	−0.4416
AK	0.6703**	0.6781**	LA	0.5913**	0.6127**	OH	0.6162**	0.6225**
AZ	0.5407*	0.5409	ME	0.7369***	0.7805***	OK	0.8091***	0.8007***
AR	0.7417***	0.7218***	MD	0.9548***	0.9485***	OR	0.8570***	0.7947***
CA	0.7892***	0.8752***	MA	0.9188***	0.9088***	PA	0.7548***	0.7885***
CO	0.7025**	0.7475***	MI	0.8174***	0.8500***	RI	0.9306***	0.9414***
CT	0.9073***	0.9342***	MN	0.8772***	0.8971***	SC	0.8078***	0.8680***
DE	0.8683***	0.8952***	MS	0.4497	0.3353	SD	0.197	0.0973
DC	0.8876***	0.8767***	MO	0.7687***	0.7644***	TN	0.6986**	0.7114***
FL	0.9141***	0.9203***	MT	0.4793	0.5216*	TX	0.6832**	0.6678**
GA	0.6711**	0.6613**	NE	0.6527**	0.6290**	UT	0.0594	0.1989
HI	0.6668**	0.6412**	NV	0.7547***	0.7992***	VT	0.3291	0.2394
ID	0.037	0.0295	NH	0.8076***	0.7846***	VA	0.4242	0.5301*
IL	0.8934***	0.8830***	NJ	0.8755***	0.8797***	WA	0.8275***	0.8204***
IN	0.8090***	0.7757***	NM	0.5913**	0.5266*	WV	0.7553***	0.8062***
IA	0.2429	0.184	NY	0.9462***	0.9479***	WI	0.6826**	0.7132***
KS	0.6121**	0.5699*	NC	0.3724	0.402	WY	−0.1721	−0.2062

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Forecasting AIDS prevalence in USA

As 'AIDS Prevalence' data are highly correlated with both 'AIDS (Search Term)' and with 'AIDS (Illness)' in all 50 States (plus DC), the next step is to examine the relationships

Table 3 Pearson correlation coefficients between 'AIDS Deaths' and (a) AIDS (Search Term)' and (b) 'AIDS (Illness)' from January 2004 to December 2015

	AIDS (search term)	AIDS (illness)		AIDS (search term)	AIDS (illness)		AIDS (search term)	AIDS (illness)
AL	0.6079**	0.7163***	KY	0.4357	0.3921	ND	0.1693	0.2203
AK	0.2352	0.0614	LA	0.7166***	0.7977***	OH	0.6211**	0.6414**
AZ	0.9078***	0.8694***	ME	0.3963	0.3518	OK	−0.0216	−0.0309
AR	0.3168	0.3696	MD	0.9157***	0.9153***	OR	0.429	0.4976*
CA	0.9748***	0.9272***	MA	0.9528***	0.9503***	PA	0.8666***	0.8707***
CO	0.6677**	0.6843**	MI	0.7982***	0.8343***	RI	0.6301**	0.6834**
CT	0.9486***	0.9502***	MN	0.1207	0.1212	SC	0.6050**	0.7163***
DE	0.7248***	0.8979***	MS	0.7318***	0.7766***	SD	0.2726	0.2971
DC	0.8975***	0.8842***	MO	0.8488***	0.8431***	TN	0.8685***	0.9143***
FL	0.8923***	0.9059***	MT	0.0672	0.2129	TX	0.8428***	0.8314***
GA	0.7522***	0.7388***	NE	−0.0976	−0.1354	UT	0.8202***	0.8492***
HI	0.48	0.5338*	NV	0.4519	0.4493	VT	0.2404	0.4347
ID	−0.0062	−0.1231	NH	0.1343	0.2082	VA	0.6974**	0.7303***
IL	0.8944***	0.8869***	NJ	0.9643***	0.9694***	WA	0.7672***	0.7778***
IN	0.3926	0.3771	NM	−0.2418	−0.1415	WV	0.3769	0.4425
IA	−0.3386	−0.3144	NY	0.9536***	0.9545***	WI	0.3208	0.3706
KS	0.0073	0.0024	NC	0.4551	0.4577	WY	0.3951	0.3269

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

between Google data and AIDS data and estimate the forecasting models. The relationship is logarithmic and of the form $y = \alpha \ln(x) + \beta$, where y (y-axis-dependent variable) denotes the 'AIDS Prevalence', x (x-axis-independent variable) denotes the respective Google Trends' data, namely 'AIDS (Search Term)' and 'AIDS (Illness)', and α and β are constants. To elaborate on the robustness of the estimated models, the R^2 is selected, as it is the statistical measure by which the variable variation is explained. R^2 takes values between 0 and 1 (i.e. 0% to 100%), and the higher the percentage, the better the fit.

Table 4 consists of the coefficients for the estimated logarithmic models for 'AIDS Prevalence' for both the examined Google Trends' terms, i.e. 'AIDS (Search Term)' and 'AIDS (Illness)', while Figs. 4, 5, 6 and 7 depict the respective relationships in the US and in each individual State.

In the US, the estimated models for 'AIDS Prevalence' based on the two examined terms have an R^2 of 0.9695 and 0.9844, which shows that the relationship between AIDS prevalence and Google Trends data is well described using the estimated equations and that AIDS prevalence can be predicted based on online search traffic data from Google. Furthermore, most States' models exhibit high R^2 in at least one Google Trends' category, which is indicative of the significance of the estimated forecasting models of AIDS prevalence in the US States.

Though in several States the R^2 is higher for the respective linear or polynomial forecasting model, the relationship is overall logarithmic as clearly shown in the case of the US. Therefore, all estimated models for all categories and all individual States are calculated based on a logarithmic relationship independent of the value of R^2, as this will be more evident when more years' data are available.

Table 4 Regression coefficients and R^2 for the estimated forecasting models for 'AIDS Prevalence'

	AIDS (search term)			AIDS (illness)		
	α	β	R^2	α	β	R^2
USA	− 100,000	881,002	0.9695	− 100,000	830,816	0.9844
Alabama	− 2483	12,707	0.8661	− 2370	11,734	0.9207
Alaska	− 90.42	625.04	0.7216	− 78.32	586.9	0.8248
Arizona	− 3851	17,270	0.8075	− 3334	15,114	0.8137
Arkansas	− 667	3962.2	0.8775	− 573.6	3870.6	0.9132
California	− 11,677	99,283	0.9684	− 10,780	96,113	0.9239
Colorado	− 1619	9209.8	0.7248	− 1508	8592.7	0.7922
Connecticut	− 370	7634	0.6253	− 278	7301	0.5456
Delaware	− 406	3222	0.8132	− 320	2901	0.813
DC	− 1313	12,424	0.7864	− 1212	12,059	0.7763
Florida	− 17,848	105,055	0.955	− 14,942	97,677	0.9571
Georgia	− 13,963	63,863	0.893	− 12,276	5852	0.8835
Hawaii	− 456.2	2903.5	0.8666	− 442	2802.8	0.8842
Idaho	− 292.3	1311.6	0.7561	− 209.2	1057	0.7494
Illinois	− 4563	29,922	0.9783	− 4151	29,124	0.9865
Indiana	− 1899	10,401	0.9330	− 1545	9239.8	0.9463
Iowa	− 622.2	3106.6	0.9292	− 548.7	2671.5	0.9462
Kansas	− 516.1	2834	0.9193	− 442.2	2589.8	0.9246
Kentucky	− 1604	8141.1	0.9557	− 1173	6561.5	0.9468
Louisiana	− 3661	20,818	0.7518	− 2882	17,158	0.8707
Maine	− 343	1787.7	0.8870	− 257.4	1419.5	0.9540
Maryland	− 4325	28,920	0.9663	− 3755	26,973	0.9788
Massachusetts	− 2369	17,052	0.9828	− 2102	16,120	0.9848
Michigan	− 2349	14,054	0.9853	− 2010	13,066	0.9752
Minnesota	− 1453	7419.2	0.9426	− 1286	6450.5	0.9704
Mississippi	− 2446	12,186	0.7234	− 2259	12,064	0.8157
Missouri	− 1824	10,892	0.9658	− 1565	10,104	0.9723
Montana	− 140.4	687.7	0.7816	− 125.7	596.03	0.8817
Nebraska	− 438.2	2148.4	0.9400	− 367.3	1954.6	0.9042
Nevada	− 2109	10,090	0.7889	− 2151	10,041	0.8836
New Hampshire	− 156.7	995.89	0.9123	− 157.5	1003.1	0.9437
New Jersey	− 2052	24,339	0.9535	− 1771	23,254	0.9361
New Mexico	− 770.7	3825.1	0.8517	− 632.1	3155.7	0.7478
New York	− 8477	97,596	0.9246	− 7652	94,878	0.9283
North Carolina	− 6723	31,000	0.9409	− 5705	26,921	0.9588
North Dakota	− 74.1	312.72	0.7160	− 75.35	302.53	0.7879
Ohio	− 4158	20,794	0.9309	− 3499	18,605	0.957
Oklahoma	− 1158	6162.9	0.8817	− 923.4	5001.5	0.9097
Oregon	− 1354	6969.1	0.9189	− 1247	6570.4	0.9129
Pennsylvania	− 4376	30,222	0.9891	− 3825	28,372	0.9914
Rhode Island	− 224	1942	0.9389	− 177	1824	0.8333
South Carolina	− 3917	20,557	0.7879	− 3277	18,920	0.8652
South Dakota	− 94.71	460	0.7941	− 72.61	361.2	0.7471
Tennessee	− 3760	19,372	0.8981	− 3136	17,266	0.949
Texas	− 17,403	88,900	0.9182	− 16,260	85,188	0.9207
Utah	− 314.3	2121.8	0.7112	− 285.2	1983.4	0.8525
Vermont	− 107	583.73	0.7082	− 91.13	545.91	0.8386

Table 4 (continued)

	AIDS (search term)			AIDS (illness)		
	α	β	R^2	α	β	R^2
Virginia	− 2376	15,862	0.8017	− 2530	16,430	0.9206
Washington	− 1696	11,095	0.9594	− 1464	10,179	0.9652
West Virginia	− 327.7	1857.3	0.6888	− 339.3	1727.3	0.7492
Wisconsin	− 1207	5836.5	0.9428	− 956.5	5201.9	0.9594
Wyoming	− 58.9	316.91	0.9289	− 49.8	253.69	0.8337

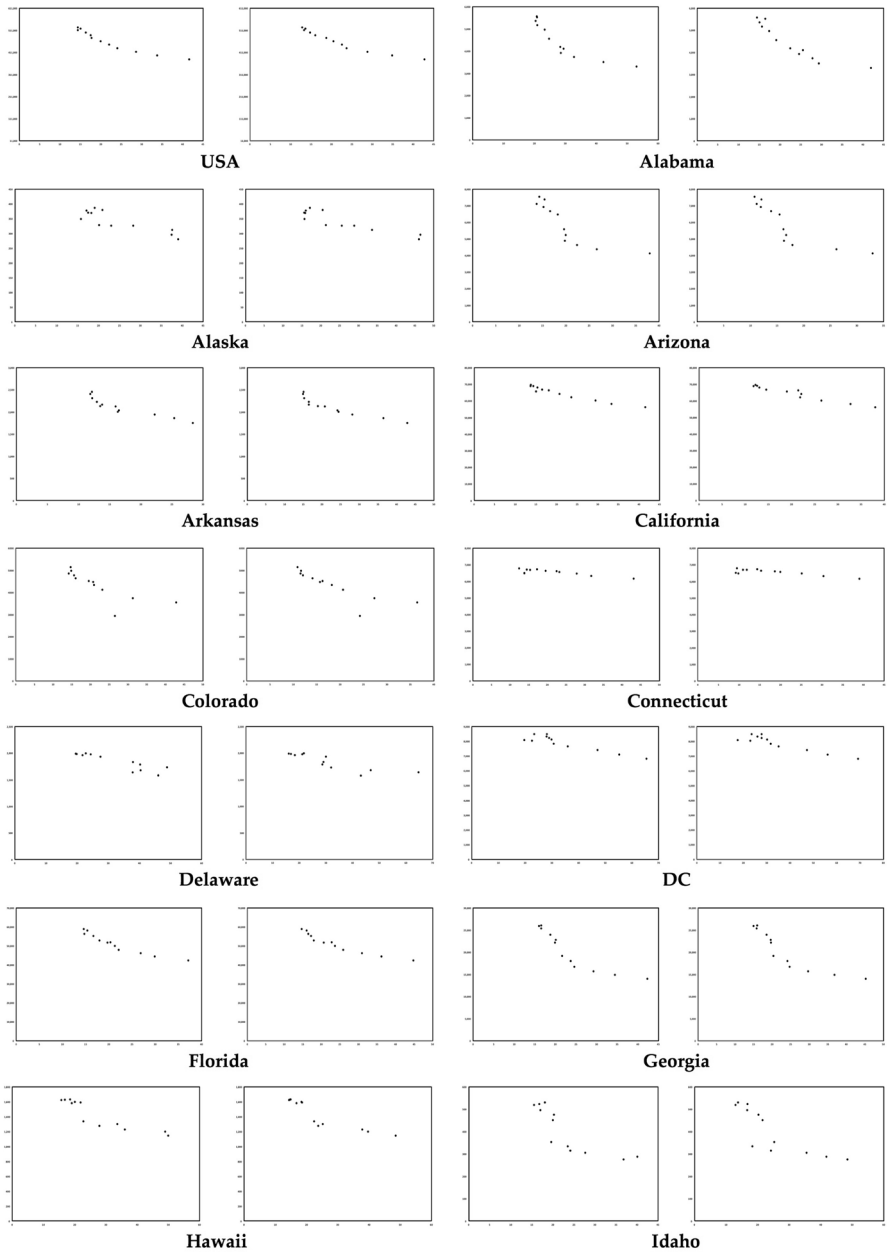


Fig. 4 'AIDS Prevalence' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015 (USA; Alabama–Idaho)

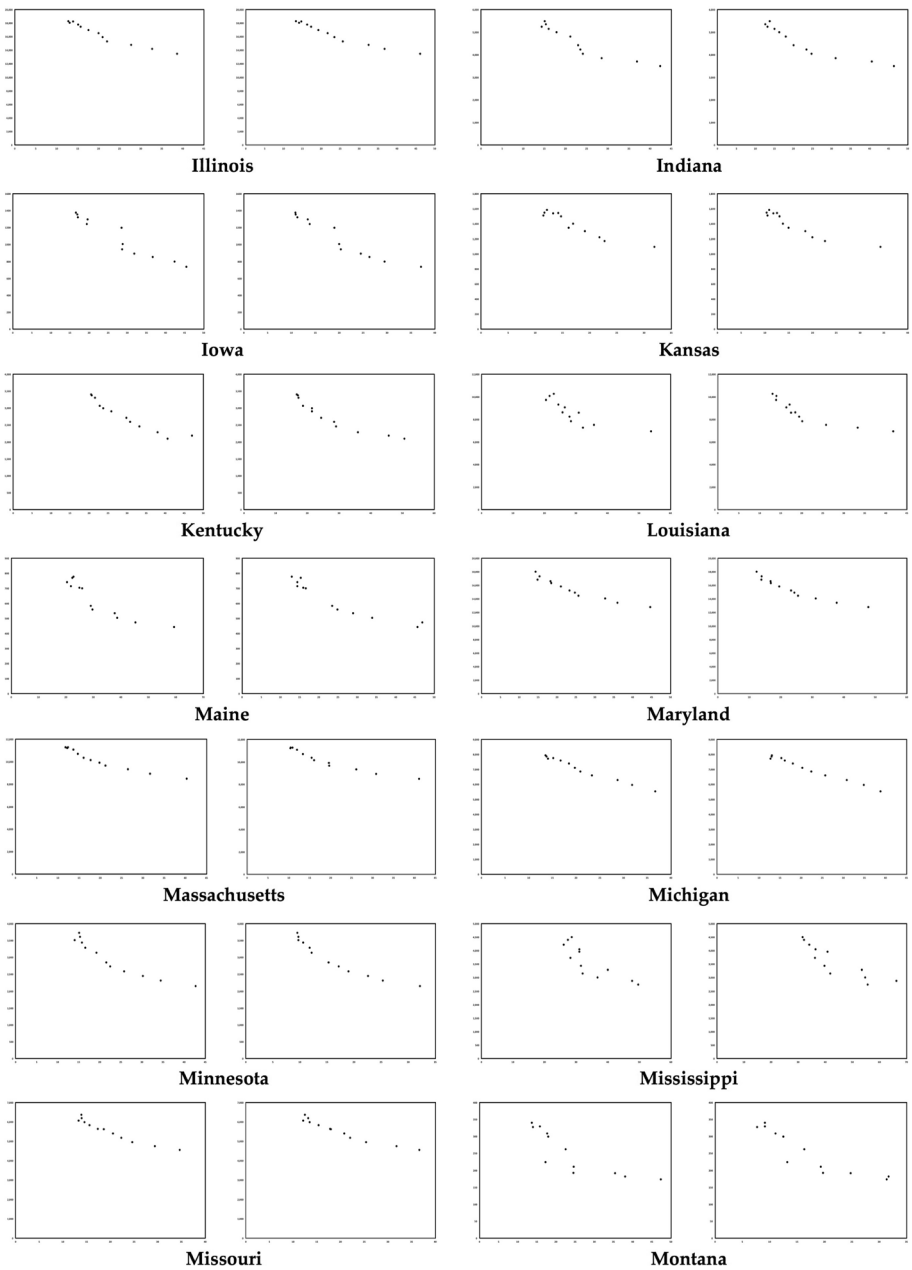


Fig. 5 'AIDS Prevalence' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015 (Illinois–Montana)

The categories 'AIDS Diagnoses' and 'AIDS Deaths', though significant correlations with Google data are identified, are not included in further analysis, as the results are not significant for all States, though the respective analyses on said categories can be found in Appendix 1 and Appendix 2.

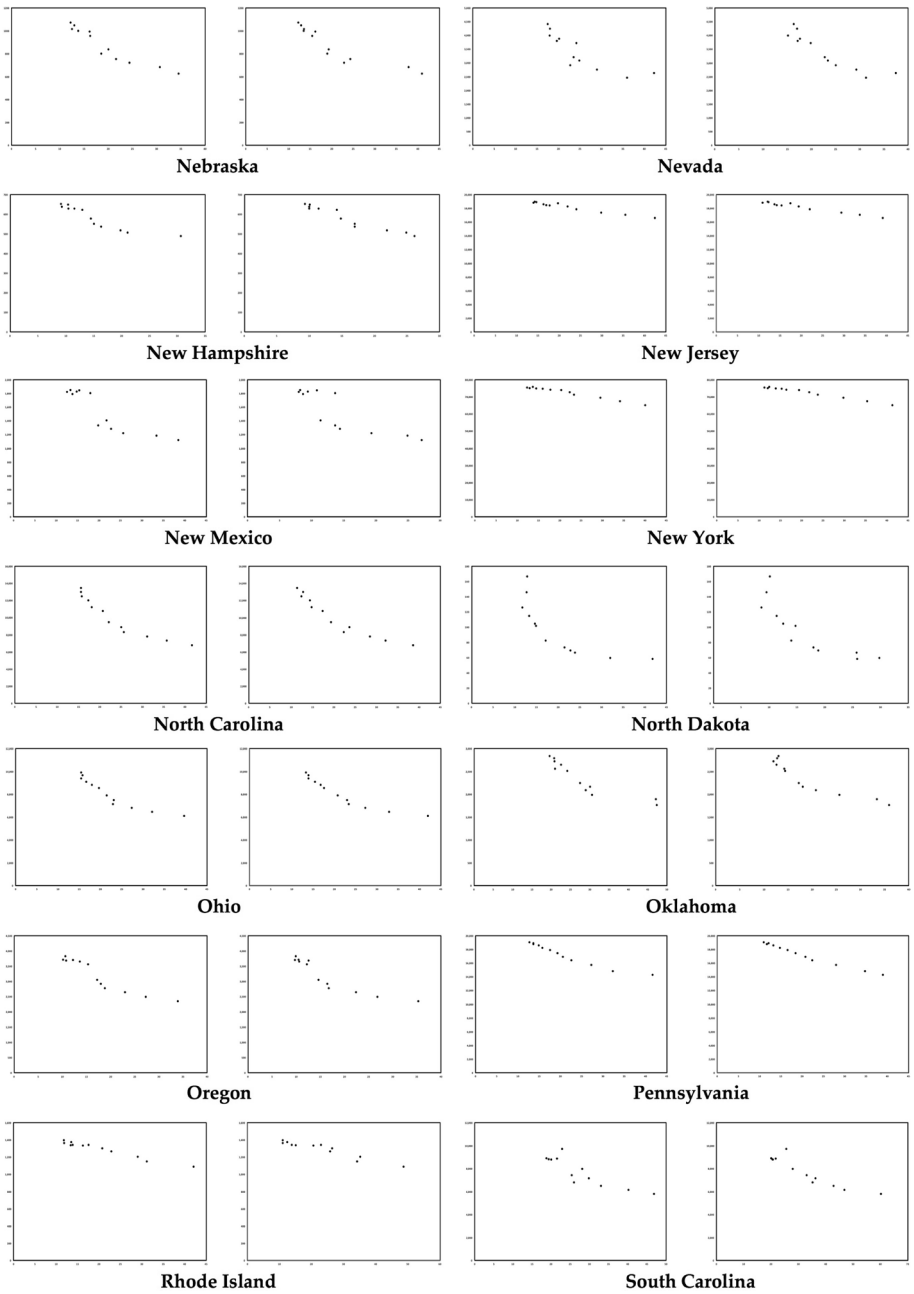


Fig. 6 'AIDS Prevalence' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015 (Nebraska–South Carolina)

Discussion

The AIDS epidemic is a serious health issue and needs immediate and constant attention. In the Internet age, new methods for the monitoring and assessment of AIDS are required, so as to decrease the numbers of AIDS prevalence and deaths around the globe, and especially in developing countries. In this study, we provide a novel approach

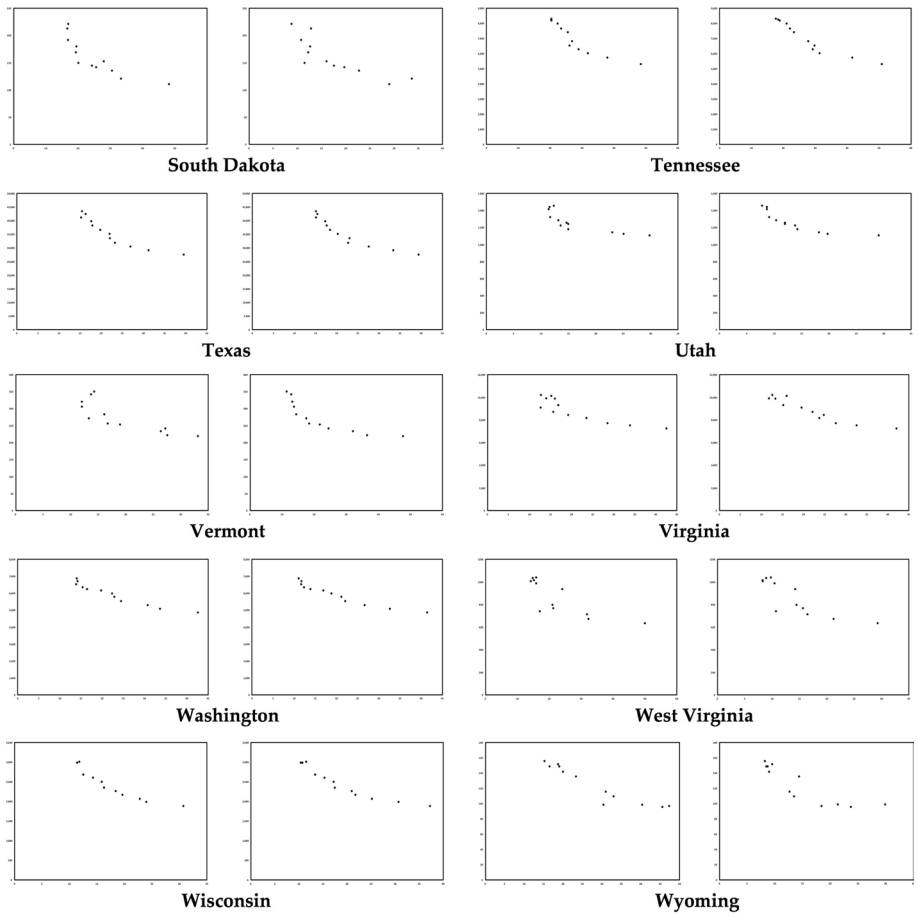


Fig. 7 'AIDS Prevalence' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015 (South Dakota–Wyoming)

of monitoring online search traffic data retrieved from Google Trends in order to develop forecasting models for AIDS prevalence in the US.

Both examined Google terms, i.e. 'AIDS (Search Term)' and 'AIDS (Illness)', exhibited significant correlations with official data on 'AIDS Prevalence', 'AIDS Diagnoses', and 'AIDS Deaths', especially in the States where the AIDS rates are higher. Despite previous concerns on the reliability of Google Trends data [38], our results support research over the last decade showing that empirical relationships widely exist between Google Trends' data and public health data records [5, 6, 9, 11, 20–22, 26, 39–42]. Therefore, the forecasting of AIDS prevalence is possible, as the estimated models for several States are robust despite the limitation of data being available for only 12 years. For 'HIV (Search Term)' and 'HIV (Illness)', though search volumes are high throughout the examined period, the correlations with official HIV data were not as statistically significant as in the case of AIDS, and were identified in fewer US States, which is an interesting topic to be examined in future research on the subject.

Table 5 consists of the coefficients and the R^2 for the estimated forecasting logarithmic forecasting models of the form $y = \alpha \ln(x) + \beta$ for States that exhibit high significance in all three categories, i.e. 'AIDS Prevalence', 'AIDS Diagnoses', and 'AIDS Deaths'.

Table 5 Estimated Logarithmic forecasting models for USA and selected states

	AIDS	AIDS (search term)			AIDS (illness)		
		α	β	R^2	α	β	R^2
USA	Prevalence	− 100,000	881,002	0.9695	− 100,000	830,816	0.9844
	Diagnoses	16,068	− 20,481	0.8548	14,525	− 15,399	0.8982
	Deaths	5859	− 2738	0.9452	5186	− 553.58	0.9524
California	Prevalence	− 11,677	99,283	0.9684	− 10,780	96,113	0.9239
	Diagnoses	1670.7	− 1852	0.7151	1733.9	− 1962.9	0.8622
	Deaths	559.95	− 168.21	0.942	481.76	87.41	0.7806
Florida	Prevalence	− 17,848	105,055	0.955	− 14,942	97,677	0.9571
	Diagnoses	2822.4	− 5025	0.8862	2396.1	− 3962	0.9133
	Deaths	970.62	− 961.25	0.8457	828.81	− 610.55	0.8816
Illinois	Prevalence	− 4563	29,922	0.9783	− 4151	29,124	0.9865
	Diagnoses	587.39	− 713.99	0.8688	530.2	− 598.58	0.8625
	Deaths	207.83	− 107.81	0.8635	187.74	− 67.41	0.8586
Maryland	Prevalence	− 4325	28,920	0.9663	− 3755	26,973	0.9788
	Diagnoses	764.14	− 1357.2	0.9226	658.98	− 999.93	0.9223
	Deaths	323.92	− 410.32	0.8647	281.24	− 264.71	0.8762
Massachusetts	Prevalence	− 2369	17,052	0.9828	− 2102	16,120	0.9848
	Diagnoses	363.09	− 566.12	0.9204	319.15	− 414.72	0.9053
	Deaths	92.75	− 19.36	0.8951	81.7	18.84	0.8841
New Jersey	Prevalence	− 2052	24,339	0.9535	− 1771	23,254	0.9361
	Diagnoses	633.68	− 928.47	0.8535	555.61	− 618.92	0.8651
	Deaths	377.47	− 468.44	0.9657	328.49	− 276.84	0.9642
New York	Prevalence	− 8477	97,596	0.9246	− 7652	94,878	0.9283
	Diagnoses	3085.9	− 6020.5	0.9607	2794.8	− 5058.3	0.9709
	Deaths	978.44	− 877.45	0.9683	885.23	− 569.63	0.9765

This study has some limitations. The estimated forecasting models are based on only 12 years' data, thus the robustness of the models will increase when more years or smaller interval data are made officially available. In addition, we do not argue that each hit on the AIDS related keywords corresponds to an AIDS case and vise versa, as hits can also be attributed to general or academic interest, or increased interest due to an event, incident, or public figure that announces something related to the disease. Overall, the online interest towards AIDS increases according to the rates of AIDS prevalence (Appendix 3), thus it is expected for the forecasting models to be robust in the States for which the rates—and the online interest—are increased. Therefore, when more data are available, the significance will most probably increase.

Overall, this study highlights the importance of the analysis of online queries in order to better and more timely assess various issues in the US Health Care System. The estimated forecasting models on AIDS prevalence have very good performance, indicating that Google data can be of value in dealing with this sensitive subject, as we can this way have access to data that would not easily or at all been accessed with conventional methods.

Conclusions

This study aimed at introducing a novel approach in forecasting AIDS prevalence in the US using data from Google Trends on related terms. The results, exhibiting significant correlations between Google Trends' data and official health data on AIDS (2004–2015) and high significance of the estimated forecasting models in several US States, support previous work on the subject suggesting that Google Trends' data have been shown to be empirically related to health data and that they can assist with the analysis, monitoring, and forecasting of several health topics. This study, however, also addresses a more important issue; that of anonymity. A Google Trends important advantage is that it uses the revealed and not the stated data [37] in general, but in the case of AIDS the latter is even more important. As HIV and AIDS testing, diagnosis, and treatment is a sensitive subject, people may less easily go to the hospital or consult a doctor, health official, especially before testing and diagnosis.

Therefore, the monitoring of the interest towards States with increased rates of AIDS prevalence is essential, so that health officials can a) make relative information available on the Internet at time point e.g. with advertisements, b) take preventive measures, e.g. organizing event etc., and c) prepare the Health Care System accordingly, e.g. organize free testing outside of the hospitals. AIDS and HIV are terms that are not translated, not easily misspelled, and do not include accents or special characters. Thus, future research can include the application of this method in other countries and regions, as well as taking into consideration data retrieved by other online sources.

Authors' contributions
AM collected the data, performed the analysis, and wrote the paper. GO had the overall supervision. Both authors read and approved the final manuscript.

Authors' information
Amaryllis Mavragani is a Ph.D. Candidate at the Department of Computing Science and Mathematics, University of Stirling. She holds a B.Sc in Mathematics from the University of Crete and an M.Sc from Democritus University of Thrace. Her research interests include Data Analysis, Mathematical Modeling, Online Behavior, Big Data, Public Health, Environmental Economics and Legislation, and Statistical Analysis.

Gabriela Ochoa is a Senior Lecturer at the Department of Computing Science and Mathematics, University of Stirling. She holds a Ph.D. in Computer Science and Artificial Intelligence from the University of Sussex, UK. Her research interests lie in the foundations and application of evolutionary algorithms and heuristic search methods, data science and visualization. She is an associated editor of both the IEEE Transactions on Evolutionary Computation and the Evolutionary Computation Journal, MIT Press.

Acknowledgements
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
All data used in this study are publicly available and accessible in the cited sources.

Consent for publication
The authors consent to the publication of this work.

Ethics approval and consent to participate
Not applicable.

Funding
Not applicable.

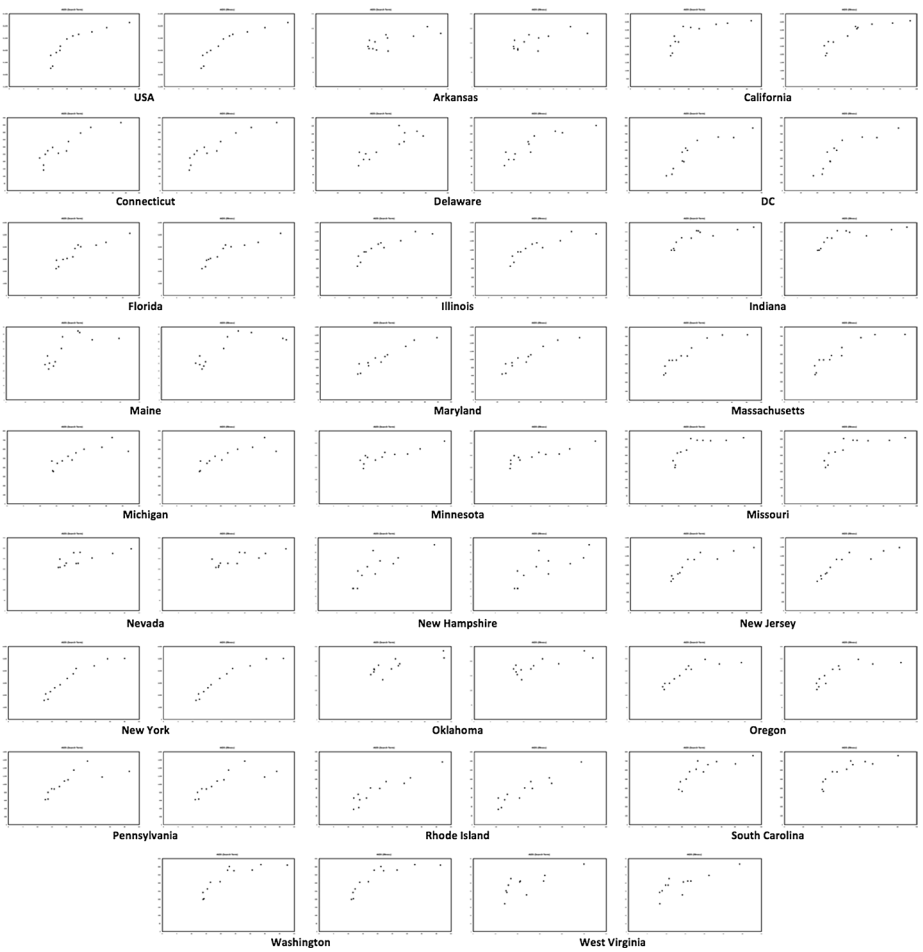


Fig. 8 'AIDS Diagnoses' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015

Appendix 1

AIDS diagnoses vs. Google Trends

Figure 8 depicts the scatterplots between 'AIDS Diagnoses' and both the examined Google terms, i.e. 'AIDS (Search Term)' and 'AIDS (Illness)', in the US and in the 25 States for which significant correlations with $p < 0.01$ were observed between AIDS and Google data. The States are Arkansas, California, Connecticut, Delaware, DC, Florida, Illinois, Indiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Missouri, Nevada, New Hampshire, New Jersey, New York, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, Washington, and West Virginia.

Table 6 consists of the coefficients for the estimated logarithmic models for 'AIDS Diagnoses' for both the examined terms, namely 'AIDS (Search Term)' and 'AIDS (Illness)'. As in 'AIDS Prevalence', the relationship between Google Trends and health data is logarithmic and of the form $y = \alpha \ln(x) + \beta$.

For 'AIDS Diagnoses', the estimated forecasting models for 'AIDS (Search Term)' and 'AIDS (Illness)' in the US have an R^2 of 0.8548 and 0.8982, respectively. It is therefore evident that the forecasting model for 'AIDS Diagnoses' in the US performs well, though

Table 6 Coefficients α and β, and R^2 for the estimated forecasting models for 'AIDS Diagnoses'

	AIDS (search term)			AIDS (illness)		
	α	β	R^2	α	β	R^2
USA	16,068	−20,481	0.8548	14,525	−15,399	0.8982
Arkansas	66.22	−29.35	0.5501	54.641	−13.149	0.5269
California	1670.70	−1852	0.7151	1733.90	−1962.90	0.8622
Connecticut	235.88	−412.74	0.8485	195.81	−251.63	0.9036
Delaware	81.88	−171.42	0.7702	68.34	−119.06	0.8612
DC	490.36	−1245.00	0.8745	451.06	−1103.10	0.8577
Florida	2822.40	−5025.00	0.8862	2396.10	−3962.00	0.9133
Illinois	587.39	−713.99	0.8688	530.20	−598.58	0.8625
Indiana	112.76	−33.04	0.7622	88.92	44.56	0.7261
Maine	23.68	−46.86	0.6465	18.11	−22.50	0.7223
Maryland	764.14	−1357.20	0.9226	658.98	−999.93	0.9223
Massachusetts	363.09	−566.12	0.9204	319.15	−414.72	0.9053
Michigan	272.70	−304.44	0.7383	241.81	−215.23	0.7847
Minnesota	69.68	−18.47	0.7487	62.60	25.53	0.7937
Missouri	177.82	−186.27	0.6939	152.91	−110.45	0.7018
Nevada	84.23	−23.00	0.5873	84.29	−16.08	0.6334
New Hampshire	23.28	−33.43	0.7263	21.90	−30.47	0.6578
New Jersey	633.68	−928.47	0.8535	555.61	−618.92	0.8651
New York	3085.90	−6020.50	0.9607	2794.80	−5058.30	0.9709
Oklahoma	72.37	−57.96	0.6557	56.03	19.55	0.6372
Oregon	101.12	−96.98	0.8577	88.18	−53.71	0.7636
Pennsylvania	655.13	−927.11	0.6999	587.89	−694.56	0.7392
Rhode Island	65.89	−119.03	0.8694	55.05	−92.83	0.8557
South Carolina	371.79	−642.02	0.7354	318.21	−511.42	0.8456
Washington	192.94	−267.10	0.8056	167.26	−165.03	0.8176
West Virginia	24.03	−15.52	0.5608	25.54	−7.69	0.6431

not as well as in the 'AIDS Prevalence' category, which could be attributed to the more narrow –compared to AIDS prevalence—field that said category covers, which is also supported by the correlations in Table 2, which show that the 'AIDS Diagnoses' are not as significantly and in less States correlated with Google Trends' data.

Appendix 2

AIDS Deaths vs. Google Trends

Figure 9 depicts the relationship between 'AIDS Deaths' and both the examined Google terms, i.e. 'AIDS (Search Term)' and 'AIDS (Illness)', in the US and in the 21 States for which significant correlations with $p < 0.01$ between AIDS data and Google Trends' data were observed. These States are Arizona, California, Connecticut, Delaware, DC, Florida, Georgia, Illinois, Louisiana, Maryland, Massachusetts, Michigan, Mississippi, Missouri, New Jersey, New York, Pennsylvania, Tennessee, Texas, Utah, and Washington.

Table 7 consists of the coefficients for the estimated logarithmic models for 'AIDS Deaths' for both the examined Google Trends' terms, i.e. 'AIDS (Search Term)' and 'AIDS (Illness)' for the aforementioned States.

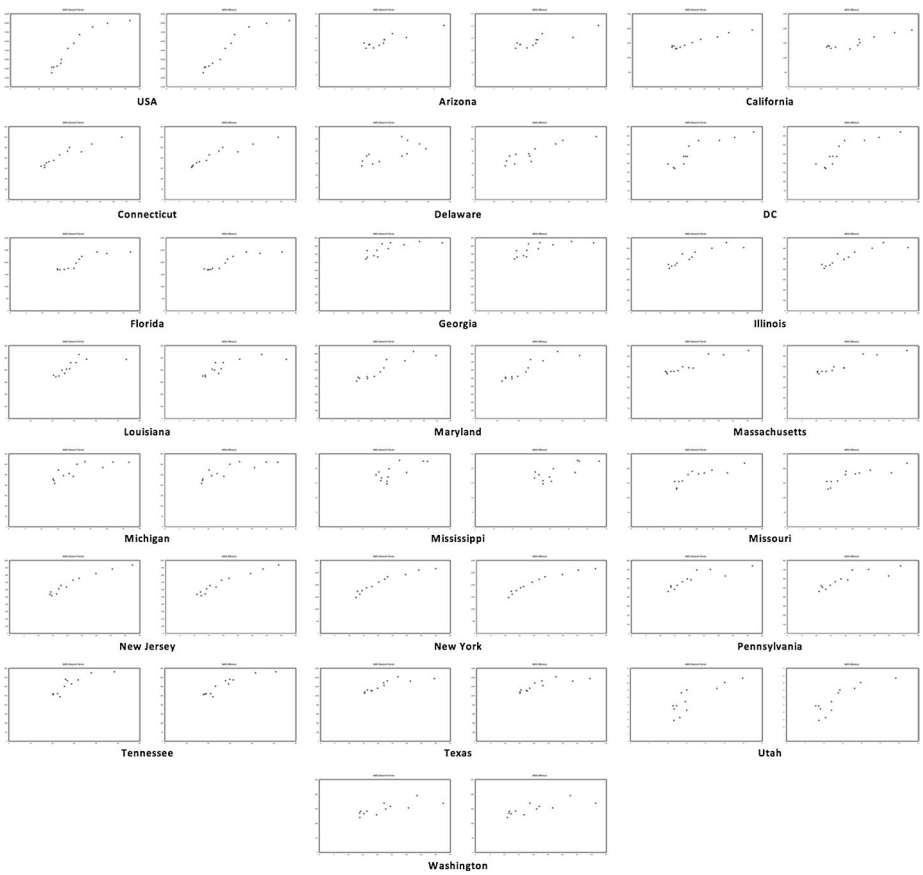


Fig. 9 'AIDS Deaths' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015

Thus, as in the case of 'AIDS Diagnoses', when the AIDS data category is narrow, the forecasting results are robust in less States. Despite this, the forecasting models for the 'AIDS Prevalence' category exhibit significant results. Therefore, as more years' data become available, the forecasting of AIDS Diagnoses and Deaths will be possible in more States.

Appendix 3

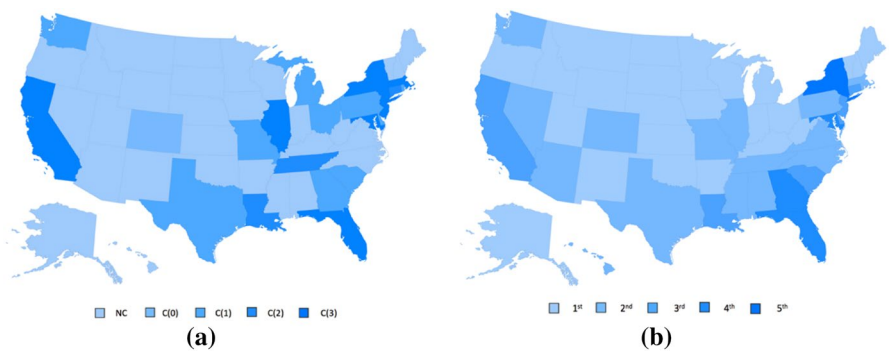
Forecasting model significance vs. AIDS rates in the US

Figure 10a maps the following five groups of significance of modeling by State: the first level—denoted by NC-consists of the States for which the correlations between health and Google data were not significant in all pairs of categories and thus not included for further analysis. The second group—denoted by C(0)-consists of the States for which significant correlations were identified in all categories, but the forecasting models had an R^2 lower than 0.85 in all AIDS data categories. The third, forth, and fifth groups—denoted by C(1), C(2), and C(3), respectively- consist of the States for which significant correlations were identified in all categories, and the forecasting models' R^2 was above 0.85 in one (1), two (2), and three (3) AIDS data categories, respectively.

Table 7 Coefficients α and β, and R^2 for the estimated forecasting models for 'AIDS Deaths'

	AIDS (search term)			AIDS (illness)		
	α	β	R^2	α	β	R^2
USA	5859	−2738	0.9452	5186	−553.58	0.9524
Arizona	82.79	−5778	0.8031	68.27	−1.97	0.7339
California	559.95	−168.21	0.9420	481.76	87.41	0.7806
Connecticut	117.33	−141.80	0.9481	94.35	−53.21	0.9474
Delaware	33.22	−38.50	0.5379	31.87	−31.08	0.7946
DC	176.46	−357.37	0.8566	161.04	−301.91	0.8267
Florida	970.62	−961.25	0.8457	828.81	−610.55	0.8816
Georgia	209.60	99.15	0.6421	182.93	183.64	0.6259
Illinois	207.83	−107.81	0.8635	187.74	−67.41	0.8586
Louisiana	189.47	−213.67	0.6370	148.89	−23.65	0.7357
Maryland	323.92	−410.32	0.8647	281.24	−264.71	0.8762
Massachusetts	92.75	−19.36	0.8951	81.70	18.84	0.8841
Michigan	91.73	−10.15	0.6962	82.42	16.64	0.7596
Mississippi	96.25	−153.79	0.5226	88.01	−145.31	0.5765
Missouri	73.36	−45.40	0.7740	62.78	−13.22	0.7753
New Jersey	377.47	−468.44	0.9657	328.49	−276.84	0.9642
New York	978.44	−877.45	0.9683	885.23	−569.63	0.9765
Pennsylvania	224.11	−85.79	0.8285	195.35	10.60	0.8257
Tennessee	164.41	−237.24	0.8190	139.23	−152.07	0.8921
Texas	378.75	48.21	0.8020	347.21	149.30	0.7741
Utah	23.82	−37.38	0.6919	20.70	−24.59	0.7611
Washington	44.27	13.46	0.6295	38.75	35.80	0.6516

In order to elaborate on why some States exhibit low correlations and not significant forecasting models and why some others show very high correlations in addition to very significant forecasting models, we calculate the average of the AIDS prevalence yearly Rates for all US States excluding DC from 2004 to 2015 and divide them into 5 classes of equal intervals. Figure 10b maps said 5 classes of AIDS prevalence Rates' in each US State. As is evident, a correspondence exists between the 1st class of AIDS prevalence rates, i.e. the group with the States that do not exhibit significant correlations between Google data in AIDS related terms with official data on AIDS prevalence, Diagnoses, and Deaths. In particular, the 1st class, i.e. with average yearly rates on AIDS prevalence from 2004 to 2015 of 16.81 to 99.10, consists of 29 out of the 51 States, namely Oregon,



(a) (b)

Fig. 10 US states categorized by (**a**) correlations-estimated models' significance and (**b**) AIDS rates

New Mexico, Arkansas, Indiana, Michigan, Ohio, Kentucky, Minnesota, Kansas, Utah, Alaska, Nebraska, West Virginia, Maine, New Hampshire, Wisconsin, Vermont, Iowa, Idaho, Montana, Wyoming, South Dakota, and North Dakota. Of those, only two exhibit significant correlations between public health and Google data, namely Michigan and Ohio. It is thus evident that the online interest towards AIDS increases according to the rates of AIDS prevalence, thus it is expected for the forecasting models to be robust in the States for which the rates are increased.

Publisher's Note

Received: 14 March 2018 Accepted: 8 May 2018
Published online: 19 May 2018

References

1. Hilbert M, Lopez P. The World's technological capacity to store, communicate, and compute information. Science. 2011;332:60–5.
2. Chen CLP, Zhang CY. Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inform Sci. 2014;275:314–47.
3. Al Nuaimi E, Al Neyadi H, Mohamed N, Al-Jaroodi J. Applications of big data to smart cities. J Int Serv App. 2015;6:25.
4. Matthew Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. J Big Data. 2014;1:2.
5. Preis T, Moat HS, Stanley HE, Bishop SR. Quantifying the advantage of looking forward. Sci Rep. 2012;2:350.
6. Preis T, Moat HS, Stanley HE. Quantifying trading behavior in financial markets using Google Trends. Sci Rep. 2013;3:1684.
7. Google Trends. https://trends.google.com/trends/explore. Accessed 7 Feb 2018.
8. Nuti SV, Wayda B, Ranasinghei I, Wang S, Dreyer RP, Chen SI, Murugiah K. The use of Google Trends in health care research: a systematic review. PLoS ONE. 2014;9:e109583.
9. Alicino C, Bragazzi NL, Faccio V, Amicizia D, Panatto D, Gasparini R, Icardi G, Orsi A. Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes. Infect Dis Poverty. 2015;4(1):54.
10. Hossain L, Kam D, Kong F, Wigand RT, Bossomaier T. Social media in Ebola outbreak. Epidemiol Infect. 2016;144:2136–43.
11. Mavragani A, Ochoa G. The internet and the anti-vaccine movement: tracking the 2017 EU measles outbreak. Big Data Cogn Comput. 2018;2(1):2.
12. Sentana-Lledo D, Barbu CM, Ngo MN, Wu Y, Sethuraman K, Levy MZ. Seasons, searches, and intentions: what the internet can tell us about the bed bug (Hemiptera: Cimicidae) epidemic. J Med Entomol. 2016;53(1):116–21.
13. Zhou X, Ye J, Feng Y. Tuberculosis surveillance by analyzing Google Trends. IEEE Trans Biomed Eng. 2011;58:2247–54.
14. Kang M, Zhong H, He J, Rutherford S, Yang F. Using Google Trends for influenza surveillance in South China. PLoS ONE. 2013;8(1):e55205.
15. Davidson MW, Haim DA, Radin JM. Using networks to combine big data and traditional surveillance to improve influenza predictions. Sci Rep. 2015;5:8154.
16. Cho S, Sohn CH, Jo MW, Shin SY, Lee JH, Ryoo SM, Kim WY, Seo DW. Correlation between national influenza surveillance data and Google Trends in South Korea. PLoS ONE. 2013;8:e81422.
17. Domnich A, Panatto D, Signori A, Lai PL, Gasparini R, Amicizia D. Age-related differences in the accuracy of web query-based predictions of influenza-like illness. PLoS ONE. 2015;10:0127754.
18. Solano P, Ustulin M, Pizzorno E, Vichi M, Pompili M, Serafini G, Amore M. A Google-based approach for monitoring suicide risk. Psychiatry Res. 2016;246:581–6.
19. Arora VS, Stuckler D, McKee M. Tracking search engine queries for suicide in the United Kingdom, 2004–2013. Public Health. 2016;137:147–53.
20. Mavragani A, Sypsa K, Sampri A, Tsagarakis KP. Quantifying the UK online interest in substances of the EU watch list for water monitoring: diclofenac, estradiol, and the macrolide antibiotics. Water. 2016;8:542.
21. Gahr M, Uzelac Z, Zeiss R, Connemann BJ, Lang D, Schönfeldt-Lecuona C. Linking annual prescription volume of antidepressants to corresponding web search query data: a possible proxy for medical prescription behavior? J Clin Psychopharmacol. 2015;235:681–5.
22. Schuster NM, Rogers MA, McMahon LF Jr. Using search engine query data to track pharmaceutical utilization: a study of statins. Am J Manag Care. 2010;16:e215–9.
23. Zhang Z, Zheng X, Zeng DD, Leischow SJ. Tracking dabbing using search query surveillance: a case study in the United States. J Med Internet Res. 2016;18(9):e252.
24. Zheluk A, Quinn C, Meylakhs P. Internet search and Krokodil in the Russian Federation: an infoveillance study. J Med Internet Res. 2014;16(9):e212.

25. Gamma A, Schleifer R, Weinmann W, Buadze A, Liebren M. Could Google Trends be used to predict methamphetamine-related crime? An analysis of search volume data in Switzerland, Germany, and Austria. PLoS ONE. 2016;11(11):e0166566.
26. Eysenbach G. Infodemiology and Infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. J Med Internet Res. 2009;11(1):e11.
27. Zhang Z, Zheng X, Zeng DD, Leischow SJ. Information seeking regarding tobacco and lung cancer: effects of seasonality. PLoS ONE. 2015;10(3):e0117938.
28. Ingram DG, Plante DT. Seasonal trends in restless legs symptomatology: evidence from internet search query data. Sleep Med. 2013;14(12):1364–8.
29. Ingram DG, Matthews CK, Plante DT. Seasonal trends in sleep-disordered breathing: evidence from Internet search engine query data. Sleep Breath. 2015;19(1):79–84.
30. Pollett S, Wood N, Boscardin WJ, Bengtsson H, Schwarcz S, Harriman K, Winter K, Rutherford G. Validating the use of Google Trends to enhance pertussis surveillance in California. PLoS Curr. 2015;19:7.
31. Wang HW, Chen DR, Yu HW, Chen YM. Forecasting the incidence of dementia and dementia-related outpatient visits with Google Trends: evidence from Taiwan. J Med Internet Res. 2015;17(11):e264.
32. Centers for Disease Control and Prevention: HIV/AIDS. https://www.cdc.gov/hiv/basics.html/. Accessed 7 Feb 2018.
33. What are HIV and AIDS? https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids. Accessed 7 Feb 2018.
34. UNAIDS. Fact sheet—latest statistics on the status of the AIDS epidemic. http://www.unaids.org/en/resources/fact-sheet. Accessed 7 Feb 2018.
35. Google. Trends help. how trends data is adjusted. https://support.google.com/trends/answer/4365533?hl=en. Accessed 7 Feb 2018.
36. Scharkow M, Vogelgesang J. Measuring the public agenda using search engine queries. Int J Public Opin Res. 2011;23:104–13.
37. Atlas Plus. Centers for disease control and prevention. https://gis.cdc.gov/grasp/nchhstpatlas/main.html. Accessed 7 Feb 2018.
38. Cervellin Gianfranco, Comelli Ivan, Lippi Giuseppe. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. J Epidemiol Global Health. 2017;7:185–9.
39. Mavragani A, Sampri A, Sypsa K, Tsagarakis KP. Integrating 'Smart Health' in the US Health Care System: asthma Monitoring in the Google Era. JMIR Public Health Surveill. 2018;4(1):e24.
40. Jun SP, Park DH. Consumer information search behavior and purchasing decisions: empirical evidence from Korea. Technol Forecast Soc Change. 2016;31:97–111.
41. Jun SP, Park DH, Yeom J. The possibility of using search traffic information to explore consumer product attitudes and forecast consumer preference. Technol Forecast Soc Change. 2014;86:237–53.
42. Mavragani A, Tsagarakis KP. YES or NO: predicting the 2015 Greferendum results using Google Trends. Technol Forecast Soc. 2016;109:1–5.

Mavragani and Ochoa *J Big Data* (2018) 5:30
https://doi.org/10.1186/s40537-018-0140-9

Journal of Big Data

RESEARCH Open Access

Infoveillance of infectious diseases in USA: STDs, tuberculosis, and hepatitis

Amaryllis Mavragani* and Gabriela Ochoa

*Correspondence:
amaryllis.mavragani1@stir.ac.uk
Department of Computing Science and Mathematics, Faculty of Natural Sciences, University of Stirling, FK9 4LA Stirling, UK

Abstract

Big Data Analytics have become an integral part of Health Informatics over the past years, with the analysis of Internet data being all the more popular in health assessment in various topics. In this study, we first examine the geographical distribution of the online behavioral variations towards Chlamydia, Gonorrhea, Syphilis, Tuberculosis, and Hepatitis in the United States by year from 2004 to 2017. Next, we examine the correlations between Google Trends data and official health data from the '*Centers for Disease Control and Prevention*' (CDC) on said diseases, followed by estimating linear regressions for the respective relationships. The results show that Infoveillance can assist with exploring public awareness and accurately measure the behavioral changes towards said diseases. The correlations between Google Trends data and CDC data on Chlamydia cases are statistically significant at a national level and in most of the states, while the forecasting exhibits good performing results in many states. For Hepatitis, significant correlations are observed for several US States, while forecasting also exhibits promising results. On the contrary, several factors can affect the applicability of this forecasting method, as in the cases of Gonorrhea, Syphilis, and Tuberculosis, where the correlations are statistically significant in fewer states. Thus this study highlights that the analysis of Google Trends data should be done with caution in order for the results to be robust. In addition, we suggest that the applicability of this method is not that trivial or universal, and that several factors need to be taken into account when using online data in this line of research. However, this study also supports previous findings suggesting that the analysis of real-time online data is important in health assessment, as it tackles the long procedure of data collection and analysis in traditional survey methods, and provides us with information that could not be accessible otherwise.

Keywords: Big data, Chlamydia, Gonorrhea, Google trends, Infodemiology, Infoveillance, Health informatics, Hepatitis, Internet behavior, Public health, Sexually transmitted diseases, Syphilis, Tuberculosis

Introduction

Over the past years, with Big Data Analytics being all the more integrated in Health Informatics research, the analysis of Internet data has become a valuable way for monitoring and analyzing the behavior towards health topics. Using data from online sources in order to inform public health and policy is called 'Infodemiology', derived from the words 'Information' and 'Epidemiology' [1]. Infodemiology and Infoveillance (information and surveillance) studies using various online sources, such as Google, Twitter, and

other Social Media [2–6], show the importance of having access to real-time data in health assessment.

Google Trends [7], the most popular tool for retrieving online information, is highly used in health care research [8]. Google Trends data main advantages are that they are real-time data, and that they provide us with the revealed and not the stated preferences [9]. Google Trends has been a useful tool for the analysis, monitoring, forecasting, and nowcasting of many health topics; in seasonal [2, 10], chronic [11–14], and infectious diseases [15–17], as well as in outbreaks and epidemics, such as in AIDS [18], Measles [19], Ebola [20, 21], MERS [22], and the Zika Virus [23–25]. Online queries have been much employed up to this point for the analysis and forecasting of Influenza Like Illness, i.e., the flu [6, 26–28], while an emerging interest in analyzing Google queries for vaccination related topics has been increasing over the last couple of years [19, 29–31]. Other topics that Google Trends data have found significant applicability, include the monitoring of cancer types and screenings [32–35], the relation between online queries and suicide rates [36–39], as well as the analysis of the online interest and its association with both legal [40–42] and illegal drugs [43, 44].

Though Google Trends data have been much employed in forecasting, a gap exists in forecasting diseases' cases using said data. This gap could be mainly attributed to low official health data openness and availability, as well as regional limitations that are due to Internet penetration and restrictions. Traditional methods, e.g., surveys and questionnaires, are time consuming for both collecting and analyzing data, therefore the results are available long after the period to which they refer. In addressing this drawback, online data have exhibited promising results up to this point in this line of research, i.e., showing that Internet data correlate with official health data and further examining the possibility of monitoring and forecasting diseases using data from online sources.

Towards the direction of examining novel, alternative methods of disease surveillance, this study provides an overview of the Infoveillance of five diseases, i.e., Chlamydia, Gonorrhea, Syphilis, Tuberculosis, and Hepatitis, using Google Trends data. Following, we explore the possibility of forecasting said diseases cases in the US at both national and state level. All examined diseases are in the 2018 list of National Notifiable Conditions for Infectious Diseases, i.e., included in the CDC list for Surveillance Case Definitions [45], defined as: "*a set of uniform criteria used to define a disease for public health surveillance. Surveillance case definitions enable public health officials to classify and count cases consistently across reporting jurisdictions*" [46].

For the diseases included in the National Notifiable Infectious Diseases list, the monitoring and analysis of the effects and trends of said diseases is achieved via public health surveillance. Despite provisional data being available in shorter time frames, the official data on the diseases are published annually. This is a long procedure involving a chain of several health officials; hence the data are far from being real time [45].

Out of the notifiable diseases, Chlamydia is the most common one, and is also the most common sexually transmitted disease (STD). It is most frequently met amongst young females, while most of infected people have no symptoms. Chlamydia can have serious effects in a woman's health, even causing infertility. There are increased risks with Chlamydia, such as getting HIV infection, or passing the disease to the baby during delivery. There is a lack of awareness on the subject, while testing does not reach as many women as it should [47].

Gonorrhea is a very common STD, transmitted through the reproductive male and female parts, but also through the mouth and anus. As in the case of Chlamydia, Gonorrhea is mostly asymptomatic, can be passed from mother to child during childbirth, and could even result in infertility. It is prevalent in young adults and African Americans. Gonorrhea also increases the risk of getting HIV [48].

Syphilis is an STD with very serious effects on human health, mainly transmitted through sexual contact or direct contact with infected genitals, anus, and mouth. Congenital Syphilis, i.e., passing the disease from mother to baby, mostly occurs in black and hispanic mothers, which is a very serious complication of the disease and can result in stillbirth or death of the baby. As in Chlamydia and Gonorrhea, the infection of Syphilis increases the risk of HIV transmission. As the symptoms can point to several other diseases, diagnosis of Syphilis can take several months, or even years. The progression of the disease consists of three stages, i.e., Primary Stage. Secondary Stage, and the Latent Stage. Tertiary Syphilis can occur even 30 years after the initial infection and could result in death, while Neurosyphilis and Ocular Syphilis can occur at any stage of the infection, causing serious complications [49].

Tuberculosis (TB) is an infectious disease that mainly affects the lungs and could result in serious complications or death. The risk of TB is higher amongst those with weakened immune systems, as, for example, those with HIV. Tuberculosis is divided in the TB disease and the latent TB infection, i.e., the disease does not develop [50].

Hepatitis is an infectious disease resulting in the inflammation of the liver. It is mainly caused by one of the three most common viruses, i.e., Hepatitis A (HAV), Hepatitis B (HBV), or Hepatitis C (HCV). Hepatitis A is a vaccine preventable, highly contagious disease, and can be transmitted through food, drinks, stool, or through close contact with an infected person. It cannot result in a chronic disease, while it is usually not fatal. On the contrary, Hepatitis B and Hepatitis C can be either acute or chronic, while they can result in serious health issues, even death. Hepatitis B is also vaccine preventable, while for Hepatitis C there is no vaccine yet. Hepatitis B is most commonly transmitted through blood, semen, sexual contact, and needles, while Hepatitis C is most commonly met amongst those who share needles or other drug related equipment [51].

The rest of the paper is structured as follows: In "Data and methods", the data collection procedure and analysis are detailed, and in "Results", the results are presented. "Discussion" consists of the discussion of the analysis, while "Conclusions" presents the overall conclusions and further research suggestions.

Data and methods

Data used in this study are retrieved online by Google Trends [7] and are normalized over the selected period as follows: "*Search results are proportionate to the time and location of a query: Each data point is divided by the total searches of the geography and time range it represents, to compare relative popularity. Otherwise places with the most search volume would always be ranked highest. The resulting numbers are then scaled on a range of 0–100 based on a topic's proportion to all searches on all topics. Different regions that show the same number of searches for a term will not always have the same total search volumes*" [52].

Data on diseases cases and rates are retrieved by CDC's AtlasPlus [53]. This database contains data for 6 infectious diseases, i.e., HIV/AIDS, Chlamydia, Gonorrhea, Syphilis, Tuberculosis, and Hepatitis. Following the well performing forecasting results for AIDS [18], in this study we use data on the rest of the diseases included in AtlasPlus. The data retrieved for Hepatitis are from January 1st, 2004 to December 31st, 2015, while for the rest of the examined diseases; the examined time frame is from January 1st, 2004 to December 31st, 2016. Note that the data may very slightly vary depending on the time of retrieval.

The steps towards examining the possibility of forecasting said diseases using Google Trends data are as follows: First, we provide an overview of the online interest variations on each of these diseases for the respective examined periods. Next, we visualize the geographical distribution of the online interest in each disease for all states for each individual year from 2004 to 2017. Following, we calculate the Pearson correlations between Google Trends data and the respective CDC data on each disease's cases. Finally, we estimate linear regressions for the examined diseases at both national and state level, in order to examine the possibility of forecasting said diseases using Google Trends data.

Results

This section consists of the analysis of the results for the five examined diseases, i.e., Chlamydia, Gonorrhea, Syphilis, Tuberculosis, and Hepatitis.

Chlamydia

Figure 1 consists of the heat map of the online interest for the term 'Chlamydia' by state from January 2004 to December 2016, while Fig. 2 depicts the online interest by state for each year from 2004 to 2017 (Additional file 1: Tables S1 and S2).

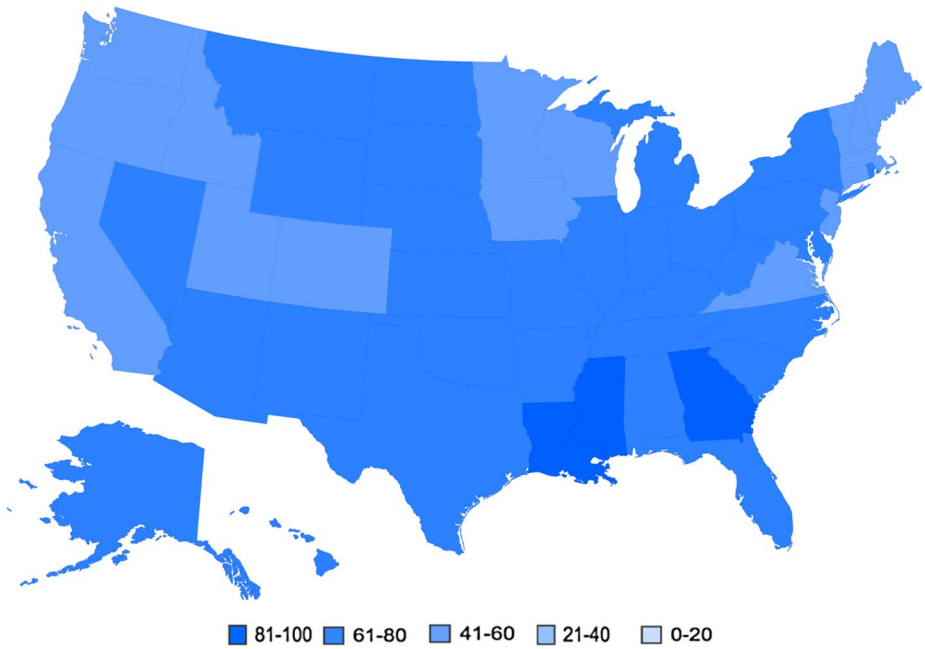


Fig. 1 Heat map of the online interest in the term 'Chlamydia' by state (2004–2016)

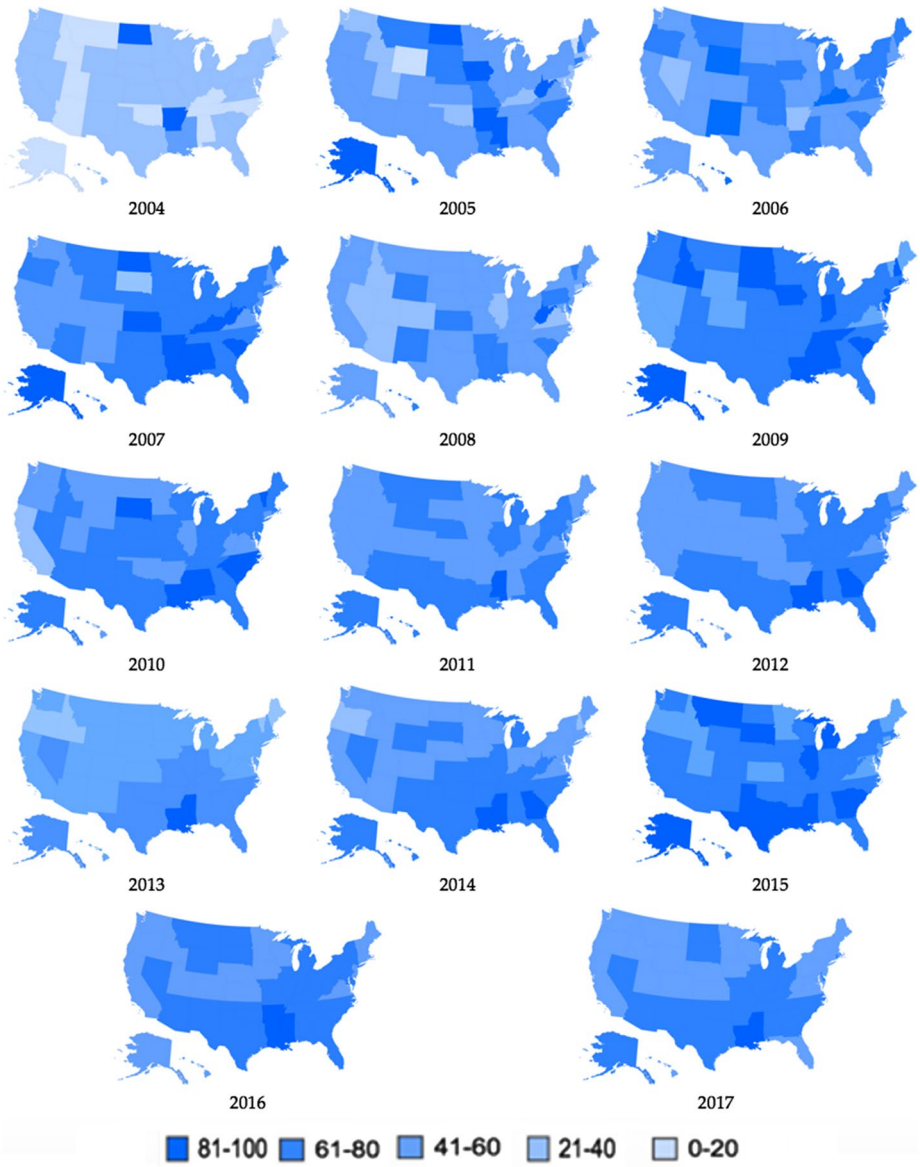


Fig. 2 Online interest heat maps for the term 'Chlamydia' by state by year (2004–2017)

It is evident that the online interest in the term 'Chlamydia' is significant throughout the examined period, i.e., from 2004 to 2017. In the US, the top related searches for the term 'Chlamydia' from 2004 to 2016 include: 'chlamydia symptoms' (100), 'chlamydia gonorrhea' (50), 'symptoms of chlamydia' (38), 'chlamydia men' (36), 'std chlamydia' (34), 'std' (33), 'chlamydia treatment' (33), 'treatment chlamydia' (33), 'chlamydia in men' (28), 'chlamydia infection' (26), 'chlamydia in women' (25), 'what is chlamydia' (24), 'chlamydia test' (22), 'chlamydia symptoms women' (19), 'chlamydia symptoms men' (18), 'chlamydia symptoms in women' (16), 'chlamydia symptoms in men' (16), 'chlamydia discharge' (15), 'chlamydia signs' (14), 'chlamydia cure' (13).

Table 1 consists of the Pearson correlation coefficients between Google Trends data on the term 'Chlamydia' and official Chlamydia cases in each US State from 2004 to 2016.

Table 1 Correlations between Google Trends data and Chlamydia cases by state

State	*r*	State	*r*	State	*r*
Alabama	0.8373***	Kentucky	0.8864***	North Dakota	0.2555
Alaska	0.7691***	Louisiana	0.8771***	Ohio	0.8742***
Arizona	0.8784***	Maine	0.6600**	Oklahoma	0.9208***
Arkansas	0.2461	Maryland	0.7906***	Oregon	0.7691***
California	0.8779***	Massachusetts	0.8744***	Pennsylvania	0.9131***
Colorado	0.8469***	Michigan	0.6276**	Rhode Island	0.8776***
Connecticut	0.7919***	Minnesota	0.7699***	South Carolina	0.6456**
Delaware	0.8278***	Mississippi	−0.1721	South Dakota	0.7496***
DC	0.6606**	Missouri	0.8484***	Tennessee	0.8973***
Florida	0.8845***	Montana	0.7411***	Texas	0.9033***
Georgia	0.9223***	Nebraska	0.9001***	Utah	0.9111***
Hawaii	0.3736	Nevada	0.8578***	Vermont	0.6280**
Idaho	0.8663***	New Hampshire	0.6281**	Virginia	0.7852***
Illinois	0.8585***	New Jersey	0.8305***	Washington	0.8578***
Indiana	0.9119***	New Mexico	0.7714***	West Virginia	0.3165
Iowa	0.6445**	New York	0.8423***	Wisconsin	0.8183***
Kansas	0.8172***	North Carolina	0.9306***	Wyoming	0.5874**

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2 Coefficients α, β, and R^2 of the linear regressions for Chlamydia cases

State	α	β	R^2	State	α	β	R^2	State	α	β	R^2
AL	253	12,263	0.7012	KY	241	2096	0.7856	ND	30	2023	0.0653
AK	88	3805	0.5915	LA	473	13351	0.7694	OH	310	31,751	0.7642
AZ	227	14,831	0.7715	ME	37	1431	0.4356	OK	183	7631	0.8479
AR	78	10,713	0.0606	MD	173	15,209	0.6250	OR	285	2630	0.5915
CA	886	103,581	0.7706	MA	215	7437	0.7646	PA	456	22,216	0.8338
CO	132	11,942	0.7172	MI	171	36,168	0.3939	RI	122	1587	0.7701
CT	148	6320	0.6272	MN	232	4976	0.5927	SC	169	17,417	0.4168
DE	47	2943	0.6852	MS	−17	21,242	0.0296	SD	87	1853	0.5619
DC	4	3887	0.4364	MO	122	19,103	0.7198	TN	151	20,748	0.8052
FL	784	22,622	0.7824	MT	62	1800	0.5493	TX	1040	46,987	0.8159
GA	351	27,004	0.8507	NE	78	3079	0.8103	UT	101	2519	0.8301
HI	34	5146	0.1396	NV	114	5152	0.7358	VT	27	745	0.3944
ID	88	1509	0.7505	NH	38	1343	0.3945	VA	259	17,559	0.6166
IL	293	45,209	0.7370	NJ	374	7834	0.6897	WA	20	11,299	0.7359
IN	309	10,914	0.8315	NM	135	5801	0.5951	WV	50	2774	0.1002
IA	132	4686	0.4154	NY	704	44,661	0.7094	WI	145	15,033	0.6696
KS	140	4467	0.6678	NC	525	11,489	0.8660	WY	43	934	0.3451

At national level, the correlation between the yearly averages of Google Trends data and yearly cases of Chlamydia from 2004 to 2016 is statistically significant ($r = 0.9096$, $p < 0.01$). The correlations are also statistically significant for all states, apart from Arkansas, Mississippi, Hawaii, North Dakota, and West Virginia.

The next step is to identify the relationship between Chlamydia cases and the online interest on the term. Table 2 consists of the coefficients α, β, and the respective R^2 for each of the linear regressions of the form $y = \alpha x + \beta$ estimated for the relationships

between Chlamydia cases (dependent variable) and Google Trends data (independent variable). For the US, the equation describing the relationship is $y = 9012x + 681655$ with an R^2 of 0.8277. Most of the respective models at state level are also performing well, indicating that the forecasting of Chlamydia cases is possible using online search traffic data.

Gonorrhea

Figure 3 depicts the heat map of the online interest in the term 'Gonorrhea' in the US from 2004 to 2016. Figure 4 consists of the heat maps for the online interest of said term for each year from 2004 to 2017 by State (full datasets available in Additional file 1: Tables S3 and S4). As shown in Fig. 4, the online interest by state by year is increasing from 2004 to 2017, with no states in the '0–20' interest group from 2008 on, and with the most states in the interest groups '81–100' and '61–80' being observed after 2014.

The top related searches for the term 'Gonorrhea' in the US from 2004 to 2016 include: 'gonorrhea symptoms' (100), 'symptoms' (98), 'chlamydia' (97), 'chlamydia gonorrhea' (97), 'std' (41), 'gonorrhea std' (40), 'treatment gonorrhea' (35), 'syphilis' (30), 'gonorrhea men' (28), 'herpes' (25), 'what is gonorrhea' (24), 'gonorrhea in women' (23), 'chlamydia and gonorrhea' (22), 'gonorrhea in men' (22), 'gonorrhea symptoms women' (19), 'gonorrhea discharge' (19), 'gonorrhea symptoms men' (18), 'gonorrhea test' (15), 'throat gonorrhea' (15), 'stds' (15).

Table 3 consists of the Pearson correlation coefficients between Google Trends data on the term 'Gonorrhea' from 2004 to 2016 and data on Gonorrhea cases from the CDC for the same period. Contrary to Chlamydia, no statistically significant correlation is observed for USA ($r = 0.0974$, $p > 0.1$), while significant correlations are only observed in the states of Michigan, South Carolina, Alabama, California, Kentucky,

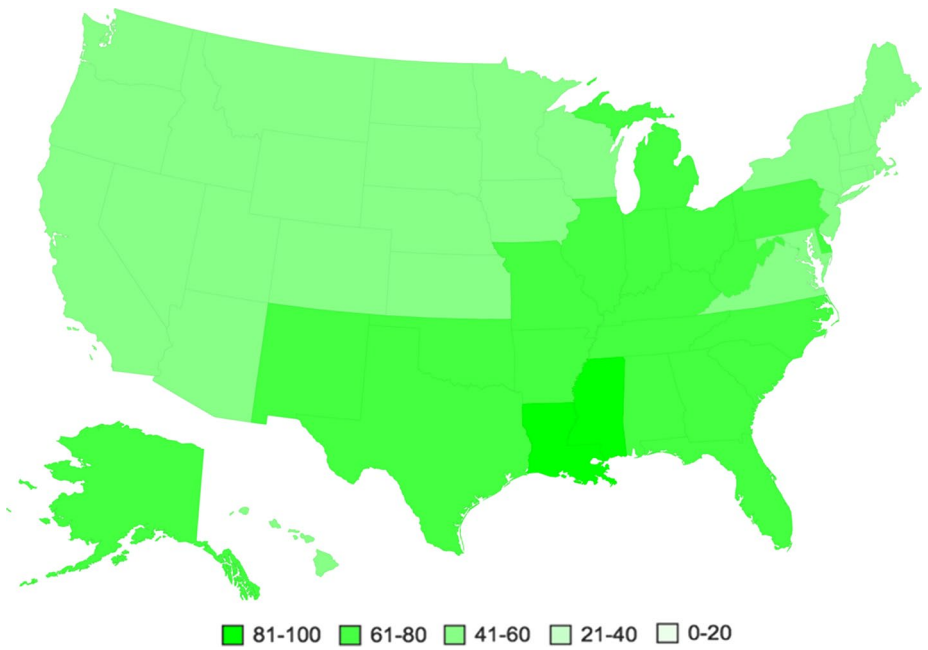


Fig. 3 Heat map of the online interest in the term 'Gonorrhea' by State (2004–2016)

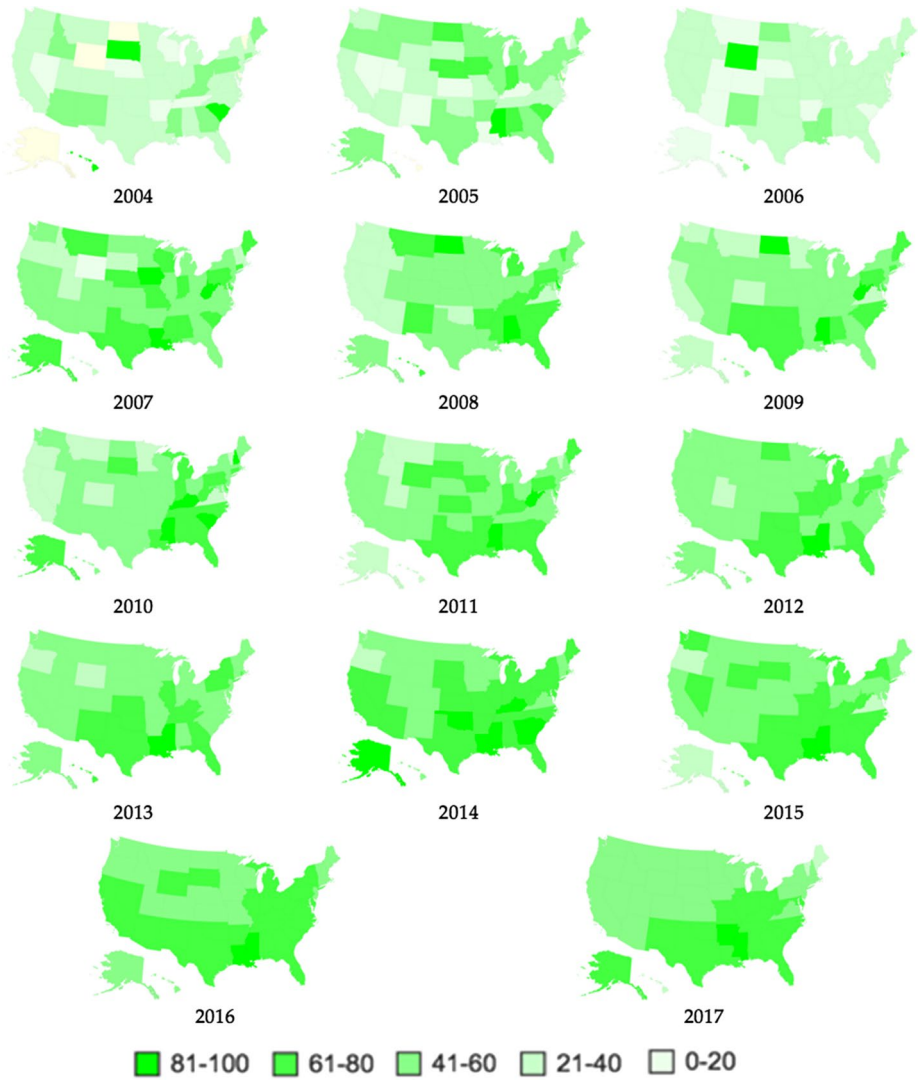


Fig. 4 Online interest heat maps for the term 'Gonorrhea' by state by year (2004–2017)

Mississippi, South Dakota, Texas, Wisconsin, Arizona, Arkansas, Illinois, Louisiana, New York, and Pennsylvania.

Table 4 consists of the coefficients α, β, and the respective R^2 for each of the linear regressions. For the US, the estimated model is $y = 325.28x + 334069$ with an R^2 of 0.0095. In the three States for which significant correlations with $p < 0.01$ are observed, i.e., in Illinois, Michigan, and South Carolina, the respective R^2 for the linear regressions for Gonorrhea cases are 0.6867, 0.5966, and 0.6556.

The R^2 of the estimated equations are not very high even in the states with significant correlations between online and official data on Gonorrhea, while for the US, the results are significantly low. Thus the forecasting of Gonorrhea cases using this method cannot be performed at this point.

Table 3 Correlations between Google Trends data and Gonorrhea cases by state

State	*r*	State	*r*	State	*r*
Alabama	−0.5996**	Kentucky	0.5928**	North Dakota	−0.1005
Alaska	0.2957	Louisiana	−0.5142*	Ohio	−0.7490
Arizona	0.4903*	Maine	0.4675	Oklahoma	0.2069
Arkansas	0.5430*	Maryland	−0.2098	Oregon	0.2629
California	0.5540**	Massachusetts	0.2573	Pennsylvania	0.5140*
Colorado	−0.1122	Michigan	−0.7357***	Rhode Island	−0.4736
Connecticut	−0.0825	Minnesota	−0.0228	South Carolina	−0.8040***
Delaware	0.0856	Mississippi	−0.5825**	South Dakota	0.5805**
DC	0.3097	Missouri	−0.3413	Tennessee	−0.4391
Florida	−0.1847	Montana	0.0953	Texas	0.5624**
Georgia	−0.3326	Nebraska	−0.0830	Utah	0.3331
Hawaii	−0.0990	Nevada	0.1814	Vermont	0.1045
Idaho	0.1987	New Hampshire	−0.0086	Virginia	−0.0348
Illinois	−0.7933*	New Jersey	0.2843	Washington	0.3453
Indiana	−0.4479	New Mexico	−0.0052	West Virginia	−0.4462
Iowa	0.3235	New York	0.5312*	Wisconsin	−0.6704**
Kansas	−0.0925	North Carolina	−0.0271	Wyoming	0.2684

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4 Coefficients α, β, and R^2 of the linear regressions for Gonorrhea cases

State	α	β	R^2	State	α	β	R^2	State	α	β	R^2
AL	−68.71	11,175	0.3595	KY	50.69	2881	0.3515	ND	−5.10	422	0.0101
AK	26.94	637	0.0874	LA	−36.98	11,268	0.2645	OH	−164.46	23,450	0.5609
AZ	59.98	2852	0.2404	ME	13.77	−42	0.2186	OK	15.80	4761	0.0428
AR	32.26	3944	0.2948	MD	−24.73	7773	0.0440	OR	30.32	853	0.0691
CA	344.22	15,916	0.3069	MA	17.81	2196	0.0662	PA	96.82	9535	0.2642
CO	−9.02	3688	0.0126	MI	−193.86	19,933	0.5413	RI	−10.07	715	0.2243
CT	−2.23	2620	0.0068	MN	−2.07	3402	0.0005	SC	−99.84	11,208	0.6464
DE	3.17	1107	0.0073	MS	−148.05	8439	0.3394	SD	17.29	226	0.3370
DC	5.91	2176	0.0959	MO	−54.90	10,334	0.1165	TN	−28.08	9489	0.1928
FL	−43.93	23,410	0.0341	MT	3.95	208	0.0091	TX	188.73	25,163	0.3163
GA	−47.51	18,229	0.1106	NE	−3.52	1518	0.0069	UT	22.724	321	0.1109
HI	−3.97	965	0.0098	NV	19.12	2326	0.0329	VT	0.528	71	0.0109
ID	6.50	141	0.0395	NH	−0.10	181	0.0001	VA	−3.43	8073	0.0012
IL	−124.49	23,886	0.6293	NJ	44.35	5337	0.0808	WA	46.32	1716	0.1192
IN	−49.10	9292	0.2006	NM	−0.64	1857	0.0000	WV	−13.93	1098	0.1991
IA	11.08	1499	0.1046	NY	151.28	13,546	0.2821	WI	−80.09	7863	0.4494
KS	−3.07	2513	0.0086	NC	−5.83	16,119	0.0007	WY	4.29	72	0.0720

Syphilis

Figure 5 depicts the heat map of the online interest in the term 'Syphilis' by state from January 2004 to December 2016, while Fig. 6 consists of the heat maps of the online interest in the term 'Syphilis' by state by year from 2004 to 2017 (Additional file 1: Tables S5 and S6).

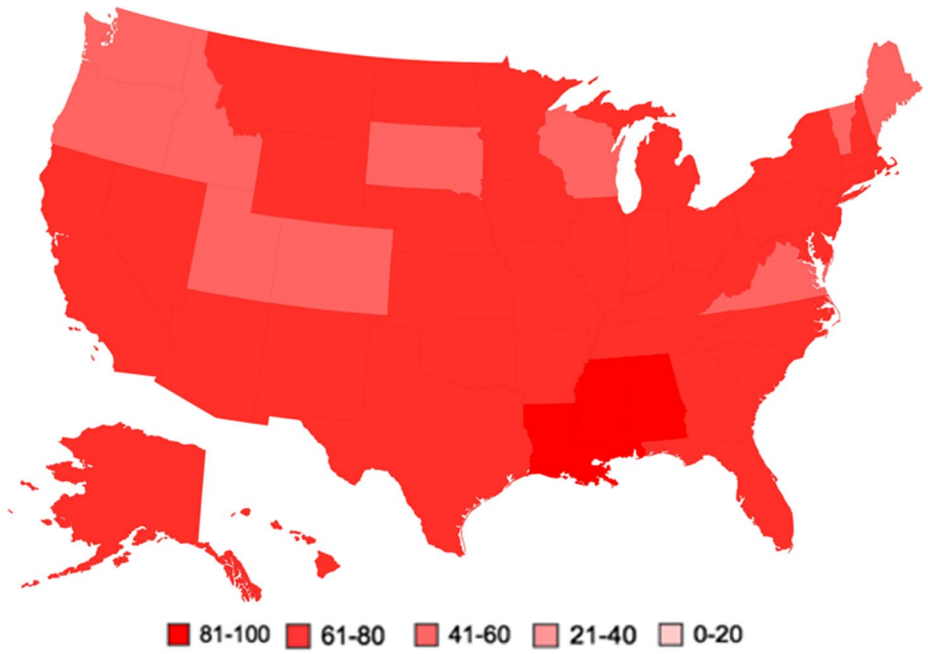


Fig. 5 Heat map of the online interest in the term 'Syphilis' by state (2004–2016)

The top related queries for the term 'Syphilis' from 2004 to 2016 in the US include: 'symptoms syphilis' (97), 'herpes' (37), 'gonorrhea' (36), 'symptoms of syphilis' (34), 'chlamydia' (33), 'std syphilis' (33), 'std' (32), 'what is syphilis' (31), 'syphilis pictures' (28), 'syphilis treatment' (27), 'tuskegee' (25), 'tuskegee syphilis' (25), 'syphilis rash' (24), 'syphilis test' (21), 'hiv' (17), 'tuskegee syphilis study' (16), 'syphilis penis' (15), 'syphilis disease' (15), 'syphilis in men' (14), 'stds' (14), 'gonorrhea symptoms' (13), 'chlamydia symptoms' (12), 'herpes symptoms' (12).

Table 5 consists of the Pearson correlation coefficients between Google Trends data and numbers of Syphilis cases for each examined state. Data on Syphilis cases for calculating the Pearson correlations are retrieved from CDC AtlasPlus [30] by adding the 'Primary and Secondary Syphilis' cases to 'Early Latent Syphilis' cases. Congenital Syphilis' cases are not included, as data are not available for most of the states for most of the years. However, by adding the Congenital Syphilis cases to the analysis, the correlations and the respective results remain significant in the same states. For the years where data for Early Latent Syphilis are not available, only data from 'Primary and Secondary Syphilis' cases are used.

For the US, the correlation between online data and Syphilis cases is statistically significant ($r = 0.6478$, $p < 0.05$). At state level, significant correlations are only observed in California, Illinois, Massachusetts, Utah, in Arkansas, Colorado, DC, Minnesota, Nevada, New Hampshire, North Carolina, Iowa, Michigan, New York, Ohio, and Washington. The states of North Dakota, South Dakota, and Wyoming are excluded from further analysis due to lack of complete datasets in all Syphilis subcategories.

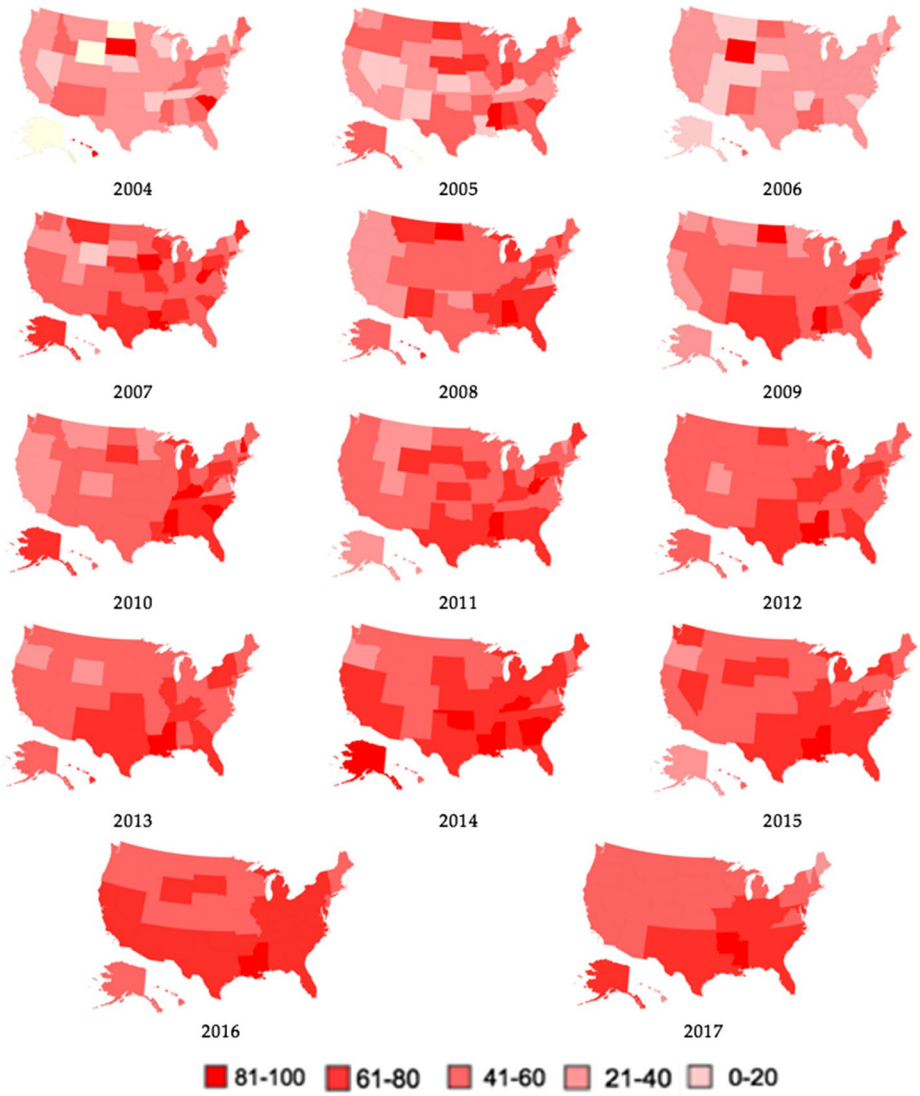


Fig. 6 Online interest heat maps for the term 'Syphilis' by state by year (2004–2017)

Table 6 consists of the coefficients α, β, and the respective R^2 for each of the linear regressions for Syphilis cases. For the US, the equation describing the linear relationship between online data and official Syphilis cases is $y = 748.65x - 26929$ with an R^2 of 0.4196, which is indicating that, though at this point the model is not performing well, we could see promising results in the future when more data are available.

The states where the estimated models perform relatively well are Illinois and Massachusetts, for both of which the estimated correlations between online and official data were high ($p < 0.01$). It is thus evident that, as in the case of Gonorrhea, Syphilis cases cannot be forecasted using this method at this point.

Table 5 Correlations between Google Trends data and Syphilis cases by state

State	*r*	State	*r*	State	*r*
Alabama	−0.2414	Kansas	0.2949	New York	0.5173*
Alaska	0.4024	Kentucky	0.1182	North Carolina	0.6114**
Arizona	0.4722	Louisiana	0.0551	Ohio	0.5523*
Arkansas	0.5739**	Maine	−0.065	Oklahoma	0.4253
California	0.7465***	Maryland	0.2001	Oregon	0.2134
Colorado	0.5662**	Massachusetts	0.8250***	Pennsylvania	0.1238
Connecticut	−0.0757	Michigan	0.4983*	Rhode Island	−0.1962
Delaware	0.1988	Minnesota	0.5806**	South Carolina	0.5695**
DC	0.5640**	Mississippi	−0.0481	Tennessee	0.0385
Florida	0.4942*	Missouri	0.3284	Texas	0.5704**
Georgia	0.5154*	Montana	0.3894	Utah	0.7218***
Hawaii	0.0962	Nebraska	0.1133	Vermont	0.2731
Idaho	0.0983	Nevada	0.6802**	Virginia	0.4594
Illinois	0.7757***	New Hampshire	0.5888**	Washington	0.5350*
Indiana	0.1794	New Jersey	0.1485	West Virginia	0.2697
Iowa	0.5081*	New Mexico	−0.0188	Wisconsin	0.0476

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6 Coefficients α, β, and R^2 of the linear regressions for Syphilis cases

State	α	β	R^2	State	α	β	R^2	State	α	β	R^2
AL	−6.43	759.73	0.0583	KY	2.70	142.52	0.0140	ND	−0.11	10.44	0.0003
AK	0.57	5.89	0.1619	LA	2.63	920.28	0.0030	OH	15.85	−102.13	0.3050
AZ	12.71	5.18	0.2230	ME	−0.26	25.71	0.0042	OK	8.617	30.39	0.1809
AR	14.56	−94.78	0.3294	MD	5.16	488.30	0.0400	OR	8.14	6.60	0.0455
CA	178.81	−5271.90	0.5572	MA	19.09	−259.86	0.6807	PA	16.51	363.47	0.0153
CO	11.33	−156.14	0.3206	MI	14.48	−251.17	0.2484	RI	−1.58	95.88	0.0385
CT	−1.10	145.91	0.0057	MN	12.77	−353.58	0.3371	SC	21.97	−103.92	0.3244
DE	0.81	39.98	0.0395	MS	−2.01	488.04	0.0023	SD	1	8.25	0.0381
DC	5.97	77.35	0.3181	MO	12.77	−7.54	0.1079	TN	1.12	542.05	0.0015
FL	72.64	−908.29	0.2443	MT	0.33	3.42	0.1516	TX	45.87	680.43	0.3253
GA	47.28	−226.06	0.2656	NE	0.60	8.66	0.0128	UT	2.80	−44.59	0.5210
HI	0.94	38.57	0.0093	NV	25.81	−138.21	0.4627	VT	0.04	8.58	0.0003
ID	0.49	22.62	0.0097	NH	1.93	−14.70	0.3467	VA	7.33	139.54	0.2110
IL	51.56	−1385.40	0.6016	NJ	7.25	449.55	0.0221	WA	17.93	−289.54	0.2862
IN	6.47	74.26	0.0322	NM	−0.17	155.34	0.0004	WV	2.06	4.12	0.0728
IA	6.54	−108.68	0.2582	NY	104.91	−2924.50	0.2676	WI	0.54	125.34	0.0023
KS	2.74	27.68	0.0870	NC	39.93	−1288.80	0.3738	WY	−0.0029	2.73	0.00001

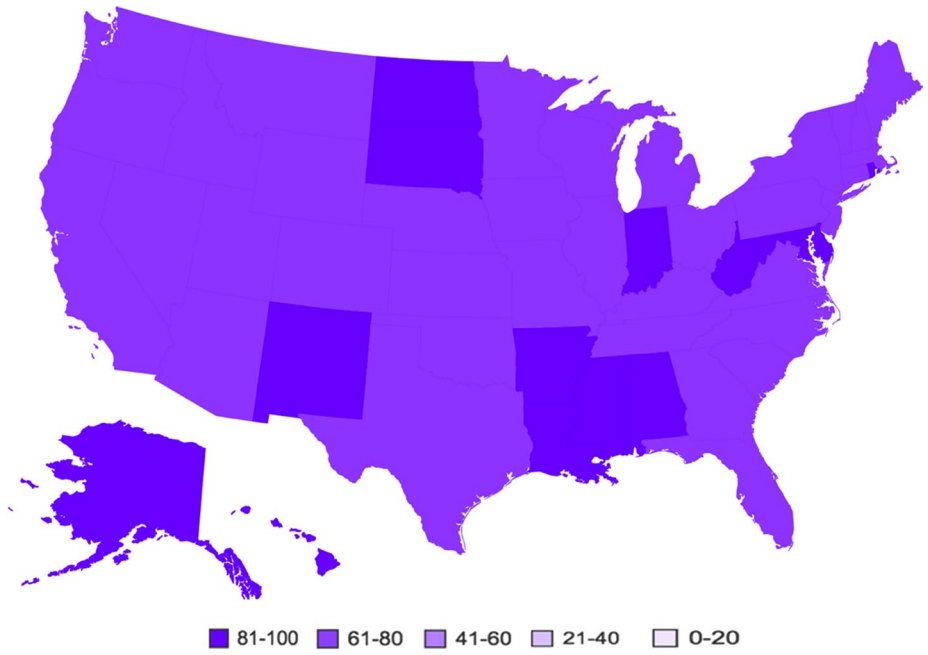


Fig. 7 Heat map of the online interest in the term 'Tuberculosis' by state (2004–2016)

Tuberculosis

Figure 7 consists of the heat map of the online interest by state from January 2004 to December 2016 for the term 'Tuberculosis', while Fig. 8 consists of the respective heat maps by state for each year from 2004 to 2017 (Additional file 1: Tables S7 and S8).

The top related searches for the term 'Tuberculosis' from 2004 to 2016 include 'symptoms tuberculosis' (77), 'tb' (72), 'tuberculosis test' (65), 'mycobacterium tuberculosis' (38), 'tuberculosis treatment' (32), 'symptoms of tuberculosis' (29), 'tuberculosis disease' (29), 'tb test' (19), 'tuberculosis vaccine' (18), 'tuberculosis causes' (14), 'who tuberculosis' (13), 'tuberculosis skin test' (13).

Table 7 consists of the Pearson correlation coefficients (r) between Google Trends data and Tuberculosis cases for each of the states, while Table 8 consists of the coefficients α, β, and the respective R^2 for each of the linear regressions for Tuberculosis cases.

For the US, statistically significant correlations are observed ($r=0.5672$, $p<0.05$) between the online interest on the term 'Tuberculosis' and official Tuberculosis cases. Statistically significant correlations with $p<0.01$ are observed for the states of DC, Louisiana, and Wisconsin, with $p<0.05$ for Illinois, Kentucky, Maryland, New Hampshire, Rhode Island, and Virginia, and with $p<0.1$ for Alabama and California. Based on the calculated correlations, the respective estimated models are not expected to perform well in most of the states.

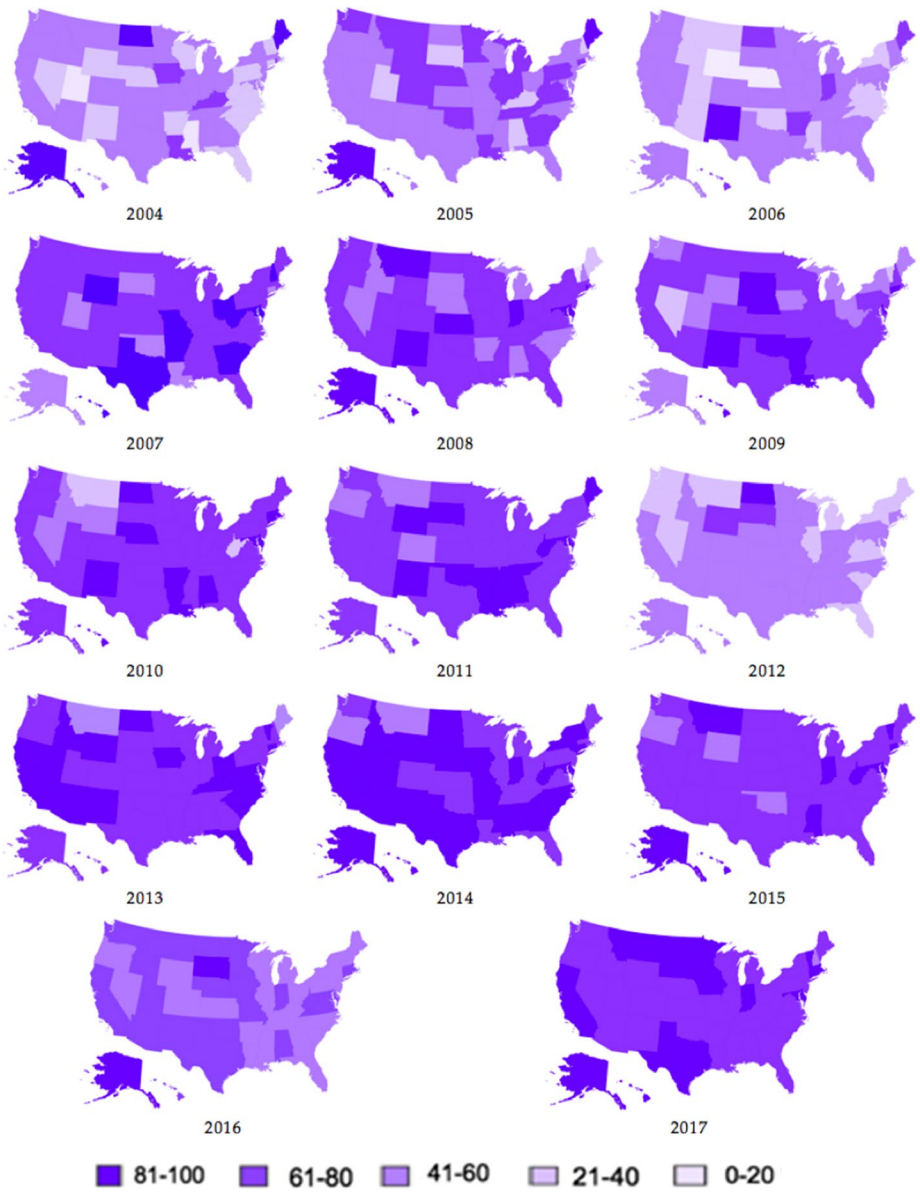


Fig. 8 Online interest heat maps for the term 'Tuberculosis' by state by year (2004–2017)

For the US, the relationship between Google Trends data and Tuberculosis cases is described by $y = 147.51x + 3787$ with an R^2 of 0.3217. The only state that shows promising results that forecasting could be possible at this point is Michigan, with an R^2 of 0.6840. Therefore, as in the case of Gonorrhea and Syphilis, Tuberculosis forecasting is not possible at this point using this method in all states.

Table 7 Correlations between google trends data and Tuberculosis cases by state

State	*r*	State	*r*	State	*r*
Alabama	0.5290*	Kentucky	0.5891**	North Dakota	0.4649
Alaska	0.0859	Louisiana	0.7141***	Ohio	0.4079
Arizona	0.3347	Maine	0.0915	Oklahoma	0.3842
Arkansas	0.3801	Maryland	0.6761**	Oregon	0.3230
California	0.5454*	Massachusetts	0.0513	Pennsylvania	0.6732**
Colorado	0.3382	Michigan	0.8271***	Rhode Island	0.5800**
Connecticut	0.5413*	Minnesota	0.1527	South Carolina	0.3933
Delaware	−0.2075	Mississippi	0.1090	South Dakota	0.2435
DC	0.7382***	Missouri	0.3436	Tennessee	0.2710
Florida	0.1885	Montana	0.2888	Texas	0.3996
Georgia	0.4886*	Nebraska	−0.3154	Utah	0.0570
Hawaii	−0.4057	Nevada	−0.0080	Vermont	0.3065
Idaho	−0.1846	New Hampshire	0.6565**	Virginia	0.5887**
Illinois	0.6608**	New Jersey	0.2505	Washington	0.1680
Indiana	0.2221	New Mexico	0.0315	West Virginia	−0.0706
Iowa	0.2460	New York	0.5450*	Wisconsin	0.7275***
Kansas	−0.0543	North Carolina	0.3604	Wyoming	0.4667

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 8 Coefficients α, β, and R^2 of the linear regressions for Tuberculosis cases

State	α	β	R^2	State	α	β	R^2	State	α	β	R^2
AL	4.40	77.90	0.2799	KY	2.74	41.20	0.3470	ND	1.06	−0.64	0.2161
AK	0.21	53.38	0.0074	LA	5.38	66.08	0.5100	OH	2.78	105.32	0.1664
AZ	2.31	173.65	0.1120	ME	0.10	13.17	0.0084	OK	2.54	36.46	0.1476
AR	0.99	65.08	0.1445	MD	4.49	91.71	0.4571	OR	0.79	60.60	0.1043
CA	25.45	905.03	0.2975	MA	0.30	214.18	0.0026	PA	5.21	50.41	0.4531
CO	1.388	50.28	0.1144	MI	5.23	−30.65	0.6840	RI	0.99	10.47	0.3364
CT	2.38	31.79	0.2931	MN	1.36	137.97	0.0233	SC	3.412	52.04	0.1547
DE	−0.14	25.61	0.0431	MS	0.38	89.77	0.0119	SD	0.21	10.01	0.0593
DC	1.64	−15.24	0.5449	MO	1.41	65.58	0.1180	TN	3.17	123.19	0.0735
FL	4.09	617.84	0.0355	MT	0.23	5.92	0.0834	TX	7.78	1022.40	0.1597
GA	6.80	158.32	0.2387	NE	−0.71	42.07	0.0995	UT	0.04	30.23	0.0033
HI	−0.67	132.59	0.1646	NV	−0.02	94.25	0.0001	VT	0.07	4.17	0.0940
ID	−0.20	18.59	0.0341	NH	0.58	0.25	0.4310	VA	5.12	85.23	0.3466
IL	7.90	48.48	0.4366	NJ	3	283.90	0.0628	WA	0.93	194.91	0.0282
IN	0.40	97.48	0.0493	NM	0.04	45.84	0.0010	WV	−0.08	19.07	0.0050
IA	0.22	40.66	0.0605	NY	17.78	198.30	0.2970	WI	1.69	15.99	0.5293
KS	−0.21	55.58	0.0030	NC	4.28	126.48	0.1299	WY	0.19	0.78	0.2178

Hepatitis

Figure 9 consists of the heat map of the online interest by state from January 2004 to December 2015 for the term 'Hepatitis', while Fig. 10 consists of the respective heat maps by state for each year from 2004 to 2017 (Additional file 1: Tables S9 and S10).

The top related queries include 'symptoms hepatitis' (100), 'hepatitis vaccine' (91), 'what is hepatitis' (66), 'hepatitis b vaccine' (56), 'hepatitis treatment' (44), 'symptoms

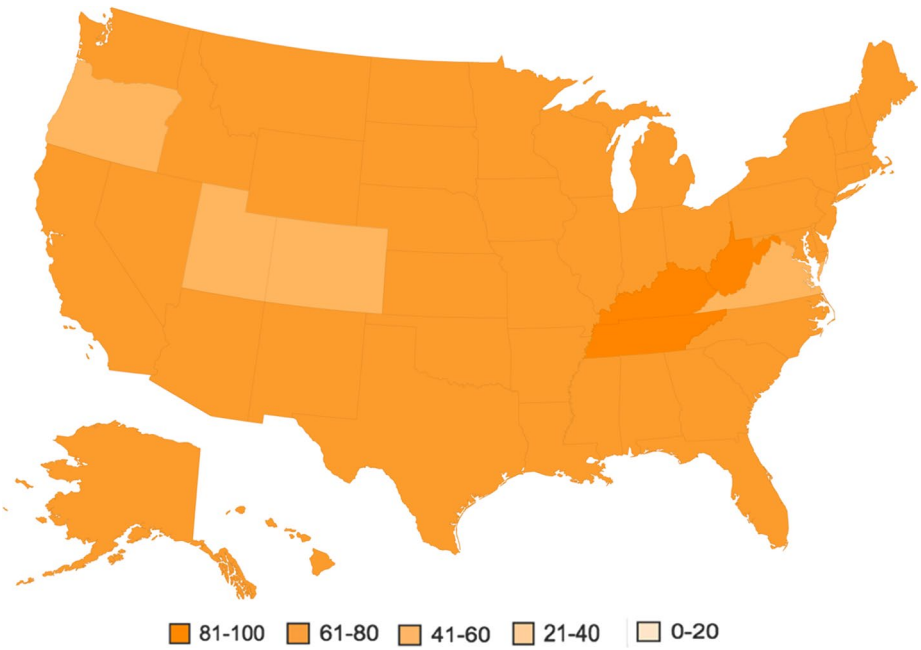


Fig. 9 Heat map of the online interest in the term 'Hepatitis' by state (2004–2015)

hepatitis c' (43), 'symptoms of hepatitis' (41), 'hep' (38), 'hepatitis a vaccine' (35), 'hepatitis test' (30), 'hepatitis virus' (27), 'hepatitis c treatment' (26), 'what is hepatitis c' (26), 'what is hepatitis a' (23), 'hepatitis b symptoms (22), 'viral hepatitis' (21), 'what is hepatitis b' (20), 'hepatitis a symptoms' (20), and 'hepatitis transmission' (17).

Table 9 consists of the Pearson correlation coefficients (r) between Google Trends data and Hepatitis cases for each of the states. For calculating the correlations, the sum of the cases for Hepatitis A, Hepatitis B, and Hepatitis C are used. Where data are not available for a category, the sum of the remaining ones is used.

For the US, statistically significant correlation was observed between Hepatitis cases and Google Trends data ($r = 0.9583$, $p < 0.01$). For Hepatitis A, statistically significant correlations were observed between Google data in the US ($r = 0.9045$, $p < 0.01$); the same for Hepatitis B ($r = 0.8922$, $p < 0.01$). On the other hand, for Hepatitis C cases, no correlation was observed with Google Trends data ($r = -0.3089$, $p > 0.1$), indicating that the latter does not contribute significantly to the high correlation between all Hepatitis cases and Google data.

Table 10 consists of the coefficients α, β, and the respective R^2 for each of the linear regressions for Hepatitis cases for all US States, apart from DC where full datasets are not available.

For the US, the equation describing the linear relationship between Hepatitis cases and Google Trends data is $y = 261.44x - 8197.4$ with an R^2 of 0.9184. The states of Arizona, Florida, Hawaii, New York, Pennsylvania, and Wisconsin exhibit good performing forecasting results. Several other states have R^2 that are relatively high, indicating that they will exhibit better results once more years' data are available.

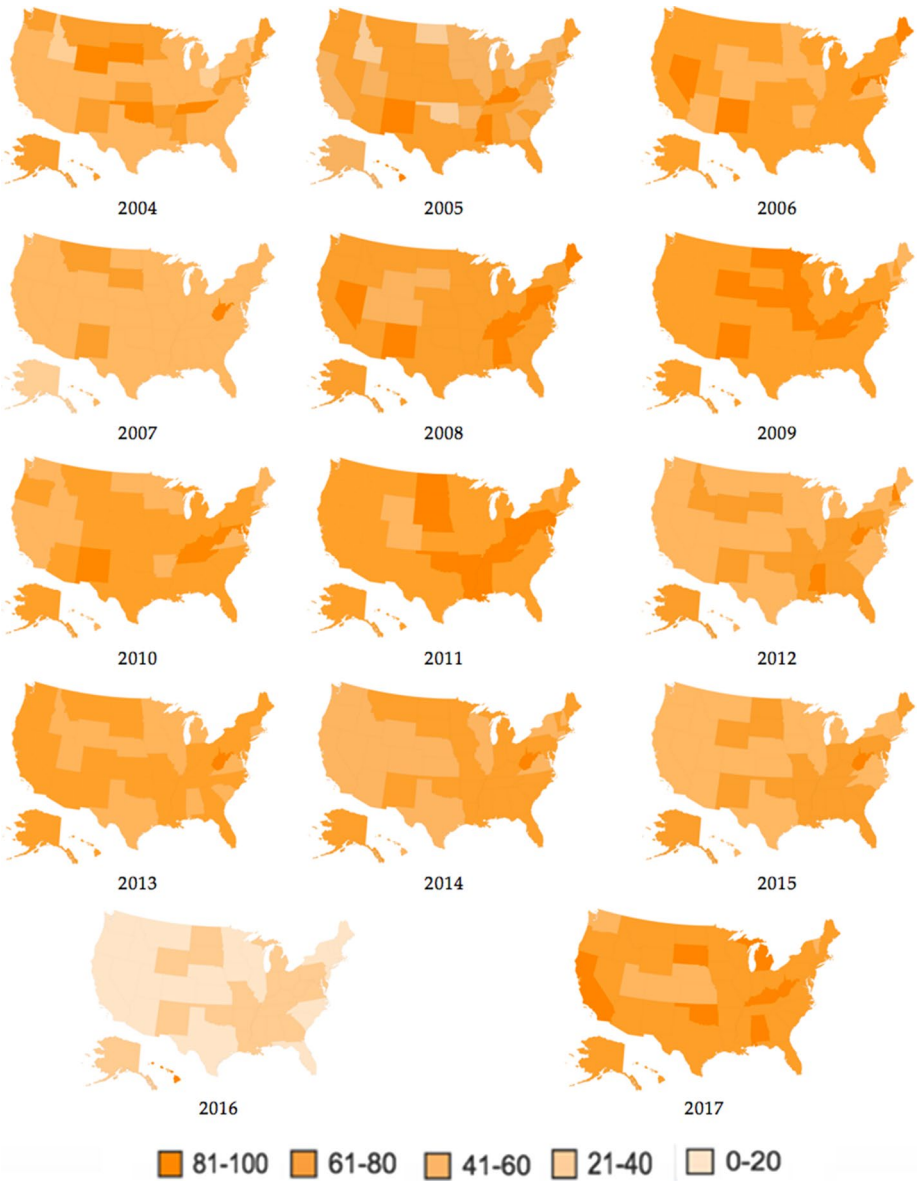


Fig. 10 Online interest heat maps for the term 'Hepatitis' by state by year (2004–2017)

As depicted in Fig. 10, in 2016 the online interest in all states but Hawaii is very low. This can be attributed to the Hepatitis A outbreak in Hawaii in August 2016, possibly linked to raw scallops that were served at a Hawaiian restaurant [54]. This is why the interest is so low in the rest of the states, constituting a good example of how an unexpected event can (negatively) affect this method of forecasting, but also how real life events are immediately and accurately depicted in online searches. The latter is very significant for the real-time examining of epidemics and outbreaks.

Table 9 Correlations between Google Trends data and Hepatitis cases by state

State	*r*	State	*r*	State	*r*
Alabama	0.0012	Louisiana	0.4745	Ohio	0.4040
Alaska	0.1039	Maine	0.3873	Oklahoma	− 0.4900
Arizona	0.9207***	Maryland	0.5980**	Oregon	0.7944***
Arkansas	0.7377***	Massachusetts	0.8010***	Pennsylvania	0.8759***
California	0.8333***	Michigan	0.5740*	Rhode Island	0.3977
Colorado	0.7206***	Minnesota	0.5583*	South Carolina	0.2419
Connecticut	0.7561***	Mississippi	0.6715**	South Dakota	− 0.3825
Delaware	− 0.3014	Missouri	0.6581**	Tennessee	0.3609
Florida	0.9151***	Montana	0.1725	Texas	0.8163***
Georgia	0.7010**	Nebraska	0.5650*	Utah	0.3074
Hawaii	0.8513***	Nevada	0.5200*	Vermont	0.2253
Idaho	0.3770	New Hampshire	0.5045*	Virginia	0.8309***
Illinois	0.5267*	New Jersey	0.7993***	Washington	0.6129**
Indiana	− 0.2965	New Mexico	− 0.4728	West Virginia	0.2579
Iowa	0.3598	New York	0.8631***	Wisconsin	0.8844***
Kansas	0.5213*	North Carolina	0.7576***	Wyoming	0.6561**
Kentucky	− 0.0950	North Dakota	0.4797		

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 10 Coefficients α, β, and R^2 of the linear regressions for Hepatitis cases

State	α	β	R^2	State	α	β	R^2	State	α	β	R^2
AL	0.01	134.60	0.000002	LA	1.50	56.25	0.2252	OH	10.63	− 351.75	0.1632
AK	0.06	7.01	0.0108	ME	0.50	12.88	0.1500	OK	− 5.16	321.44	0.2401
AZ	25.80	− 674.83	0.8477	MD	6.45	− 141.22	0.3576	OR	5.13	− 113.01	0.6311
AR	4.11	− 29.01	0.5442	MA	24.98	− 694.28	0.6416	PA	19.06	− 740.71	0.7673
CA	58.75	− 1762.5	0.6944	MI	7.47	− 102.27	0.3295	RI	1.18	− 20.87	0.1582
CO	2	9.47	0.5192	MN	1.98	15.78	0.3117	SC	1.79	23.98	0.0585
CT	3.75	− 53.98	0.5716	MS	4.37	− 78.62	0.4509	SD	− 0.30	15.12	0.1463
DE	− 1.42	78.13	0.0909	MO	3.55	− 86.62	0.4331	TN	6.21	81.14	0.1303
FL	19.75	− 373.34	0.8374	MT	0.31	9.82	0.0298	TX	44.73	− 1423.4	0.6663
GA	8.28	− 155.43	0.4914	NE	1.01	− 5.35	0.3192	UT	1.29	− 3.21	0.0945
HI	1.41	− 6.58	0.7248	NV	3.08	15.25	0.2704	VT	0.42	0.13	0.0508
ID	0.38	15.24	0.1421	NH	2.59	− 46.34	0.2545	VA	10.11	− 283.41	0.6905
IL	4.10	19.74	0.2775	NJ	10.99	− 237.27	0.6389	WA	1.73	60.33	0.3757
IN	− 3.87	360.46	0.0879	NM	− 0.85	70.28	0.2235	WV	6.34	− 19.29	0.0665
IA	2.19	− 35.28	0.1295	NY	19.72	− 859.60	0.7450	WI	4.80	− 121.40	0.7821
KS	0.75	6.01	0.2718	NC	6.62	− 56.20	0.5739	WY	0.41	− 0.86	0.4305
KY	− 1.81	322.09	0.0090	ND	0.22	0.28	0.2302				

Discussion

The surveillance of diseases using information available online, i.e., Infoveillance, has become an integral part of Health Informatics over the past years. Internet data can provide a large amount of information that could not be accessed through traditional surveillance methods, such as questionnaires, surveys, and registries. New methods and approaches are constantly discovered and used in order to take advantage of what the Internet has to offer.

Table 11 CDC reported cases for the infectious diseases included in AtlasPlus in 2016

Disease	Reported cases
Chlamydia	1,598,354
Gonorrhea	468,514
Primary and Secondary Syphilis	27,814
Tuberculosis	9272
Hepatitis (A, B, and C)	7170

In this study, we assessed the online interest in the US at both national and state level in five infectious diseases, in order to show how Internet data can be used in the Infoveillance of said diseases, and explore the possibility of forecasting cases using online search traffic data.

Yearly Data from the Atlas CDC website [53] were used, which are available for up to 2015 or 2016 (depending on the disease) for Chlamydia, Gonorrhea, Syphilis, Tuberculosis, and Hepatitis. In the case of AIDS, the estimated forecasting models of AIDS Prevalence in the US exhibited very good performance [18], supporting previous work on the subject suggesting that empirical relationships between online data and official health data exist, and highlighting the usefulness of this tool in health assessment.

As is evident from the geographical distribution of the online interest towards the examined diseases in each state per year since 2004, Google Trends data are an accurate and valuable way to measure public interest and awareness on the subject. This is essential especially for STDs, since new innovative public surveillance methods, preventive measures, and increased public information via traditional and new channels can increase awareness, particularly in the regions where said diseases' rates are higher.

Table 11 consists of the US CDC reported cases for the diseases included in Atlas for the year 2016, apart from Hepatitis for which data refer to the year 2015. As is evident, Chlamydia cases are by far the most. The latter could explain why statistically significant correlations are observed between Google Trends data and reported Chlamydia cases in most US States, and the forecasting models are performing well. All diseases apart from Tuberculosis are experiencing an increase since the previous year, indicating that probably better- and for more diseases- forecasting will be possible in the future using this method.

Table 12 consists of the USA yearly rates (per 100,000) for Chlamydia, Gonorrhea, Syphilis, Tuberculosis from 2004 to 2016, and Hepatitis from 2004 to 2015. For Hepatitis, the reported rate is the sum of rates from Hepatitis A, Hepatitis B, and Hepatitis C, while for Syphilis, the rate is the sum of Primary and Secondary Syphilis, Early Latent Syphilis, and Congenital Syphilis.

As shown in Table 12, Chlamydia rates in the US are significantly higher than the rates for the rest of the examined diseases. This partly explains why Chlamydia cases exhibit so high correlations with online search traffic data and why the forecasting of Chlamydia is possible in many states using Google Trends data. For Syphilis and Tuberculosis, the rates included in Table 12 show that said diseases have very decreased rates, with Tuberculosis showing a downward trend since 2004. The low rates can partly explain why this method does not apply to these diseases. This is contrary to the case of Hepatitis, which may have the lowest numbers of reported cases (Table 11) and a downward rate trend

Table 12 CDC reported yearly rates in USA for the examined diseases from 2004 to 2016

	Chlamydia	Gonorrhea	Syphilis	Tuberculosis	Hepatitis
2004	317.3	112.7	14.5	5	4.3
2005	330.3	114.9	14	4.8	3.5
2006	345.4	120.1	15.1	4.6	3.1
2007	367.7	118.1	17.5	4.4	2.8
2008	398	110.7	19	4.2	2.4
2009	405.7	98.2	19.3	3.8	2
2010	422.8	100	18.6	3.6	1.9
2011	453.2	103.2	17.8	3.4	1.7
2012	453	106.6	18	3.2	2
2013	443	105.2	20.1	3	2.1
2014	452.1	109.8	24	3	2
2015	475	123	27.2	3	2.2
2016	497.3	145.8	33.4	2.9	–

(Table 12), but it shows more promising results in forecasting. Based on the observations for Tuberculosis and Syphilis, however, and as in 29 out of 50 states significant correlations are observed for Hepatitis cases and online queries, there is a slight possibility that what is observed is a decrease in significance of the reported results instead of a projected increase in the future. For Gonorrhea, the online behavioral assessment is not trivial, as it is a word that is often misspelled, mostly for 'Gonorrea', contrary to e.g., AIDS, which is a word that is not misspelled, and for which the forecasting results exhibit good performance.

Many factors should be taken into account when using online search traffic data in health assessment, and the results should be interpreted carefully. This study is an overview of how infoveillance methods can be applied in monitoring and forecasting diseases cases using online search traffic data. In this analysis, we highlight not only what studies in this field normally highlight, i.e., the usefulness of Internet data in the monitoring and forecasting of diseases' prevalence, but also provide examples of cases where this method does not work. In fact, we emphasize on how the suitability of this method along with the respective forecasting results can be affected by low rates or other factors.

However, despite previous concerns on the reliability using Google data as a means for disease monitoring [55], including the case of *Google Flu Trends* [56] which is now not available [57], the use of Google Trends data in health and medicine has exhibited very promising results so far. Nevertheless, it is essential to understand that this method cannot be applied in every case, and, more importantly, that the methodology should be designed cautiously and that the results must always be interpreted accordingly. Taking into account these limitations, future research should focus on employing more detailed and complicated mathematical modeling in order to improve diseases' and epidemics' forecasting, as, in order for all available information to be integrated in health research, both online data and data from traditional sources should be combined [56].

The overall assessment of the diseases examined in this study indicate the usefulness of Google Trends as a tool for disease surveillance, providing real-time data and thus

tackling the disadvantage of time consuming traditional data collection and analysis methods.

Conclusions

Over the past decade, the analysis of online search traffic data has been shown valuable and useful in the assessment of public health issues. In this study, by examining the geographical distribution of the online behavioral variations towards Chlamydia, Gonorrhea, Syphilis, Tuberculosis, and Hepatitis in the US by year since 2004, we showed how Infoveillance can explore public awareness and accurately measure the behavior towards said diseases. Next, we examined the correlations between Google Trends data and CDC data for the reported diseases. For Chlamydia, statistically significant correlations were observed for the US as a whole and most of the states, while their relationship was well described by the linear regressions estimated for many states. For Hepatitis, significant correlations were observed in 29 states, while forecasting seems to be exhibiting promising results at this point. On the contrary, for Syphilis and Tuberculosis the correlations were statistically significant in less states, which can be partly explained by the very low rates of said diseases in the US. For Gonorrhea, however, though rates are high in the US, the results were not significant as well. The latter could be due to the high volumes of Internet users that search for the disease with incorrect spelling, highlighting one of the main limitations of the tool, and being a good example of why the selection of keywords and the interpretation of the results when using online search traffic data are crucial for the robustness of the analysis. Overall, this study indicates that the analysis of real time data of diseases is important for obtaining information that cannot be accessible through traditional survey methods. Future research on the subject could focus on developing new methods of monitoring and analysis of health issues, as well as overcoming the limitations highlighted in this study.

Additional file

Additional file 1. Additional tables.

Authors' contributions
AM collected the data, performed the analysis, and wrote the paper. GO had the overall supervision. Both authors read and approved the final manuscript.

Acknowledgements
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
All data used in this study are publicly available and accessible in the cited sources.

Consent for publication
The authors consent to the publication of this work.

Ethics approval and consent to participate
Not applicable.

Funding
Not applicable.

Publisher's Note

Received: 8 June 2018 Accepted: 22 August 2018
Published online: 06 September 2018

References

1. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. J Med Internet Res. 2009;11(1):e11.
2. Mavragani A, Sampri A, Sypsa K, Tsagarakis KP. Integrating Smart Health in the US Health Care system: infodemiology Study of asthma monitoring in the Google era. JMIR Public Health Surveill. 2018;4(1):e24.
3. Roccetti M, Marfia G, Salomoni P, Prandi C, Zagari MR, Gningaye Kengni LF, et al. Attitudes of Crohn's disease patients: infodemiology case study and sentiment analysis of facebook and twitter posts. JMIR Public Health Surveill. 2017;3(3):e51.
4. van Lent GGL, Sungur H, Kunneman AF, van de Velde B, Das E. Too far to care? Measuring public attention and fear for ebola using twitter. J Med Internet Res. 2017;19(6):e193.
5. Wongkoblap A, Vadillo AM, Curcin V. Researching mental health disorders in the era of social media: systematic review. J Med Internet Res. 2017;19(6):e228.
6. Lu SF, Hou S, Baltrusaitis K, Shah M, Leskovec J, Sosic R, et al. Accurate influenza monitoring and forecasting using novel internet data streams: a case study in the Boston Metropolis. JMIR Public Health Surveill. 2018;4(1):e4.
7. Google Trends. https://trends.google.com/trends/explore. Accessed 8 May 2018.
8. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of google trends in health care research: a systematic review. PLoS ONE. 2014;9(10):e109583.
9. Mavragani A, Tsagarakis KP. YES or NO: predicting the 2015 GReferendum results using Google Trends. Technol Forecast Soc. 2016;109:1–5.
10. Ingram DG, Matthews CK, Plante DT. Seasonal trends in sleep-disordered breathing: evidence from Internet search engine query data. Sleep Breath. 2015;19(1):79–84.
11. Wang HW, Chen DR, Yu HW, Chen YM. Forecasting the incidence of dementia and dementia-related outpatient visits with google trends: evidence from Taiwan. J Med Internet Res. 2015;17(11):e264.
12. Brigo F, Lochner P, Tezzon F, Nardone R. Web search behavior for multiple sclerosis: an infodemiological study. Multiple Sclerosis and Related Disorders. 2014;3(4):440–3.
13. Bragazzi NL. Infodemiology and Infoveillance of Multiple Sclerosis in Italy. Multiple Scler Int. 2013;2013:9.
14. Bragazzi NL, Bacigaluppi S, Robba C, Nardone R, Trinka E, Brigo F. Infodemiology of status epilepticus: a systematic validation of the Google Trends-based search queries. Epilepsy Behav. 2016;55:120–3.
15. Zhou X, Ye J, Feng Y. Tuberculosis surveillance by analyzing google trends. IEEE Trans Biomed Eng. 2011;58(8):2247–54.
16. Johnson AK, Mehta SD. A comparison of internet search trends and sexually transmitted infection rates using google trends. Sex Transm Dis. 2014;41(1):61–3.
17. Rohart F, Milinovich GJ, Avril SMR, Lê Cao K-A, Tong S, Hu W. Disease surveillance based on Internet-based linear models: an Australian case study of previously unmodeled infection diseases. Sci Rep. 2016;6:38522.
18. Mavragani A, Ochoa G. Forecasting AIDS prevalence in the united states using online search traffic data. J Big Data. 2018;5:17.
19. Mavragani A, Ochoa G. The internet and the anti-vaccine movement: tracking the 2017 EU measles outbreak. Big Data Cog Comput. 2018;2(1):2.
20. Alicino C, Bragazzi NL, Faccio V, Amicizia D, Panatto D, Gasparini R, et al. Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes. Infect Dis Poverty. 2015;4(1):54.
21. Hossain L, Kam D, Kong F, Wigand RT, Bossomaier T. Social media in Ebola outbreak. Epidemiol Infect. 2016;144(10):2136–43.
22. Poletto C, Bolle PY, Colizza V. Risk of MERS importation and onward transmission: a systematic review and analysis of cases reported to WHO. BMC Infect Dis. 2016;16(1):448.
23. Farhadloo M, Winneg K, Chan MPS, Albarracin D. Associations of topics of discussion on twitter with survey measures of attitudes, knowledge, and behaviors related to Zika: probabilistic study in the United States. JMIR Public Health Surveill. 2018;4(1):e16.
24. Majumder SM, Santillana M, Mekaru RS, McGinnis PD, Khan K, Brownstein SJ. 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. 2016;2(1):e30.
25. Scatà M, Di Stefano A, Liò P, La Corte A. The impact of heterogeneity and awareness in modeling epidemic spreading on multiplex networks. Sci Rep. 2016;6:37105.
26. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. Proc Natl Acad Sci. 2015;112(47):14473.
27. Kang M, Zhong H, He J, Rutherford S, Yang F. Using Google Trends for influenza surveillance in South China. PLoS ONE. 2013;8(1):e55205.
28. Domnich A, Panatto D, Signori A, Lai PL, Gasparini R, Amicizia D. Age-related differences in the accuracy of web query-based predictions of Influenza-Like Illness. PLoS ONE. 2015;10(5):e0127754.
29. Bragazzi NL, Barberis I, Rosselli R, Gianfredi V, Nucci D, Moretti M, et al. How often people google for vaccination: qualitative and quantitative insights from a systematic search of the web-based activities using Google Trends. Hum Vaccines Immunotherap. 2017;13(2):464–9.

30. Warren KE, Wen LS. Measles, social media and surveillance in Baltimore City. J Public Health. 2017;39(3):e73–8.
31. Berlinberg EJ, Deiner MS, Porco TC, Acharya NR. Monitoring Interest in Herpes Zoster Vaccination: Analysis of Google Search Data. JMIR Public Health Surv. 2018;4(2):e10180.
32. Phillips CA, Barz Leahy A, Li Y, Schapira MM, Bailey LC. Merchant RM relationship between state-level google online search volume and cancer incidence in the united states: retrospective study. J Med Internet Res. 2018;20(1):e6.
33. Schootman M, Toor A, Cavazos-Rehg P, Jeffe DB, McQueen A, Eberth J, et al. The utility of Google Trends data to examine interest in cancer screening. BMJ Open. 2015;5(6):e006678.
34. Zhang Z, Zheng X, Zeng DD, Leischow SJ. Information seeking regarding tobacco and lung cancer: effects of seasonality. PLoS ONE. 2015;10(3):e0117938.
35. Foroughi F, Lam KYA, Lim SCM, Saremi N, Ahmadvand A. Googling for Cancer: An Infodemiological Assessment of Online Search Interests in Australia, Canada, New Zealand, the United Kingdom, and the United States. JMIR Cancer. 2016;2(1):e5.
36. Solano P, Ustulin M, Pizzorno E, Vichi M, Pompili M, Serafini G, et al. A Google-based approach for monitoring suicide risk. Psychiatry Res. 2016;246:581–6.
37. Arora VS, Stuckler D, McKee M. Tracking search engine queries for suicide in the United Kingdom, 2004–2013. Public Health. 2016;137:147–53.
38. Fond G, Gaman A, Brunel L, Haffen E, Llorca PM. Google Trends®: ready for real-time suicide prevention or just a Zeta-Jones effect? An exploratory study. Psychiatry Res. 2015;228(3):913–7.
39. Parker J, Cuthbertson C, Loveridge S, Skidmore M, Dyar W. Forecasting state-level premature deaths from alcohol, drugs, and suicides using Google Trends data. J Affect Disord. 2017;213:9–15.
40. Mavragani A, Sypsa K, Sampri A, Tsagarakis KP. Quantifying the UK online interest in substances of the EU watch list for water monitoring: diclofenac, estradiol, and the macrolide antibiotics. Water. 2016;8(11):542.
41. Schuster NM, Rogers MA, McMahon LF Jr. Using search engine query data to track pharmaceutical utilization: a study of statins. Am J Manag Care. 2010;16(8):e215–9.
42. Gahr M, Uzelac Z, Zeiss R, Connemann BJ, Lang D, Schönfeldt-Lecuona C. Linking annual prescription volume of antidepressants to corresponding web search query data: a possible proxy for medical prescription behavior? J Clin Psychopharmacol. 2015;35(6):681–5.
43. Zhang Z, Zheng X, Zeng DD, Leischow SJ. Tracking dabbing using search query surveillance: A case study in the United States. J Med Internet Res. 2016. https://doi.org/10.2196/jmir.5802.
44. Zheluk A, Quinn C, Meylakhs P. Internet search and Krokodil in the Russian Federation: an infoveillance study. J Med Internet Res. 2014. https://doi.org/10.2196/jmir.3203.
45. Centers for Disease Control and Prevention. National notifiable diseases surveillance system (NNDSS). About notifiable infectious diseases and conditions data. https://wwwn.cdc.gov/nndss/infectious.html. Accessed 1 June 2018.
46. Centers for Disease Control and Prevention. National notifiable diseases surveillance system (NNDSS). surveillance case definitions. https://wwwn.cdc.gov/nndss/case-definitions.html. Accessed 1 June 2018.
47. Centers for Disease Control and Prevention. Sexually transmitted diseases (STDs). Chlamydia. Available at: https://www.cdc.gov/std/stats16/chlamydia.htm. Accessed 1 June 2018.
48. Centers for Disease Control and Prevention. Sexually transmitted diseases (STDs). Gonorrhea. https://www.cdc.gov/std/gonorrhea/stdfact-gonorrhea.htm. Accessed 1 June 2018.
49. Centers for Disease Control and Prevention. Sexually transmitted diseases (STDs). Syphilis. https://www.cdc.gov/std/syphilis/stdfact-syphilis-detailed.htm. Accessed 1 June 2018.
50. Centers for Disease Control and Prevention. Tuberculosis (TB). https://www.cdc.gov/tb/default.htm. Accessed 1 June 2018.
51. Centers for Disease Control and Prevention. Viral Hepatitis. https://www.cdc.gov/hepatitis/index.htm. Accessed 1 June 2018.
52. Google Trends. How data is adjusted. https://support.google.com/trends/answer/4365533?hl=en. Accessed 22 May 2018.
53. Centers for Disease Control and Prevention. NCHHSTP Atlas Plus. https://www.cdc.gov/nchhstp/atlas/index.htm. Accessed 8 May 2018.
54. Centers for Disease Control and Prevention. Viral hepatitis. https://www.cdc.gov/hepatitis/outbreaks/2016/hav-hawaii.htm. Accessed 30 May 2018.
55. Cervellin G, Comelli I, Lippi G. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. J Epidemiol Global Health. 2017;7:185–9.
56. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science. 2017;343(6176):1203–5.
57. Google Flu Trends. https://www.google.org/flutrends/about/. Accessed 8 Aug 2018.

Original Paper

Tracking COVID-19 in Europe: Infodemiology Approach

Amaryllis Mavragani, BSc, MSc

Department of Computing Science and Mathematics, Faculty of Natural Sciences, University of Stirling, Stirling, United Kingdom

Corresponding Author:
Amaryllis Mavragani, BSc, MSc
Department of Computing Science and Mathematics
Faculty of Natural Sciences
University of Stirling
University Campus
Stirling, FK94LA
United Kingdom
Phone: 44 7523782711
Email: amaryllis.mavragani1@stir.ac.uk

Abstract

Background: Infodemiology (ie, information epidemiology) uses web-based data to inform public health and policy. Infodemiology metrics have been widely and successfully used to assess and forecast epidemics and outbreaks.

Objective: In light of the recent coronavirus disease (COVID-19) pandemic that started in Wuhan, China in 2019, online search traffic data from Google are used to track the spread of the new coronavirus disease in Europe.

Methods: Time series from Google Trends from January to March 2020 on the Topic (Virus) of "Coronavirus" were retrieved and correlated with official data on COVID-19 cases and deaths worldwide and in the European countries that have been affected the most: Italy (at national and regional level), Spain, France, Germany, and the United Kingdom.

Results: Statistically significant correlations are observed between online interest and COVID-19 cases and deaths. Furthermore, a critical point, after which the Pearson correlation coefficient starts declining (even if it is still statistically significant) was identified, indicating that this method is most efficient in regions or countries that have not yet peaked in COVID-19 cases.

Conclusions: In the past, infodemiology metrics in general and data from Google Trends in particular have been shown to be useful in tracking and forecasting outbreaks, epidemics, and pandemics as, for example, in the cases of the Middle East respiratory syndrome, Ebola, measles, and Zika. With the COVID-19 pandemic still in the beginning stages, it is essential to explore and combine new methods of disease surveillance to assist with the preparedness of health care systems at the regional level.

(JMIR Public Health Surveill 2020;6(2):e18941) doi: 10.2196/18941

KEYWORDS

big data; coronavirus; COVID-19; infodemiology; infoveillance; Google Trends

Introduction

In December 2019, Chinese researchers identified a novel coronavirus in humans that caused acute respiratory syndrome—officially called coronavirus disease (COVID-19) as of February 11, 2020 [1]. China reported its first death on January 11, 2020, and Wuhan in the Hubei province, which was identified as the epicenter of the epidemic, was cut off by Chinese authorities on January 23, 2020 [2].

COVID-19 quickly surpassed the death toll of the severe acute respiratory syndrome (SARS) pandemic on February 9, 2020 [2]. The virus had already spread to several other Chinese regions, quickly affecting many neighboring countries as well, like the Philippines and South Korea [2]. Several cases of COVID-19 were reported throughout Europe over the next days without causing any regional epidemic at the time; although this did not last long, with Italy having its first death on February 21, 2020 [3], which in a short time spread to all European countries, resulting in the World Health Organization declaring it a pandemic on March 11, 2020 [4].

As of March 25, 2020, COVID-19 cases have surpassed 471,000 worldwide, with more than 335,000 still active, and with more than 21,000 deaths. The country with the most confirmed COVID-19 cases is the United States with 81,864, almost half of which are in the state of New York. Italy is the most affected country in number of deaths as of March 25, with 74,386 cases and 7503 deaths. Lombardy, the origin of the Italy epidemic, is the most affected region, followed by Emilia-Romagna,

Veneto, Piedmont, Marche, Tuscany, and Liguria. In Europe, Spain is unfortunately following Italy's curve, with 49,515 cases and 3647 deaths. Both countries have surpassed China's 3287 reported COVID-19 death toll. France and Germany are also facing a difficult situation, with more than 29,155 and 43,646 confirmed cases, respectively. All European countries have COVID-19 cases, and most countries have at least one death. However, there is a clear geographical distribution of COVID-19 cases in Europe, with central and southwest Europe being the most affected. Figure 1 depicts the current situation in COVID-19 cases worldwide up to March 25, 2020, while Figure 2 shows the COVID-19 (total cumulative, not per capita) deaths by country up to March 25, 2020. All data on COVID-19 cases and deaths were retrieved from Worldometer [5].

Figure 1. Worldwide heat map for total COVID-19 cases by country (as of March 25, 2020).

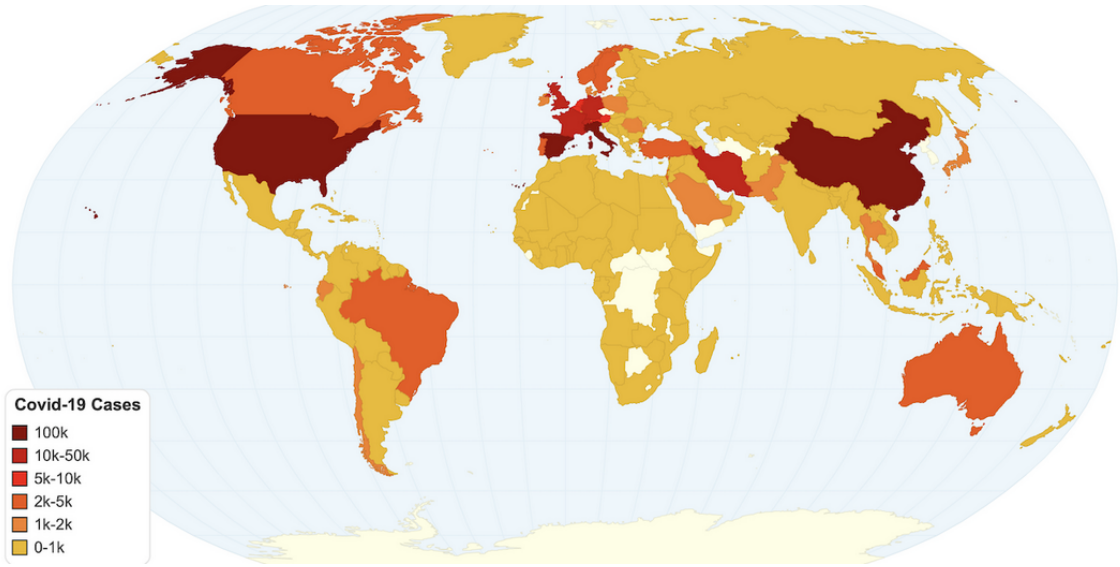
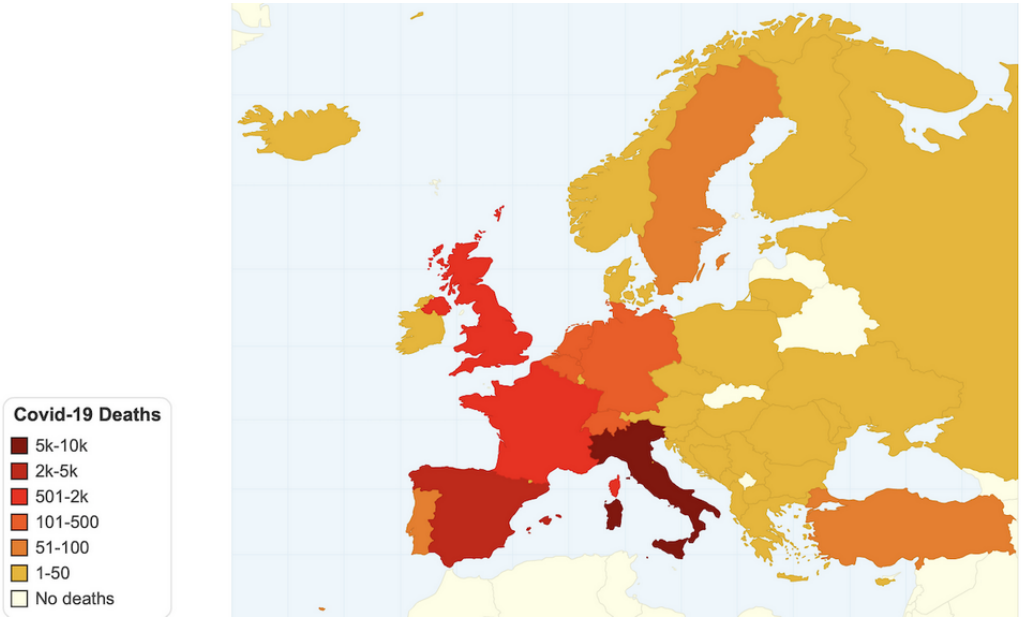


Figure 2. European heat map for total COVID-19 deaths by country (as of March 25th, 2020).



Italy is the first country facing serious issues and a large number of deaths due to COVID-19 in Europe, followed by Spain, France, Germany, and the United Kingdom [5]. The main issue in all affected countries is that of the health systems' capabilities and performance. Toward this direction and based on early Italian data about the spread of the disease, all European countries have taken measures aiming at "flattening the curve" [6], meaning to spread the cases—and, consequently, the patients that need to be admitted to the intensive care unit—over a longer period of time.

Said measures mainly consist of flight restrictions, borders closing, shutting down cafes and restaurants, closing of schools, and self-isolation at first and restriction of movement afterwards, with a total lockdown being the last resort, which has unfortunately been taken in several cases, like that of Lombardy and Spain. The United Kingdom and the Netherlands followed a different approach at first, despite the Imperial College's Response Team's reports led by Prof Ferguson [7-9], with many claiming that they were aiming at herd immunity, which also posed several ethical concerns. Even these two countries,

however, resorted to some measures and restrictions at the end [10,11].

As Gunther Eysenbach, who first proposed the concept of infodemiology (ie, information epidemiology [12-14]), suggested during the SARS pandemic, the use of population health technologies such as the internet can assist with the detection of diseases during an early stage [15]. Given the serious impact of the novel coronavirus and toward the direction of using new methods and approaches for the nowcasting and forecasting of this pandemic, in this paper, Google Trends data are used to explore the relationship between online interest in COVID-19 and cases and deaths in severely affected European countries (ie, Italy, Spain, France, Germany, and the United Kingdom). During these times, infodemiology metrics, especially if combined with traditional data, can be an integral part of the surveillance of the virus at the regional level.

Methods

Data from Google Trends [16] are normalized and retrieved online in .csv format. Note that data may slightly vary based on the time of retrieval. Time series from Google Trends for various time intervals from January to March 2020 on the Topic (Virus) of "Coronavirus" are used, combined with official data on COVID-19 cases and deaths retrieved from Worldometer [5]. The aim is to track the spread of the disease in the European countries that have been affected the most (ie, Italy, Spain, France, Germany, and the United Kingdom). Regional analysis is performed in Italy (data from the Ministry of Health [17]), and the Pearson correlation coefficients between COVID-19 cases and deaths and Google Trends time series are calculated. The Topic of "Coronavirus" was selected instead of the "COVID-19" search term, as the latter was not widely used up to the point of the analysis.

For the general worldwide interest and correlation analysis, the period was set from January 22 to March 17, 2020, while for the rest of the European countries it was set from February 15 to March 17. For the detailed European countries' correlation analysis, case and death data from March 2 to 17 were used. A new data set was retrieved for each time frame, which matched the official COVID-19 case data. The default "All categories" and "Web search" were selected. Note that each country, region, and county were examined individually, and no comparisons between countries in COVID-19 data or Google data were made. The heat maps are based on absolute numbers for COVID-19 cases and deaths, and not according to the respective population. The methodology was designed based on the Google Trends methodology framework in infodemiology and infoveillance [18].

Results

Table 1 consists of the Pearson correlation coefficients (*r*) between Google Trends data and the respective categories of total (cumulative) and daily cases and deaths (where applicable), worldwide (January 22 to March 17) and in the five most affected European countries (February 15 to March 17) (ie, Italy, Spain, France, Germany, and the United Kingdom). Note that for the total worldwide cases excluding China, the Pearson correlation coefficient (*r*) is .9430, with *P*<.001.

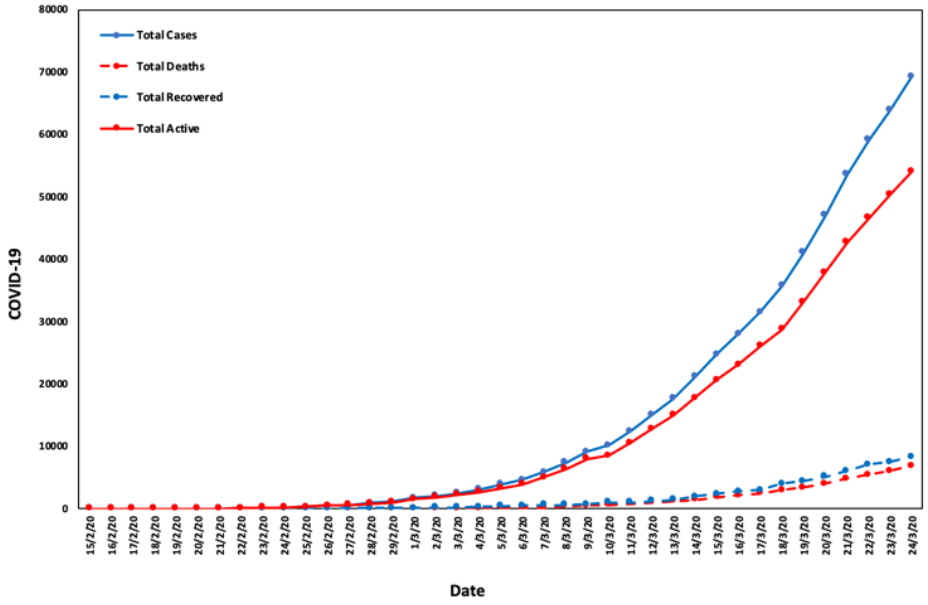
Table 1. Pearson correlation coefficients (*r*) between Google Trends and COVID-19 data.

Variables	Worldwide		Italy		Spain		France		Germany		United Kingdom	
	r	*P* value	*r*	*P* value	*r*	*P* value	*r*	*P* value	*r*	*P* value	*r*	*P* value
Total cases	0.8293	<.001	0.3301	.07	0.7363	<.001	0.8709	<.001	0.674	<.001	0.8956	<.001
Total deaths	0.8917	<.001	0.2837	.12	N/A[a]	N/A	0.8542	<.001	N/A	N/A	N/A	N/A
Daily new cases	0.7575	<.001	0.3931	.03	0.8342	<.001	N/A	N/A	N/A	N/A	0.8479	<.001
Daily new deaths	0.8536	<.001	0.3474	.05	N/A	N/A	0.8554	<.001	N/A	N/A	N/A	N/A

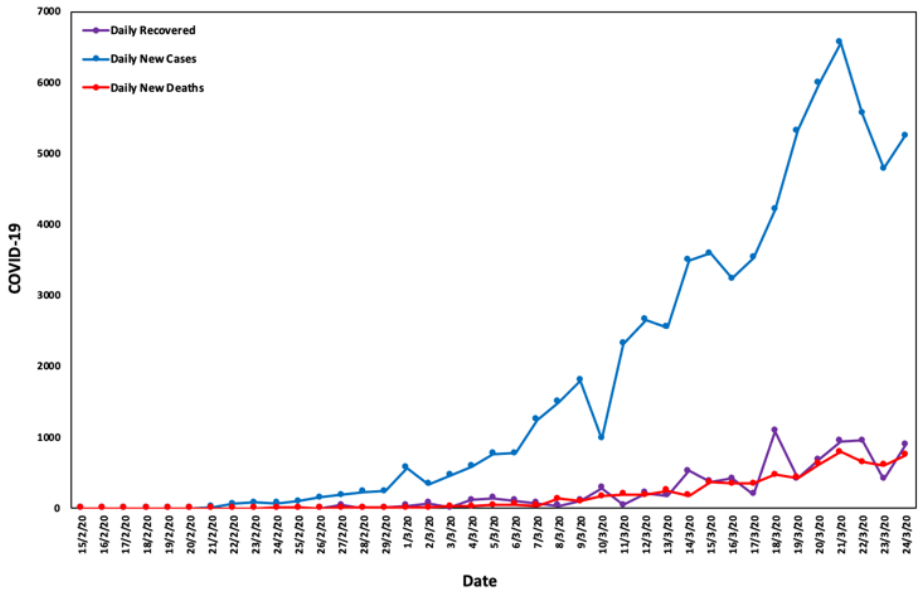
[a]N/A: not applicable.

Based on the results, high statistical significance was observed for the correlations between Google and COVID-19 data for all countries and all applicable categories, apart from Italy, where Google data and COVID-19 total deaths were not correlated. In Italy, total cases and daily deaths were statistically significant but with lower significance, which is not in line with the results for the rest of the countries. The latter could be due to Italy's current special circumstances; it is the first European country to experience such severe consequences from COVID-19 and is further along the line compared with the rest of the countries. Figure 3 depicts the cumulative and daily cases, recoveries, and deaths from February 15 to March 24 in Italy.

Figure 3. (a) Cumulative and (b) daily cases, recoveries, and deaths (Italy; February 15-March 24).



(a)



(b)

Thus, what is essential at this point is to examine if there had been periods for which COVID-19 cases and deaths in Italy correlated with Google query data. The following time frames were selected: March 2-9, March 2-10, March 2-11, March 2-12, March 2-13, March 2-13, March 2-14, March 2-15, March 2-16, and March 2-17.

Table 2 consists of the correlations between Google Trends data and cases, deaths, daily new cases, and daily new deaths in Italy for the aforementioned time frames. Tables 3-4 consist of the individual regions' correlations between COVID-19 cases and Google data.

Table 2. Pearson correlation coefficients (*r*) between COVID-19 cases and deaths and Google Trends data in Italy.

Time frames	Cases		Deaths		Daily Cases[a]		Daily Deaths[a]	
	r	*P* value	*r*	*P* value	*r*	*P* value	*r*	*P* value
March 2-9	0.9484	<.001	0.9336	<.001	0.9574	<.001	0.8097	.02
March 2-10	0.9157	<.001	0.8593	.003	0.8796	.002	0.7901	.01
March 2-11	0.8951	<.001	0.8261	.003	0.8473	.002	0.7979	.006
March 2-12	0.7942	.004	0.7279	.01	0.7644	.006	0.7792	.005
March 2-13	0.6357	.03	0.5605	.06	0.6768	.02	0.6401	.03
March 2-14	0.5067	.08	0.4537	.12	0.5394	.06	0.6223	.02
March 2-15	0.4417	.11	0.3949	.16	0.4828	.08	0.5071	.06
March 2-16	0.2944	.29	0.2410	.39	0.4065	.13	0.3678	.18
March 2-17	0.1588	.56	0.1036	.70	0.0388	.89	0.2624	.33

[a]Refers to daily new cases and deaths.

Table 3. Pearson correlation coefficients (*r*) between COVID-19 cases and Google Trends data in the 20 Italian regions for March 2-9; March 2-10; March 2-11; March 2-12.

Region	March 2-9		March 2-10		March 2-11		March 2-12	
	r	*P* value	*r*	*P* value	*r*	*P* value	*r*	*P* value
Lombardia	0.8987	.002	0.8876	.001	0.8625	.001	0.7502	.008
Emilia-Romagna	0.9017	.002	0.8839	.002	0.8798	<.001	0.8292	.002
Veneto	0.9117	.002	0.9230	<.001	0.9139	<.001	0.7960	.003
Piedmont	0.9494	<.001	0.8690	.002	0.8545	.002	0.7537	.007
Marche	0.8770	.005	0.8301	.006	0.8384	.002	0.7551	.007
Liguria	0.8739	.002	0.8451	.004	0.8042	.005	0.6810	.02
Campania	0.9506	<.001	0.9289	<.001	0.9175	<.001	0.8616	<.001
Toscana	0.9073	.002	0.8279	.006	0.8274	.003	0.7529	.007
Lazio	0.9458	<.001	0.9243	<.001	0.8883	<.001	0.7712	.005
Friuli	0.9310	<.001	0.9407	<.001	0.9284	<.001	0.8493	<.001
Trento	0.8722	.005	0.7934	.01	0.7364	.02	0.6978	.02
Apulia	0.9092	.002	0.9005	<.001	0.8573	.002	0.7894	.004
Sicily	0.9725	<.001	0.9691	<.001	0.9510	<.001	0.8604	<.001
Abruzzo	0.8720	.005	0.8523	.004	0.8685	.001	0.6261	.04
Umbria	0.8775	.004	0.8636	.003	0.8158	.004	0.7104	.01
Aosta	0.8704	.005	0.8179	.007	0.7870	.007	0.5679	.07
Sardinia	0.9170	.001	0.9047	<.001	0.7676	.009	0.7268	.01
Calabria	0.9054	.002	0.9004	<.001	0.8413	.002	0.7197	.01
Molise	0.7101	.048	0.7382	.02	0.7160	.02	0.6764	.02
Basilicata	0.8881	.003	0.7884	.01	0.8306	.003	0.8278	.002

Table 4. Pearson correlation coefficients (*r*) between COVID-19 cases and Google Trends data in the 20 Italian regions for March 2-13; March 2-14; March 2-15; March 2-16; March 2-17.

Region	March 2-13		March 2-14		March 2-15		March 2-16		March 2-17	
	r	*P* value	*r*	*P* value	*r*	*P* value	*r*	*P* value	*r*	*P* value
Lombardia	0.5864	.045	0.4216	.15	0.348	.22	0.1676	.55	0.0693	.80
Emilia-Romagna	0.6471	.02	0.5013	.08	0.442	.11	0.2773	.32	0.1406	.60
Veneto	0.6557	.02	0.4931	.09	0.4900	.08	0.3542	.20	0.2286	.39
Piedmont	0.5599	.06	0.3969	.18	0.3181	.27	0.1341	.63	0.0329	.90
Marche	0.4817	.11	0.2615	.39	0.1687	.56	–0.0869	.76	–0.1932	.47
Liguria	0.5682	.05	0.4111	.16	0.3145	.27	0.2237	.42	0.1166	.67
Campania	0.7073	.01	0.5285	.06	0.4668	.09	0.2611	.35	0.0789	.77
Toscana	0.5822	.047	0.4447	.13	0.396	.16	0.2228	.43	0.115	.67
Lazio	0.4665	.13	0.3157	.29	0.27	.35	0.0683	.81	–0.0746	.78
Friuli	0.6211	.03	0.4791	.097	0.4274	.13	0.2872	.30	0.1774	.51
Trento	0.4813	.11	0.3592	.23	0.2652	.36	0.0553	.85	–0.0388	.89
Apulia	0.6426	.02	0.4421	.13	0.3555	.21	0.2495	.37	0.0419	.88
Sicily	0.7720	.003	0.7055	.007	0.6291	.02	0.5398	.04	0.4332	.09
Abruzzo	0.5535	.06	0.4495	.12	0.4362	.12	0.2808	.31	0.1717	.53
Umbria	0.6088	.04	0.4299	.14	0.3501	.21	0.2063	.46	0.0649	.81
Aosta	0.5123	.09	0.3779	.20	0.2761	.34	0.1942	.49	0.114	.67
Sardinia	0.6188	.03	0.5551	.049	0.5808	.03	0.4049	.13	0.3125	.24
Calabria	0.6272	.03	0.5594	.047	0.5310	.05	0.4234	.12	0.2467	.36
Molise	0.7222	.008	0.4785	.098	0.4498	.12	0.3883	.15	0.232	.39
Basilicata	0.7522	.005	0.7239	.005	0.6253	.02	0.5945	.02	0.4291	.097

As is evident, the strength of the correlation decreases as the time frame includes days when the disease was already widespread, both for cumulative and daily cases and deaths. This is due to the critical point during the spreading of the disease, after which the online interest in the virus starts declining. This is apparent especially for the cumulative cases and deaths, where one function is monotonous (increasing), while the other starts exhibiting a decrease after reaching a peak. Thus, said critical point should be identified in countries and regions with fewer cases to examine the possibility of using Google Trends data to nowcast the spread of COVID-19.

Figures 4 and 5 depict the changes in the Pearson correlation coefficients (*r*) between Google Trends data and COVID-19 cases and deaths for the aforementioned time periods in Italy and Lombardy, respectively. Graphs for the respective changes in the Pearson correlation coefficients for the 20 Italian regions can be found in Multimedia Appendix 1.

Based on these results, it is suggested that regional nowcasting of COVID-19 is possible by simply monitoring Google Trends data until that critical point. This is of high significance if it is applied locally, as it could indicate the regions that will exhibit an increase in COVID-19 cases, thus increasing the preparedness of the health care systems, while, most importantly, taking the needed measures to minimize disease spreading.

In Europe, the countries experiencing the highest case and death counts (after Italy) are Spain, France, Germany, and the United Kingdom, with Spain being in an extremely difficult position with plane traffic being restricted and the army regulating local and regional movement. Thus, for the same time frames as for the Italian regions, the correlations between COVID-19 cases and deaths (where applicable) and the online interest in COVID-19 were calculated. Figures 6-8 depict the changes in the Pearson correlation coefficients for the selected time frames for Spain, Germany, and France.

Figure 4. Changes in the Pearson correlation coefficients (*r*) for Italy.

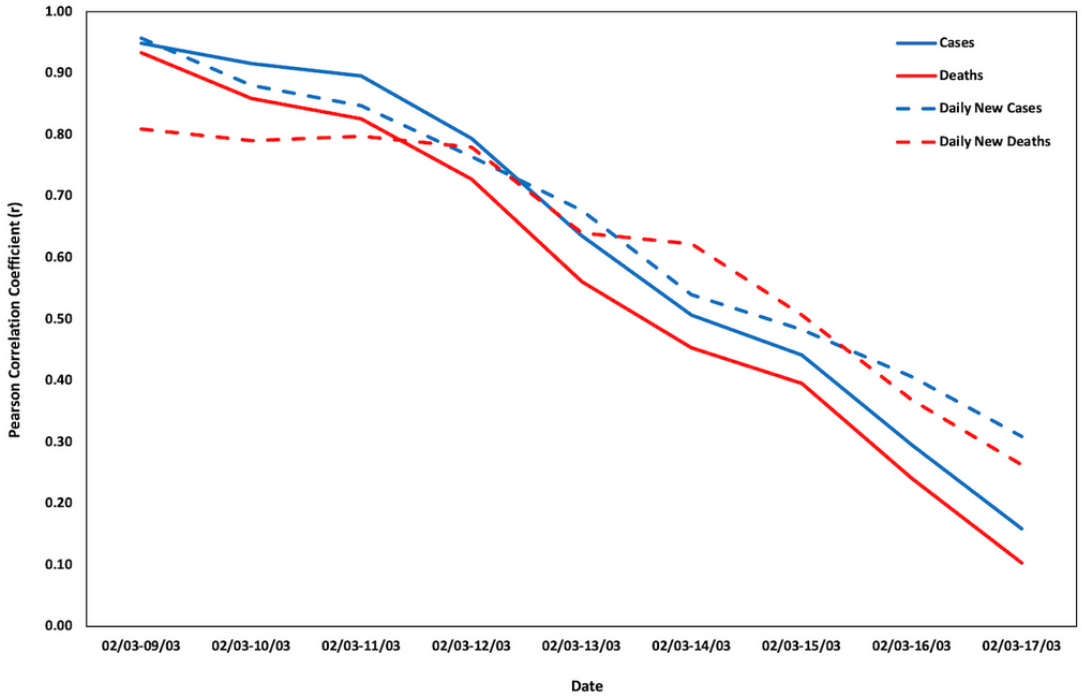


Figure 5. Changes in the Pearson correlation coefficients (*r*) for Lombardy.

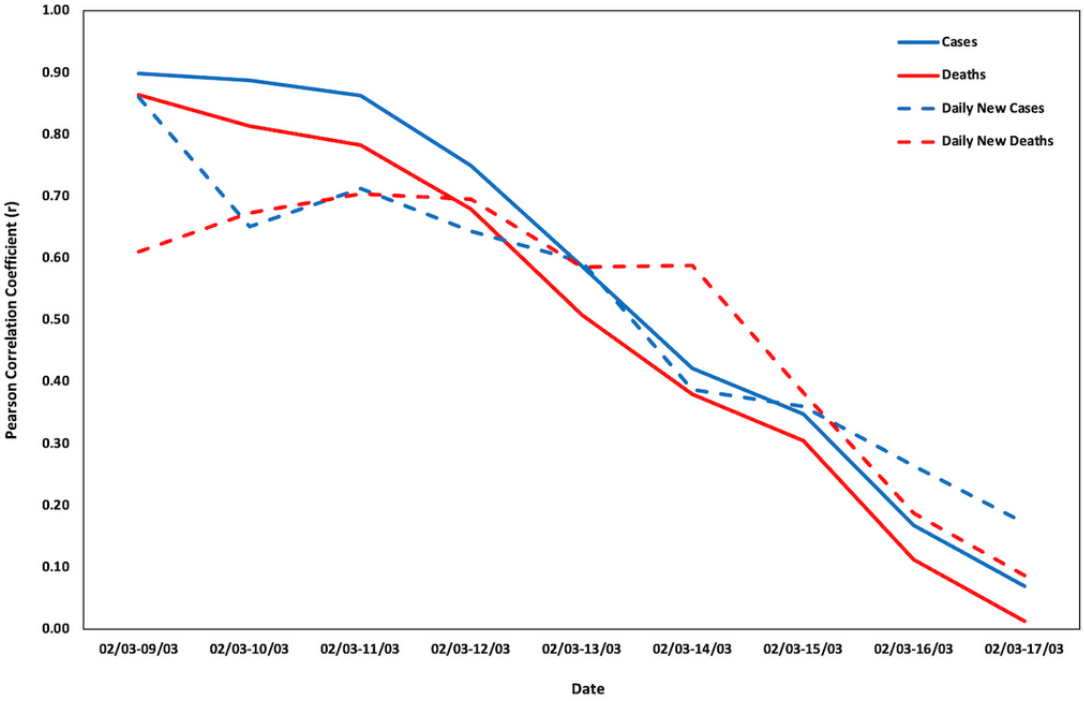


Figure 6. Changes in the Pearson correlation coefficients (*r*) for Spain.

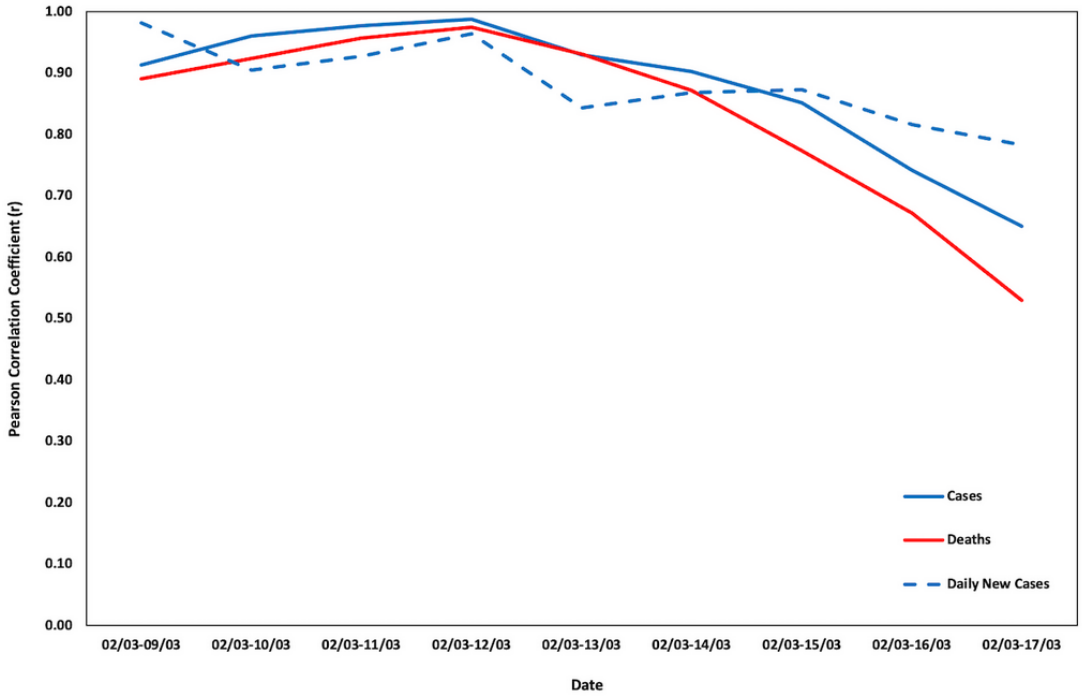


Figure 7. Changes in the Pearson correlation coefficients (*r*) for Germany.

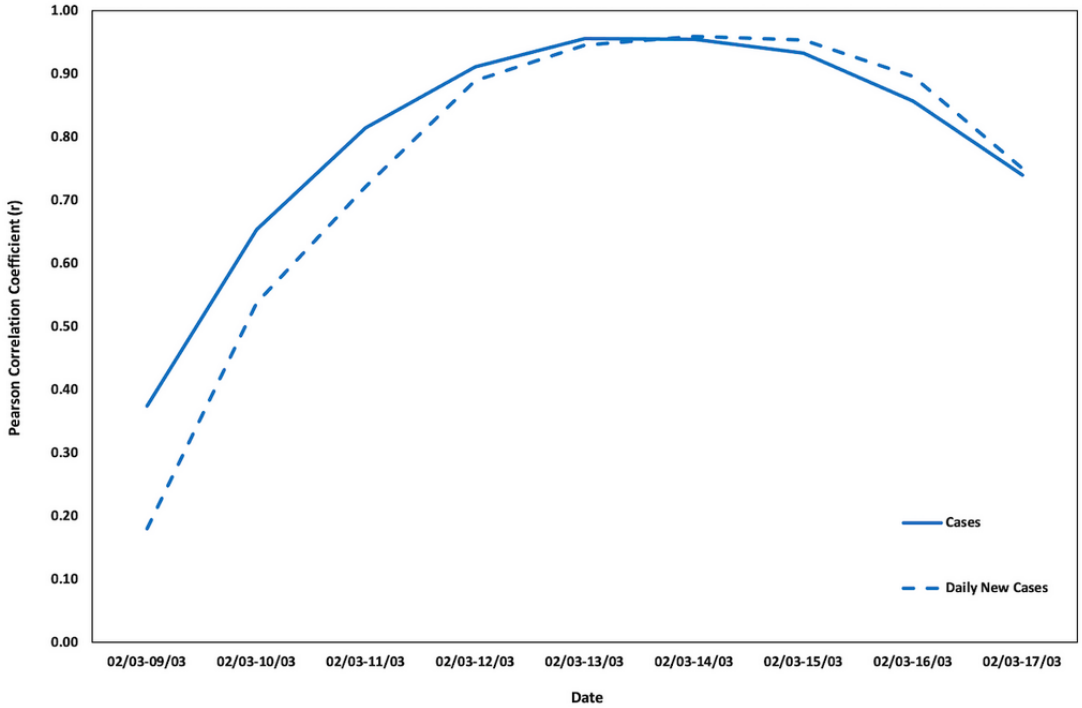
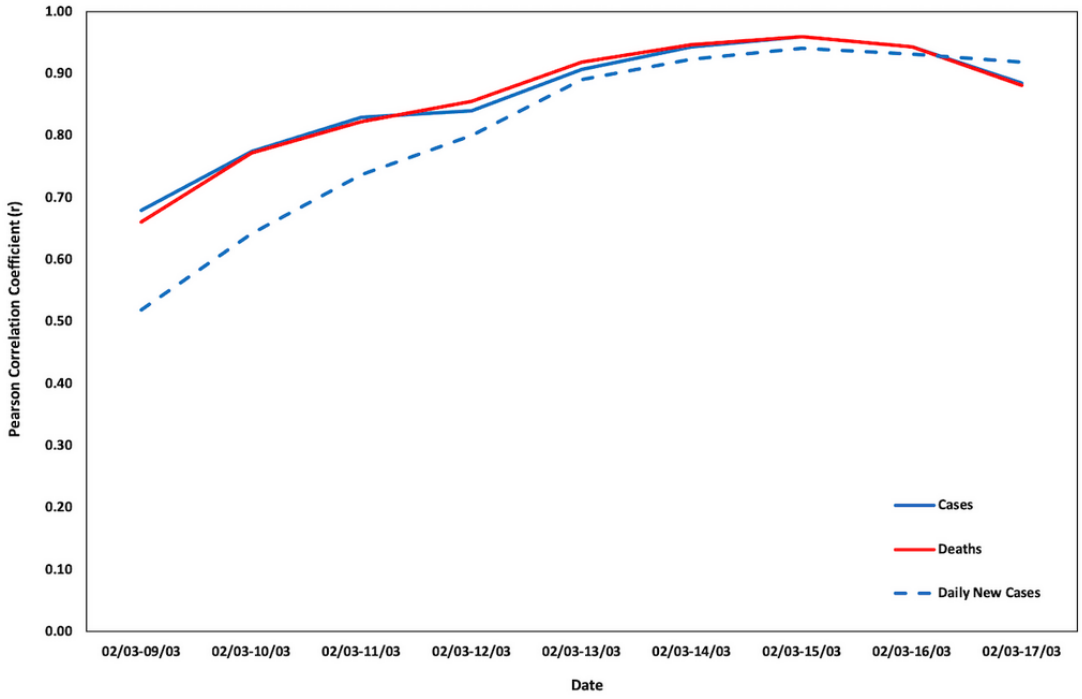


Figure 8. Changes in the Pearson correlation coefficients (*r*) for France.



For Spain, which is closely following Italy in COVID-19 cases and deaths, the Pearson correlation coefficient starts declining after March 13, 2020, which is when Spain's death toll reached 100. In France, the curve still has an increasing trend (150 total deaths as of March 16, 2020), while Germany's curve has started declining since March 15, which is when the country's casualties from COVID-19 passed 10.

Next, the most affected European country (ie, the United Kingdom with more than 10,000 cases) was selected to elaborate on the relationship between COVID-19 cases and deaths and the online interest in the topic. The United Kingdom followed a different approach than most European countries, by not taking preventive measures at an early stage. Figure 9 depicts the changes in the Pearson correlation coefficients for the same time frames selected previously. As is evident, the United Kingdom is still exhibiting high and statistically significant correlations (Table 5).

Figure 9. Changes in the Pearson correlation coefficients (*r*) for the United Kingdom.

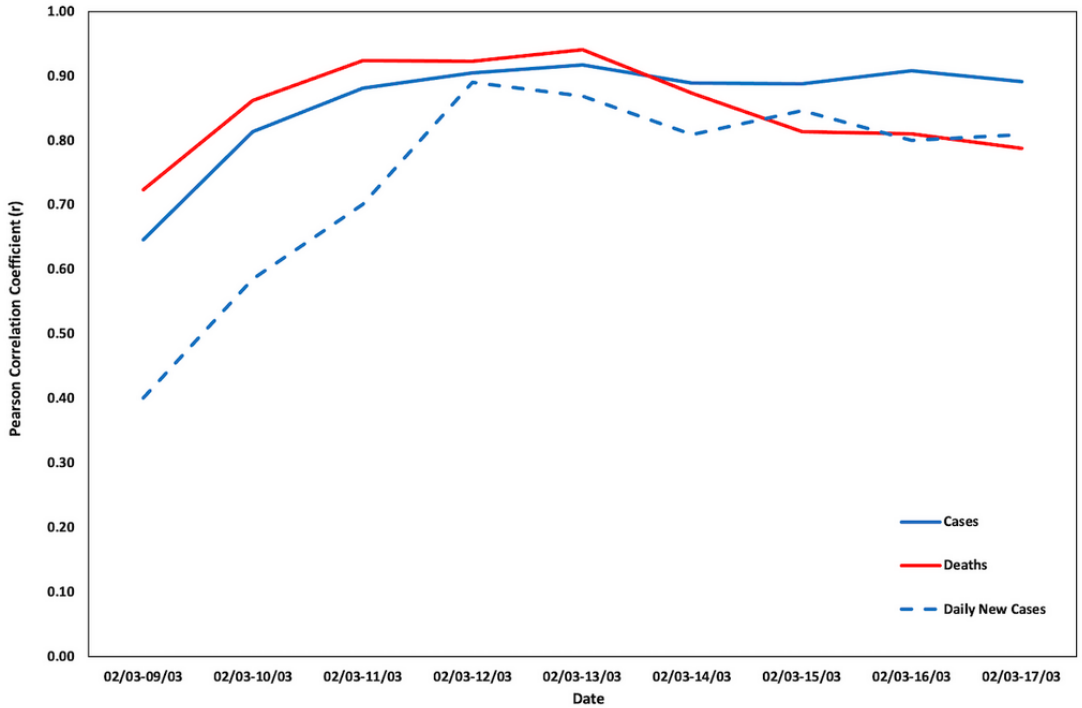


Table 5. Pearson correlation coefficients (*r*) between COVID-19 cases and deaths and Google Trends data for the United Kingdom.

Time Frames	Cases		Deaths		Daily Cases	
	r	*P* value	*r*	*P* value	*r*	*P* value
March 2-9	0.6470	.08	0.7241	.04	0.4008	.33
March 2-10	0.8144	.008	0.8629	.003	0.5863	.097
March 2-11	0.8811	<.001	0.9244	<.001	0.7021	.02
March 2-12	0.9053	<.001	0.9229	<.001	0.8907	<.001
March 2-13	0.9177	<.001	0.9408	<.001	0.8689	<.001
March 2-14	0.8896	<.001	0.8742	<.001	0.8091	<.001
March 2-15	0.8878	<.001	0.8145	<.001	0.8470	<.001
March 2-16	0.9083	<.001	0.8110	<.001	0.8010	<.001
March 2-17	0.8920	<.001	0.7878	<.001	0.8100	<.001

The relationship between COVID-19 cases and deaths shows an increasing trend over the examined period and stays high afterwards. Note that the United Kingdom had zero deaths March 2-4, 2020. The decrease is also evident in Table 5, which consists of the Pearson correlation coefficients and their significance, the latter also exhibiting increased rates as time moves forward, contrary to Italy, Spain, and all Italian regions.

Therefore, it is evident that a correlation between COVID-19 and Google Trends data exists, but the critical point, after which the online interest starts declining, should be identified in each individual case to proceed with regional nowcasting. Toward this direction, the data period should be shortened and applied to regions that have not yet been as severely affected. Google Trends provides a detailed regional break down for most countries, as well as real time and 1-hour interval data over the past week; this gives the opportunity of nowcasting users' search patterns and online behavior toward the disease.

Discussion

Principal Findings

Infodemiology metrics and approaches are an integral part of health informatics, with the most popular sources being Twitter and Google [19,20], which have been successfully employed in the past to track and forecast outbreaks and epidemics (eg, Middle East respiratory syndrome [21], measles [22,23], Ebola [24,25], the swine flu [26], and the Zika epidemic [27,28]).

However, the case of the new coronavirus is somewhat different both in terms of the qualitative and quantitative approach than the previously examined epidemics. COVID-19 has been the subject of several controversial discussions. Since China's first death report on January 11, 2020 [2], there have been several controversies regarding how China has handled the epidemic. There are ongoing debates as to whether there had been an attempt to hide the beginning of the outbreak, which became public by whistleblower Dr Li Wenliang who was reported dead as of February 7 due to COVID-19 complications [29]. There has been information about reporters being expelled from China as brought forward by New York Times reporter Amy Qin [30]. Most importantly though, there have been doubts about the accuracy of the data and results that the Chinese authorities and scientists have provided, with a much discussed incident being the announcement that "*Preliminary investigations conducted by the Chinese authorities have found no clear evidence of human-to-human transmission of the novel #coronavirus (2019-nCoV) identified in #Wuhan, #China*" [31].

However, the case of Italy, which is the country with the highest death toll and should perhaps be treated as the first case of what to expect from the virus spread, shows that the epidemic is far more serious than what the officials originally suggested, with a record daily death toll of 919 reported on March 27 [32] and total deaths slightly less than 10,000. Based on Italy's data, many European countries acted fast in imposing measures for slowing down the spread of the disease, and the next 2-3 weeks could exhibit nonexponential curves in terms of daily casualties.

Toward the direction of finding new methods for nowcasting COVID-19 to increase the preparedness of health care systems, this study suggests that Google Trends data strongly correlates with COVID-19 cases and deaths worldwide and in the examined countries. Most importantly though, there is a critical point, after which the relationship's strength (in almost all cases) monotonously decreases, even if the correlation remains statistically significant, with Italy having the sharpest downward curve.

Limitations

This study has limitations. First, since the pandemic not only is ongoing but has not reached its peak yet, the data are limited; thus, the correlations are based on fewer observations, and the results are only preliminary and subject to change as we move forward. Second, only a few countries provided, at the time of writing, sufficient data for analysis or a regional break down of the cases and deaths. Third, only the interest in the "Coronavirus (Virus)" Topic was explored, but future reports should also elaborate on more complicated search patterns, especially using the official name of the disease (ie, COVID-19) once it is used by a significant part of the population. Fourth, there are significant changes in cases, deaths, and rates even between 2 consecutive days in many regions and countries; even at the time of writing, the data can significantly vary from those at the time of retrieval.

Conclusions

In line with previous studies that have indicated that Google Trends data can assist with the tracking and nowcasting of epidemics and outbreaks, the results of this paper show that online search traffic data are highly correlated with COVID-19 cases and deaths in the examined countries and regions. Furthermore, a critical point, up to which regions not severely affected exhibit the strongest relationship between Google and COVID-19 data, was identified. This suggests that focus should shift towards these regions to make full use of what real time data assessment can offer. The latter is essential for increasing the preparedness and responsiveness of local health institutions, which is the most important aspect in handling the current pandemic.

As of March 27, the center of the COVID-19 pandemic is the United States, with New York being the most affected, and it is imperative to perform similar analyses regionally, at state, metro, and city levels. Data from the disease spread and casualties in Europe will provide a better picture as to the characteristics of the virus as well as detailed data—both traditional and infodemiological—to estimate nowcasting models.

Despite the limited data availability at this stage of the pandemic, it is essential that all results are shared and rapid publications on the topic of infodemiology are accessible. Infodemiology results from various sources such as Google, Twitter, Facebook, or other social media are valuable variables in epidemiology. It is crucial to use such preliminary findings to build novel approaches that make use of real time data for the tracking and nowcasting of COVID-19.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Changes in the Pearson correlation coefficients (r) for the 20 Italian regions.
[PDF File (Adobe PDF File), 1310 KB-Multimedia Appendix 1]

References

1. World Health Organization. Rolling updates on coronavirus disease (COVID-19) URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen [accessed 2020-03-27]
2. Secon H, Woodward A, Mosher D. Business Insider. A comprehensive timeline of the new coronavirus pandemic, from China's first COVID-19 case to the present URL: https://tinyurl.com/r6johyw [accessed 2020-03-23]
3. Worldometers. Coronavirus. Italy URL: https://www.worldometers.info/coronavirus/country/italy/ [accessed 2020-03-27]
4. World Health Organization. 2020 Mar 12. WHO announces COVID-19 outbreak a pandemic URL: http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic [accessed 2020-03-27]
5. Worldometers. COVID-19 coronavirus pandemic URL: https://www.worldometers.info/coronavirus/ [accessed 2020-03-19]
6. Specktor B. LiveScience. 2020 Mar. Coronavirus: what is 'flattening the curve,' and will it work? URL: https://www.livescience.com/coronavirus-flatten-the-curve.html [accessed 2020-04-07]
7. van Elsland SL, O'Hare R. Imperial College London. 2020 Mar 17. COVID-19: Imperial researchers model likely impact of public health measures URL: https://www.imperial.ac.uk/news/196234/covid-19-imperial-researchers-model-likely-impact/ [accessed 2020-03-27]
8. Walker PGT, Whittaker C, Watson O, Baguelin M, Ainslie KEC, Bhatia S. Imperial College London. The global impact of COVID-19 and strategies for mitigation and suppression URL: https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-Global-Impact-26-03-2020.pdf [accessed 2020-03-27]
9. Ferguson N, Laydon D, Nedjati-Gilani G, Imai N, Ainslie K, Baguelin M, et al. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID- 19 mortality and healthcare demand. Imperial College London 2020 Mar 16 [FREE Full text] [doi: 10.25561/77482]
10. BBC News. 2020 Mar 26. Coronavirus: UK before and after 'lockdown' URL: https://www.bbc.com/news/uk-52051468 [accessed 2020-03-27]
11. Reuters. 2020 Mar 23. Dutch PM Rutte: ban on public gatherings is "intelligent lockdown" URL: https://tinyurl.com/ubx65qg [accessed 2020-03-27]
12. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009 Mar 27;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
13. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med 2011 May;40(5 Suppl 2):S154-S158. [doi: 10.1016/j.amepre.2011.02.006] [Medline: 21521589]
14. Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA Annu Symp Proc 2006:244-248 [FREE Full text] [Medline: 17238340]

15. Eysenbach G. SARS and population health technology. J Med Internet Res 2003;5(2):e14 [FREE Full text] [doi: 10.2196/jmir.5.2.e14] [Medline: 12857670]
16. Google Trends Explore. URL: https://trends.google.com/trends/explore; [accessed 2020-03-26]
17. Ministero della Salute. Nuovo Coconavirus URL: http://www.salute.gov.it/nuovocoronavirus; [accessed 2020-03-19]
18. Mavragani A, Ochoa G. Google Trends in infodemiology and infoveillance: methodology framework. JMIR Public Health Surveill 2019 May 29;5(2):e13439 [FREE Full text] [doi: 10.2196/13439] [Medline: 31144671]
19. Mavragani A. Infodemiology and infoveillance: a scoping review [accepted manuscript]. J Med Internet Res 2020.
20. Mavragani A, Ochoa G, Tsagarakis KP. Assessing the methods, tools, and statistical approaches in Google Trends research: systematic review. J Med Internet Res 2018 Nov 06;20(11):e270 [FREE Full text] [doi: 10.2196/jmir.9366] [Medline: 30401664]
21. Poletto C, Boëlle PY, Colizza V. Risk of MERS importation and onward transmission: a systematic review and analysis of cases reported to WHO. BMC Infect Dis 2016 Aug 25;16(1):448 [FREE Full text] [doi: 10.1186/s12879-016-1787-5] [Medline: 27562369]
22. Mavragani A, Ochoa G. The internet and the anti-vaccine movement: tracking the 2017 EU measles outbreak. BDCC 2018 Jan 16;2(1):2. [doi: 10.3390/bdcc2010002]
23. Du J, Tang L, Xiang Y, Zhi D, Xu J, Song HY, et al. Public perception analysis of Tweets during the 2015 measles outbreak: comparative study using convolutional neural network models. J Med Internet Res 2018 Jul 09;20(7):e236 [FREE Full text] [doi: 10.2196/jmir.9413] [Medline: 29986843]
24. Hossain L, Kam D, Kong F, Wigand RT, Bossomaier T. Social media in Ebola outbreak. Epidemiol Infect 2016 Mar 04;144(10):2136-2143. [doi: 10.1017/s095026881600039x]
25. van Lent LG, Sungur H, Kunneman FA, van de Velde B, Das E. Too far to care? Measuring public attention and fear for Ebola using Twitter. J Med Internet Res 2017 Jun 13;19(6):e193 [FREE Full text] [doi: 10.2196/jmir.7219] [Medline: 28611015]
26. Bentley RA, Ormerod P. Social versus independent interest in 'bird flu' and 'swine flu'. PLoS Curr 2009 Sep 03;1:RRN1036 [FREE Full text] [doi: 10.1371/currents.rrn1036] [Medline: 20025200]
27. Farhadloo M, Winneg K, Chan MS, Hall Jamieson K, Albarracin D. Associations of topics of discussion on Twitter with survey measures of attitudes, knowledge, and behaviors related to Zika: probabilistic study in the United States. JMIR Public Health Surveill 2018 Feb 09;4(1):e16 [FREE Full text] [doi: 10.2196/publichealth.8186] [Medline: 29426815]
28. Chen S, Xu Q, Buchenberger J, Bagavathi A, Fair G, Shaikh S, et al. Dynamics of health agency response and public engagement in public health emergency: a case study of CDC Tweeting patterns during the 2016 Zika epidemic. JMIR Public Health Surveill 2018 Nov 22;4(4):e10827 [FREE Full text] [doi: 10.2196/10827] [Medline: 30467106]
29. BBC News. 2020 Feb 07. Li Wenliang: coronavirus kills Chinese whistleblower doctor URL: https://www.bbc.com/news/world-asia-china-51403795 [accessed 2020-03-27]
30. Withnall A. Independent. 2020 Mar 18. Coronavirus: China expels 13 American reporters amid ‘unparalleled global crisis’ of pandemic URL: https://tinyurl.com/rmq7vkp [accessed 2020-03-27]
31. Twitter. World Health Organization (WHO) (@WHO) URL: https://twitter.com/who/status/1217043229427761152?lang=en
32. The Guardian. Coronavirus live news: record rise in Italy death toll takes total to 9,134, as France extends lockdown by two weeks URL: https://tinyurl.com/w9cw2x7 [accessed 2020-03-27]

Abbreviations

COVID-19: coronavirus disease
SARS: severe acute respiratory syndrome

Edited by M Focsa, G Eysenbach, T Sanchez; submitted 28.03.20; peer-reviewed by E Da Silva, V Gianfredi; accepted 02.04.20; published 20.04.20

Please cite as:
Mavragani A
Tracking COVID-19 in Europe: Infodemiology Approach
JMIR Public Health Surveill 2020;6(2):e18941
URL: http://publichealth.jmir.org/2020/2/e18941/
doi: 10.2196/18941
PMID:

scientific reports

Check for updates

OPEN

COVID-19 predictability in the United States using Google Trends time series

Amaryllis Mavragani[1]✉ & Konstantinos Gkillas[2]

During the unprecedented situation that all countries around the globe are facing due to the Coronavirus disease 2019 (COVID-19) pandemic, which has also had severe socioeconomic consequences, it is imperative to explore novel approaches to monitoring and forecasting regional outbreaks as they happen or even before they do so. To that end, in this paper, the role of Google query data in the predictability of COVID-19 in the United States at both national and state level is presented. As a preliminary investigation, Pearson and Kendall rank correlations are examined to explore the relationship between Google Trends data and COVID-19 data on cases and deaths. Next, a COVID-19 predictability analysis is performed, with the employed model being a quantile regression that is bias corrected via bootstrap simulation, i.e., a robust regression analysis that is the appropriate statistical approach to taking against the presence of outliers in the sample while also mitigating small sample estimation bias. The results indicate that there are statistically significant correlations between Google Trends and COVID-19 data, while the estimated models exhibit strong COVID-19 predictability. In line with previous work that has suggested that online real-time data are valuable in the monitoring and forecasting of epidemics and outbreaks, it is evident that such infodemiology approaches can assist public health policy makers in addressing the most crucial issues: flattening the curve, allocating health resources, and increasing the effectiveness and preparedness of their respective health care systems.

In December 2019, a novel coronavirus of unknown source was identified in a cluster of patients in the city of Wuhan, Hubei, China[1]. The outbreak first came to international attention after the World Health Organization (WHO) reports said that there was a cluster of pneumonia cases on Twitter on January 4th[2], followed by the release of an official report on January 5th[3]. China reported its first COVID-19-related death on January 11th, while on January 13th, the first case outside China was identified[4]. On January 14th, the World Health Organization (WHO) tweeted that Chinese preliminary investigations reported that no human-to-human transmission had been identified[5]. However, the virus quickly spread to other Chinese regions and neighboring countries, while Wuhan, identified as the epicenter of the outbreak, was cut off by authorities on January 23rd, 2020[6]. On January 30th, the WHO declared the epidemic to be a public health emergency[1], and the disease caused by the virus received its official name, that is, COVID-19, on February 11th[7].

The first serious COVID-19 outbreak in Europe was identified in northern Italy during February, with the country recording its first death on February 21st[8]. The novel coronavirus was transmitted to all parts of Europe within the next few weeks, and as a result, the WHO declared COVID-19 to be a pandemic on March 11th, 2020. As of 16:48 GMT on April 18th, 2020[9], there were 2,287,369 confirmed cases worldwide, with 157,468 confirmed deaths and 585,838 recovered patients. The most affected countries with more than 100 k cases (in absolute numbers, not divided by population) were the US, with 715,105 confirmed cases and 37,889 deaths; Spain, with 191,726 confirmed cases and 20,043 deaths; Italy, with 175,925 confirmed cases and 23,227 deaths; France, with 147,969 confirmed cases and 18,681 deaths; Germany, with 142,614 confirmed cases and 4405 deaths; and the UK, with 114,217 confirmed cases and 15,464 deaths. The worldwide geographical distribution of COVID-19 cases and deaths by country is depicted in Fig. 1.

As shown, Europe has been severely affected by COVID-19. However, the spread of the disease now indicates that the center of the epidemic has moved to the US, with the state of New York counting more than 240 k cases and 17 k deaths. Figure 2 shows the distribution of COVID-19 cases and deaths in the United States by state as of April 18th, 2020[10].

[1]Department of Computing Science and Mathematics, Faculty of Natural Sciences, University of Stirling, Stirling FK9 4LA, Scotland, UK. [2]Department of Management Science and Technology, University of Patras, Patras, Greece. ✉email: amaryllis.mavragani1@stir.ac.uk

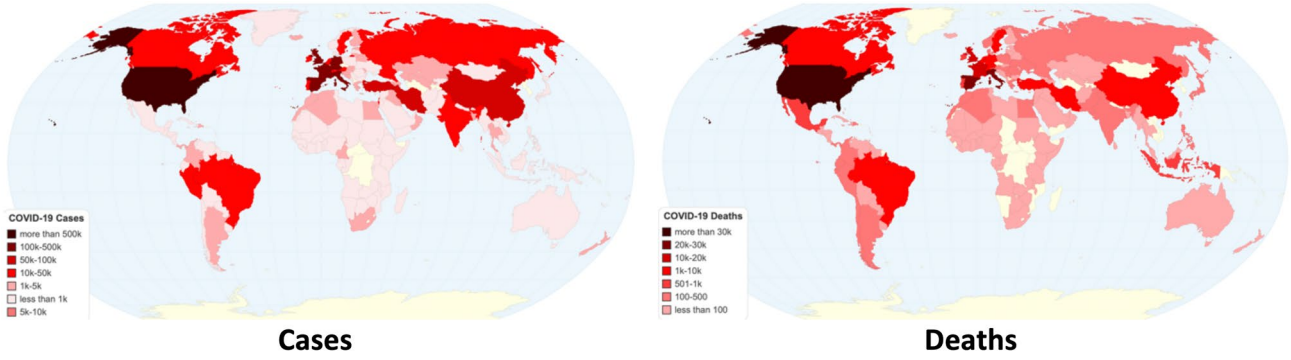


Figure 1. Geographical distribution of worldwide COVID-19 cases and deaths as of April 18th (Chartsbin[43]).

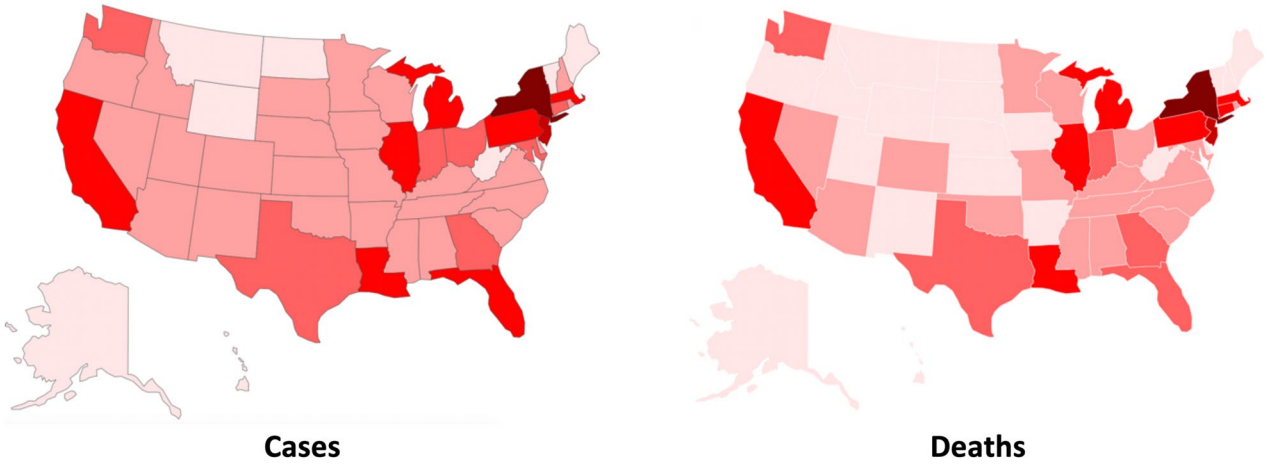


Figure 2. Geographical distribution of COVID-19 cases and deaths in the US as of April 18th (Pixelmap[42]).

To find new methods and approaches for disease surveillance, it is crucial to take advantage of real-time internet data. Infodemiology, i.e., information epidemiology, is a concept that was introduced by Gunther Eysenbach[11,12]. In the field of infodemiology, internet sources and data are employed to inform public health and policy[13,14]. These approaches have been suggested to be valuable for the monitoring and forecasting of outbreaks and epidemics[15], such as Ebola[16], Zika[17], MERS[18], influenza[19], and measles[20,21].

During the COVID-19 pandemic, several research studies using web-based data have been published. Google Trends, the most popular infodemiology source along with Twitter, has been widely used in health and medicine for the analysis and forecasting of diseases and epidemics[22]. As of April 20, 2020, seven (7) papers on the topic of monitoring, tracking, and forecasting COVID-19 using Google Trends data had already appeared online in PubMed (advanced search: covid AND google trends)[23] for several regions: Taiwan[24], China[25,26], Europe[27,28], the US[28,29], and Iran [28,30]. Note that for Twitter publications related to the COVID-19 pandemic, eight papers (8) published from March 13, 2020 to April 20, 2020[31–38] are available online (PubMed advanced search: covid AND twitter[23]). Table 1 systematically reports these COVID-19 Google Trends studies, in order of the reported publication date.

In this paper, Google Trends data on the topic of "Coronavirus (virus)" in the United States are employed at both the national and state levels to explore the relationship between COVID-19 cases and deaths and online interest in the virus. First, a correlation analysis between Google Trends and COVID-19 data is performed; then, the role of Google Trends data in the predictability of COVID-19 is explored. To the best of our knowledge, this paper is the first attempt of this kind performed for the United States.

The rest of the paper is structured as follows. The Methods section details the data collection procedure and the statistical analysis tools and methods. The Results section consists of the correlation analysis and of the forecasting models at both national and state levels. The Discussion section presents the main findings of this work, along with the limitations of this paper and future research suggestions.

Methods

Data from the Google Trends platform are retrieved in .csv[39] and are normalized over the selected period. Google Trends reports the adjustment procedure as follows: "*Search results are normalized to the time and location of a query by the following process: Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity. Otherwise, places with the most search volume would always be ranked*

Authors	Date	Region	Objective	Publisher	Journal
Husnayain et al.[24]	March 12	Taiwan	Analyzing COVID-19 related searches	Elsevier	International Journal of Infectious Diseases
Li et al.[25]	March 25	China	Correlating Internet searches with COVID-19 cases	Eurosurveillance	Eurosurveillance
Mavragani[27]	April 2	Europe	Correlating Google Trends data with COVID-19 cases and deaths	JMIR	JMIR Public Health and Surveillance
Hong et al.[29]	April 7	USA	Relationship between telehealth searches and COVID-19	JMIR	JMIR Public Health and Surveillance
Walker et al.[28]	April 11	USA, Iran, Europe	Exploring of the online activity related to loss of smell	Wiley	International Forum of Allergy and Rhinology
Ayyoubzadeh et al.[30]	April 14	Iran	Prediction of COVID-19 cases	JMIR	JMIR Public Health and Surveillance
Effenberger et al.[26]	April 16	China	Correlation between Google Trends data and COVID-19 cases	Elsevier	International Journal of Infectious Diseases

Table 1. Systematic reporting of publications on COVID-19 using Google Trends as of April 20th, 2020.

March 4th–April 15th	USA; Arizona; California; Florida; Georgia; Illinois; Massachusetts; New Hampshire; New York; North Carolina; Oregon; Texas; Washington; Wisconsin
March 5th–April 15th	Nevada; New Jersey; Tennessee
March 6th–April 15th	Colorado; Indiana; Maryland; Pennsylvania
March 7th–April 15th	Hawaii; Kentucky; Minnesota; Nebraska; Oklahoma; Rhode Island; South Carolina; Utah
March 8th–April 15th	Connecticut; District of Columbia; Kansas; Missouri; Vermont; Virginia
March 9th–April 15th	Iowa; Louisiana; Ohio
March 11th–April 15th	Delaware; Michigan; New Mexico; South Dakota
March 12th–April 15th	Arkansas; Maine; Mississippi; Montana; North Dakota; Wyoming
March 13th–April 15th	Alabama; Alaska
March 14th–April 15th	Idaho
March 18th–April 15th	West Virginia

Table 2. Timeframes for which Google Trends data are retrieved by state.

highest. The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics. Different regions that show the same search interest for a term don't always have the same total search volumes"[40]. The data collection methodology is designed based on the Google Trends Methodology Framework in Infodemiology and Infoveillance[41]. Note that the data may slightly vary based on the time of retrieval.

For keyword selection, the online interest in all commonly used variations is examined, and the variations are compared, i.e., "coronavirus (virus)"; "COVID-19 (search term)"; "SARS-COV-2 (search term)"; "2019-nCoV (search term)"; and "coronavirus (search term)". Only "coronavirus (virus)" and "coronavirus (search term)" yield, as expected, considerably high online interest. Between the two, i.e., the topic (virus) and the search term, "coronavirus (virus)" is selected for further analysis.

Data on the worldwide distribution of COVID-19 cases and deaths are retrieved from Worldometer[9]. Data for the United States analysis of COVID-19 are retrieved from "The COVID Tracking Project", which provides detailed structured data on COVID-19 cases and deaths nationally and at state level[10]. Maps of COVID-19 cases and deaths and online interest are created by the authors using the free online tools Pixelmap[42] and Chartsbin[43], with data from the respective sources[9,10], while graphs, spider web charts, and maps of the correlation coefficients are created by the authors using Microsoft Excel (version 16.39).

As Google Trends data are normalized, the timeframe for which search traffic data are retrieved should exactly match the period for which COVID-19 data are available. Therefore, the timeframes for which analysis is performed are different among states, starting either on March 4th (for most cases) or on the date on which the first confirmed case was identified in each state, as shown in Table 2.

Each variable used in this study is divided by its full-sample standard deviation, estimated or calculated based on the basic formula of the standard deviation of a variable. By doing so, the inherent variability of each variable was moved, and thus, all variables have a standard deviation equal to 1. This equivalence makes it possible to compare the strength of the impact of the explanatory variables used on the dependent variable. The nonparametric[44] unit root test is also applied to reveal whether or not the variables are stationary. The results suggest that both variables can be used directly in the present analysis without further transformation.

The first step in exploring the role of Google Trends in the predictability of COVID-19 is to examine the relationship between Google Trends and the incidence of COVID-19. As Pearson correlation analysis is the benchmark analysis in this kind of approach, the Pearson correlation coefficients (r) between the ratio (COVID-19 deaths)/(COVID-19 cases) and Google Trends data are calculated. In particular, a minimum variance bias-corrected Pearson correlation coefficient[45,46] via a bootstrap simulation is applied to deal with the limited number of observations and, therefore, small sample estimation bias (also see[45,47]). The bias-corrected bootstrap coefficient ρ for the Pearson correlation is given as follows:

$$\overset{\sim b}{\rho} = B^{-1}\sum_{j=1}^{B}\overset{\sim b}{\rho}_j(\rho)$$

where B corresponds to the length of the bootstrap samples; in this case, it is set equal to 999[48]. Note that the terms "COVID-19 deaths" and "COVID-19 cases" refer to the cumulative (total) COVID-19 deaths and cases in the United States and that this terminology is used hereafter unless otherwise stated.

Next, secondary correlation analysis is performed using the Kendall rank correlation, which is a nonparametric test that measures the strength of dependence between two variables. The Kendall rank correlation is distribution free and is considered robust in ratio data. Considering two samples with sample sizes n, the total number of pairings is $\frac{1}{2}n(n-1)$. The following formula is used to calculate the value of the bias-corrected Kendall rank correlation:

$$\overset{\sim b}{\tau} = B^{-1}\sum_{j=1}^{B}\overset{\sim b}{\tau}_j(\tau)$$

where τ is given by $\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$, n_c is the concordant value, and n_d is the discordant value.

Following, a COVID-19 predictability analysis approach based on Google Trends time series for the United States and all US states (plus DC) is performed. The predictability model is a quantile regression, which is considered to be a robust regression analysis against the presence of outliers in the sample; it was introduced by[49]. Building on the study conducted by[46], a quantile regression that is bias corrected via balanced bootstrapping is employed. Such a model is the appropriate statistical approach for mitigating small sample estimation bias and the presence of outliers in the dataset, as it combines the advantages of bootstrap standard errors and the merits of quantile regression. Additional knowledge on quantile regression can be found in the studies conducted by[50] and[51], while recent applications of quantile regression can be found in[52,53]. More recently[54] introduced unconditional quantile regression, while the study by[55] provides further insights into robust estimates of regressions.

Let Y_t, with $t \in T$, be a time series that represents the dependent variable, supposing a bivariate specification. Quantile regression estimates the impact of the explanatory variable X_t, with $t \in T$, on the variable Y_t at different points of the conditional q-quantile, with $q \in (0,1)$, of the conditional distribution. A value of the q-quantile close to zero and a value of the q-quantile close to one represent the left (lower) and right (upper) tails of the conditional distribution, respectively. The conditional quantile function is defined as follows:

$$Q_{Y|X}(q) = \mathrm{X}'\beta_q$$

Given the distribution of Y_t, the estimation of the conditional quantile functions β_q can be obtained by solving the following minimization problem:

$$\beta_q = \arg\min_{\beta\in\mathbb{R}^k} \mathrm{E}\left(\rho_q(\mathrm{Y} - \mathrm{X}\beta)\right)$$

where $\rho_q(y) = y\left(q - 1_{\{y<0\}}\right)$ represents the loss function.

By minimizing the sample analog $\{y_1, \ldots, y_n\}$ that corresponds to a q^{th} quantile sample, the estimator β_q takes the following form:

$$\beta_q = \arg\min_{\beta\in\mathbb{R}^k}\sum_{t=1}^{n}\rho_q\left(Y_t - X_t'\beta\right) = \arg\min_{\beta\in\mathbb{R}^k}\left[q\sum_{Y_t\geq\beta X_t}|Y_t - \beta X_t| + (1-q)\sum_{Y_t<\beta X_t}|Y_t - \beta X_t|\right]$$

where βX_t is an approximation of the conditional q-quantile of the variable Y_t.

In our analysis, Y_t stands for the ratio (COVID-19 deaths)/(COVID-19 cases), X_{t-1} is the respective Google Trends value in lag order, and $t = 1, \ldots, T$, with T being the respective number of observations. A linear trend is used as well.

Finally, the bias-corrected parameter is estimated as follows:

$$\overset{\sim b}{\beta}(q) = \widehat{\beta}(q) - \widehat{bias}\left(\widehat{\beta}(q)\right)$$

where $\widehat{bias}\left(\widehat{\beta}(q)\right)$ is given by $B^{-1}\sum_{j=1}^{B}\widehat{\beta}_j^*(q) - \widehat{\beta}(q)$ and $q \in (0,1)$ denotes the quantile considered and, in this case, is set equal to 0.5 (median). Median regression is considered more robust to outliers than, for example, least squares regression. Finally, it also avoids assumptions about the error parametric distribution[56].

All estimation results reported in this paper were computed in the R programming environment[57]. In particular, we employed the R packages "quantreg" and "boot" to compute the quantile regression estimates and to perform the bootstrapping, respectively. The code is available in a "Supplementary Online Material file".

Results

Figure 3 depicts the worldwide and US online interest in terms of Google queries in the "coronavirus (virus)" topic from January 22nd to April 15th, 2020. It shows that this topic is very popular, especially in Europe and North America. Specifically, interest in the United States is considerably high (above 70) for all US states.

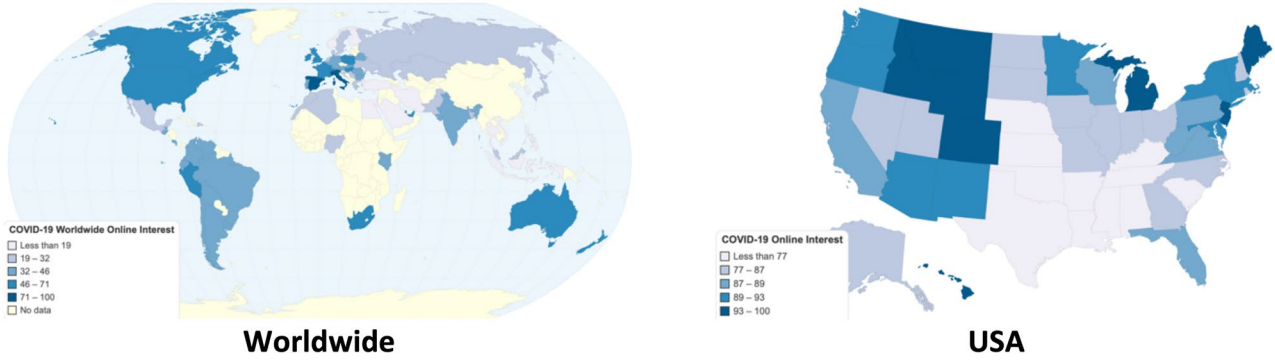


Figure 3. Heat maps of the worldwide and US online interest in "Coronavirus (Virus)" (Chartsbin[43]).

State	Pearson correlation	Standard error	Wald test ($r=0$)	p-value	State	Pearson correlation	Standard error	Wald test ($r=0$)	p-value
USA	−0.7054***	(0.0536)	[13.1672]	<0.0001	Missouri	−0.2627	(0.1608)	[1.6333]	0.1024
Alabama	−0.6896***	(0.0748)	[9.2185]	<0.0001	Montana	−0.063	(0.1727)	[0.3651]	0.7151
Alaska	−0.1162	(0.1276)	[0.9107]	0.3625	Nebraska	−0.2763*	(0.1503)	[1.8381]	0.0661
Arizona	−0.313**	(0.1292)	[2.4225]	0.0154	Nevada	−0.3452**	(0.1519)	[2.273]	0.0230
Arkansas	0.4282***	(0.1105)	[3.8742]	0.0001	New Hampshire	−0.406***	(0.1432)	[2.8349]	0.0046
California	−0.4123***	(0.1300)	[3.1711]	0.0015	New Jersey	−0.065	(0.2013)	[0.3227]	0.7469
Colorado	0.435**	(0.1761)	[2.4694]	0.0135	New Mexico	−0.1474	(0.1367)	[1.0783]	0.2809
Connecticut	−0.1266	(0.1895)	[0.668]	0.5041	New York	−0.5925***	(0.0790)	[7.5016]	<0.0001
Delaware	0.182	(0.2004)	[0.908]	0.3639	North Carolina	−0.3172**	(0.1561)	[2.032]	0.0421
DC	−0.3464**	(0.1632)	[2.1219]	0.0338	North Dakota	0.2567	(0.1705)	[1.5056]	0.1322
Florida	−0.3171**	(0.1559)	[2.034]	0.0420	Ohio	−0.1645	(0.1979)	[0.8311]	0.4059
Georgia	−0.3467**	(0.1462)	[2.3708]	0.0178	Oklahoma	−0.1703	(0.1713)	[0.9944]	0.3200
Hawaii	−0.1591	(0.1692)	[0.9405]	0.3470	Oregon	0.4605***	(0.1432)	[3.2154]	0.0013
Idaho	0.0614	(0.1436)	[0.4276]	0.6689	Pennsylvania	−0.3645**	(0.1446)	[2.5218]	0.0117
Illinois	0.2501*	(0.1512)	[1.6541]	0.0981	Rhode Island	−0.0366	(0.1805)	[0.2031]	0.8391
Indiana	0.0162	(0.1884)	[0.086]	0.9314	South Carolina	−0.2094	(0.1400)	[1.4958]	0.1347
Iowa	−0.2172	(0.1539)	[1.4112]	0.1582	South Dakota	0.3518*	(0.1920)	[1.8323]	0.0669
Kansas	0.1141	(0.1748)	[0.6531]	0.5137	Tennessee	−0.3878***	(0.1495)	[2.5937]	0.0095
Kentucky	−0.2789*	(0.1663)	[1.677]	0.0935	Texas	0.0223	(0.1931)	[0.1157]	0.9079
Louisiana	−0.2422	(0.1713)	[1.4141]	0.1573	Utah	−0.2135	(0.1448)	[1.4749]	0.1402
Maine	−0.1811	(0.1387)	[1.3062]	0.1915	Vermont	−0.3255**	(0.1549)	[2.1007]	0.0357
Maryland	−0.0385	(0.2045)	[0.1884]	0.8505	Virginia	−0.286**	(0.1414)	[2.0228]	0.0431
Massachusetts	−0.4285***	(0.1421)	[3.0152]	0.0026	Washington	−0.5805***	(0.0835)	[6.9492]	<.0001
Michigan	−0.1045	(0.1757)	[0.5949]	0.5519	West Virginia	0.0033	(0.0426)	[0.0781]	0.9378
Minnesota	−0.3513**	(0.1550)	[2.2657]	0.0235	Wisconsin	−0.3972***	(0.1285)	[3.09]	0.002
Mississippi	0.308	(0.1975)	[1.5599]	0.1188	Wyoming	0.396**	(0.1840)	[2.1524]	0.0314

Table 3. Pearson correlation analysis by state. *$p<0.1$; **$p<0.05$; ***$p<0.01$.

To perform a first assessment of the relationship between Google Trends and COVID-19 data, the Pearson and Kendall rank correlations between the two variables are calculated, and the results are further compared. Tables 3 and 4 present the results of the Pearson and Kendall correlation analysis by state, respectively.

As reported in Table 3, statistically significant correlations are observed for the United States and for the states of Alabama, Arkansas, California, Colorado, Florida, Georgia, Illinois, Kentucky, Massachusetts, Minnesota, Nebraska, Nevada, New Hampshire, New York, North Carolina, Oregon, Pennsylvania, South Dakota, Tennessee, Vermont, Virginia, Washington, Wisconsin, and Wyoming as well as DC. The states of Iowa, Louisiana, Maine, Mississippi, Missouri, North Dakota, South Carolina, and Utah do not marginally reach the $p<0.1$ threshold of statistical significance, i.e., $p \in (0.1, 0.2)$.

Based on the Kendall correlation analysis, statistically significant correlations are observed for the United States and for the states of Alaska, Arizona, Arkansas, California, Connecticut, Florida, Georgia, Hawaii, Iowa, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Missouri, Montana, Nebraska, Nevada, New Hampshire, New Mexico, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Tennessee, Utah, Vermont, Virginia, Washington, and Wisconsin as well as DC. Figure 4 depicts

State	Kendall correlation	Standard error	Wald test (r = 0)	*p*-value	State	Kendall correlation	Standard error	Wald test (*r* = 0)	*p*-value
USA	−0.6230***	(0.0780)	[7.9891]	1.36E−15	Missouri	−0.2919**	(0.1187)	[2.4585]	0.0140
Alabama	−0.0679	(0.1389)	[0.4887]	0.6251	Montana	−0.2903**	(0.1405)	[2.0660]	0.0388
Alaska	−0.2713**	(0.1279)	[2.1218]	0.0339	Nebraska	−0.3589***	(0.1216)	[2.9517]	0.0032
Arizona	−0.3372**	(0.1313)	[2.5684]	0.0102	Nevada	−0.2989**	(0.1424)	[2.0996]	0.0358
Arkansas	0.4083***	(0.1497)	[2.7278]	0.0064	New Hampshire	−0.3397***	(0.1313)	[2.5884]	0.0096
California	−0.2801**	(0.1285)	[2.1794]	0.0293	New Jersey	−0.0690	(0.1451)	[0.4759]	0.6342
Colorado	0.0510	(0.1459)	[0.3498]	0.7265	New Mexico	−0.2851**	(0.1184)	[2.4070]	0.0161
Connecticut	−0.3060**	(0.1371)	[2.2320]	0.0256	New York	−0.4379***	(0.0871)	[5.0283]	0.0000
Delaware	−0.0095	(0.1545)	[0.0618]	0.9507	North Carolina	−0.2817**	(0.1305)	[2.1582]	0.0309
DC	−0.4986***	(0.1119)	[4.4565]	0.0000	North Dakota	0.2737*	(0.1507)	[1.8160]	0.0694
Florida	−0.3247**	(0.1323)	[2.4538]	0.0141	Ohio	−0.4007***	(0.1350)	[2.9683]	0.0030
Georgia	−0.3262**	(0.1290)	[2.5291]	0.0114	Oklahoma	−0.2902**	(0.1400)	[2.0725]	0.0382
Hawaii	−0.2372*	(0.1262)	[1.8805]	0.0600	Oregon	0.2751**	(0.1320)	[2.0830]	0.0373
Idaho	−0.1065	(0.1435)	[0.7425]	0.4578	Pennsylvania	−0.4173***	(0.1192)	[3.5013]	0.0005
Illinois	−0.1379	(0.1369)	[1.0077]	0.3136	Rhode Island	−0.1088	(0.1497)	[0.7266]	0.4675
Indiana	−0.0738	(0.1344)	[0.5491]	0.5830	South Carolina	−0.1900	(0.1172)	[1.6215]	0.1049
Iowa	−0.4162***	(0.1172)	[3.5507]	0.0004	South Dakota	−0.1255	(0.1641)	[0.7645]	0.4446
Kansas	−0.0851	(0.1480)	[0.5752]	0.5651	Tennessee	−0.3333***	(0.1236)	[2.6974]	0.0070
Kentucky	−0.3496***	(0.1275)	[2.7423]	0.0061	Texas	0.0202	(0.1346)	[0.1502]	0.8806
Louisiana	−0.3701***	(0.1345)	[2.7529]	0.0059	Utah	−0.3029***	(0.1138)	[2.6617]	0.0078
Maine	−0.3012**	(0.1388)	[2.1690]	0.0301	Vermont	−0.3658***	(0.1298)	[2.8179]	0.0048
Maryland	−0.2630**	(0.1301)	[2.0218]	0.0432	Virginia	−0.4270***	(0.1141)	[3.7409]	0.0002
Massachusetts	−0.3833***	(0.1377)	[2.7829]	0.0054	Washington	−0.4560***	(0.0909)	[5.0152]	0.0000
Michigan	−0.3908***	(0.1466)	[2.6658]	0.0077	West Virginia	−0.0733	(0.1126)	[0.6515]	0.5147
Minnesota	−0.3785***	(0.1383)	[2.7372]	0.0062	Wisconsin	−0.3506***	(0.1191)	[2.9441]	0.0032
Mississippi	0.0992	(0.1486)	[0.6679]	0.5042	Wyoming	−0.0416	(0.1481)	[0.2811]	0.7786

Table 4. Kendall rank correlation analysis by state. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

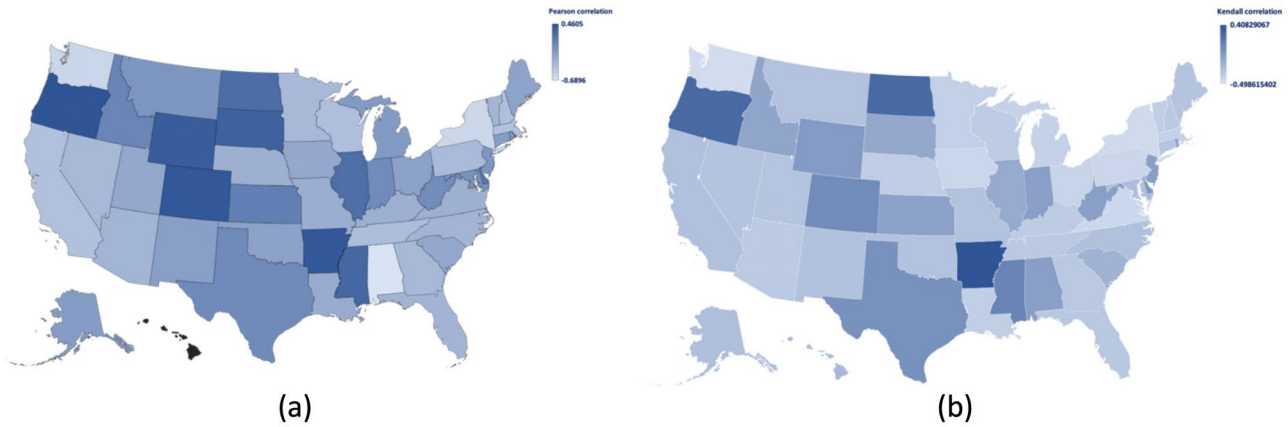


Figure 4. Heat map of the (**a**) Pearson and (**b**) Kendall correlation coefficients by state (Microsoft Excel).

the heat map of the (a) Pearson and (b) Kendall correlation coefficients in the United States by state over the period examined.

As depicted in the heat maps and in the spider web charts for the respective correlation analyses in Fig. 5, visual comparison of the two approaches indicates that the results are consistent in both analyses.

However, the main purpose of this study is to explore the predictability of COVID-19 using Google Trends data in the United States. Proceeding with the results of the predictability analysis, Fig. 6 depicts the heat map for $\boldsymbol{\beta}_1$ by state, while Table 5 presents the quantile regression estimated predictability models for the US and for each US state (plus DC). As shown, the estimated Google Trends models exhibit strong COVID-19 predictability.

Note that due to the low number of observations, the states of Maine, Montana, North Dakota, West Virginia, and Wyoming are not included in the predictability analysis results, but they are given the value "zero (0)" to be included in the heat map for purposes of uniformity.

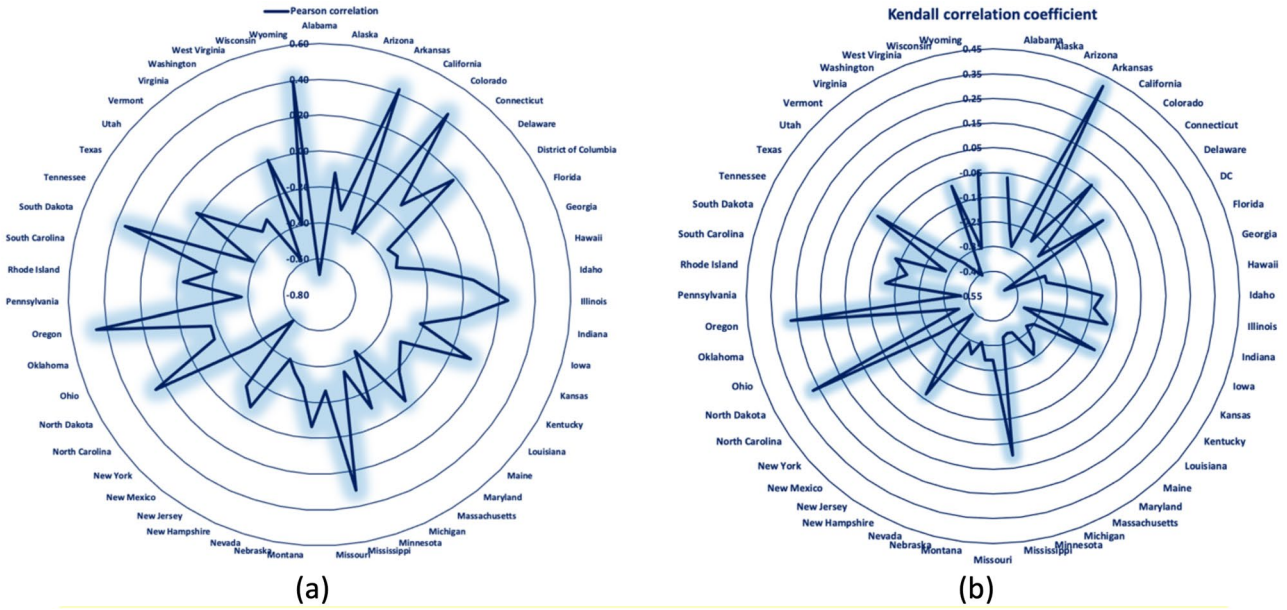


Figure 5. Radar chart of the (**a**) Pearson and (**b**) Kendall correlation coefficients by state (Microsoft Excel).

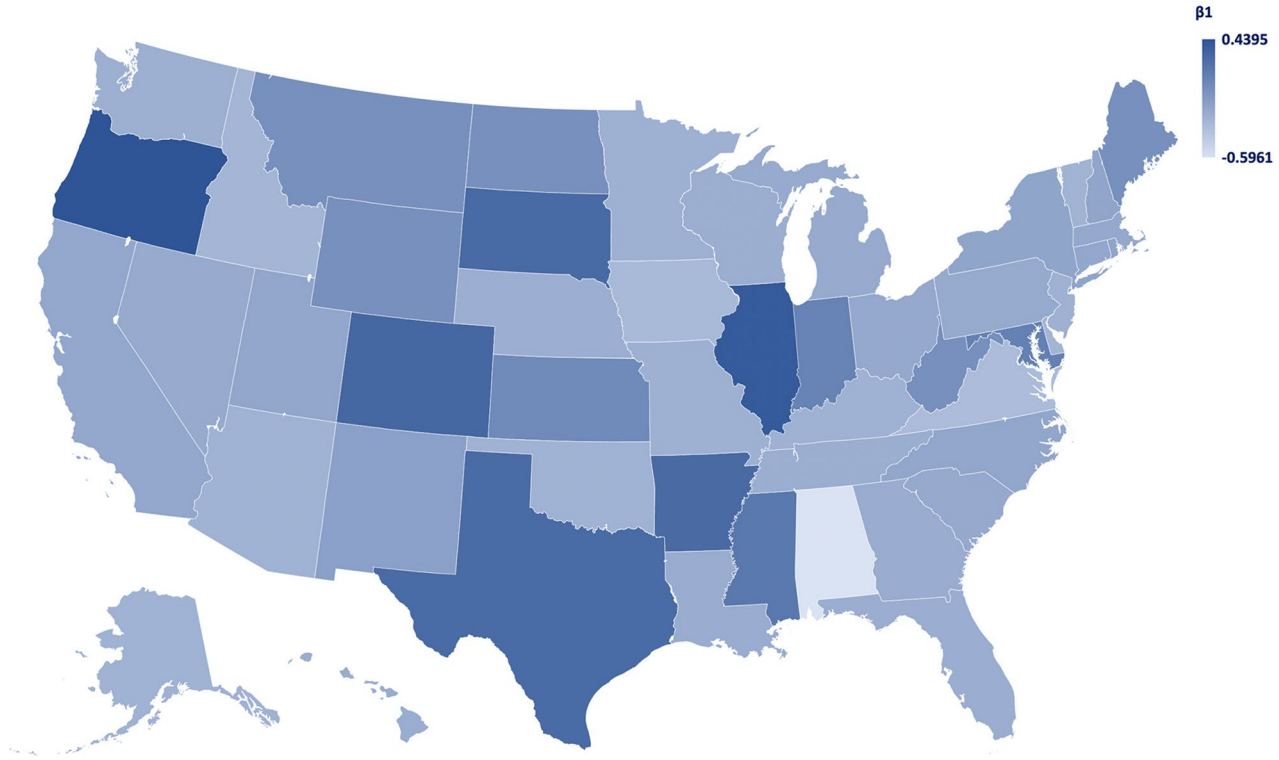


Figure 6. Heat map of β_1 of the predictability analysis models by state (Microsoft Excel).

Discussion

As of July 29th, 2020, there were 16,920,857 COVID-19 recorded cases worldwide, with the reported death toll at 664,141 and the number of recovered patients at 10,485,316[9]. In light of the COVID-19 pandemic and to find new ways of forecasting the spread of the disease, infodemiology approaches have provided valuable input in monitoring and forecasting the development of the COVID-19 pandemic over time and in measuring and analyzing the public's awareness and response. Google Trends and Twitter have been identified as the most popular infodemiology sources, while other social media, such as Facebook and Instagram, exhibit promising results in analyzing users' online behavioral patterns[13].

	β_0			β_1			β_2		
USA	−0.0509	(0.4339)	[−0.1172]	−0.7506***	(0.2197)	[−3.4173]	−0.0014	(0.0169)	[−0.0831]
AL	0.8944***	(0.2176)	[4.1099]	−0.5961***	(0.1160)	[−5.1383]	−0.0413***	(0.0070)	[−5.8850]
AK	−1.4528***	(0.2003)	[−7.2539]	−0.2449**	(0.1006)	[−2.4341]	0.0663***	(0.0087)	[7.6030]
AZ	−1.4183***	(0.1309)	[−10.8362]	−0.2429***	(0.0817)	[−2.9745]	0.0637***	(0.0049)	[12.8777]
AR	−0.2565	(0.4658)	[−0.5507]	0.2785	(0.2531)	[1.1004]	0.0023	(0.0124)	[0.1825]
CA	−1.4274***	(0.0936)	[−15.2521]	−0.1634***	(0.0539)	[−3.0325]	0.0642***	(0.0046)	[13.8481]
CO	−0.9688***	(0.1916)	[−5.0561]	0.3007	(0.2587)	[1.1623]	0.0290***	(0.0074)	[3.9132]
CT	−1.7866***	(0.0654)	[−27.3353]	−0.1645***	(0.0470)	[−3.4989]	0.0782***	(0.0026)	[30.6221]
DE	−2.0415***	(0.4639)	[−4.4003]	−0.2687	(0.2446)	[−1.0987]	0.0715***	(0.0110)	[6.4873]
DC	−1.3077***	(0.1980)	[−6.6064]	−0.1548*	(0.0849)	[−1.8228]	0.0578***	(0.0094)	[6.1513]
FL	−1.5483***	(0.0766)	[−20.2209]	−0.2128***	(0.0431)	[−4.9412]	0.0715***	(0.0024)	[29.3170]
GA	−1.5727***	(0.0808)	[−19.4690]	−0.2047***	(0.0570)	[−3.5898]	0.0721***	(0.0042)	[17.2658]
HI	−1.6732***	(0.0873)	[−19.1647]	−0.2083***	(0.0470)	[−4.4343]	0.0758***	(0.0041)	[18.3027]
ID	−1.8929***	(0.1465)	[−12.9167]]	−0.2686***	(0.0663)	[−4.0507]	0.0866***	(0.0067)	[12.8631]
IL	−1.4466***	(0.1404)	[−10.3063]	0.3943***	(0.0707)	[5.5764]	0.0680***	(0.0056)	[12.2022]
IN	−1.4674***	(0.2157)	[−6.8020]	0.0977	(0.1624)	[0.6018]	0.0693***	(0.0065)	[10.7392]
IA	−1.5912***	(0.1402)	[−11.3507]	−0.2957***	(0.0733)	[−4.0346]	0.0732***	(0.0042)	[17.3342]
KS	−1.5579***	(0.2298)	[−6.7799]	0.0463	(0.1101)	[0.4204]	0.0635***	(0.0106)	[5.9774]
KY	−1.5530***	(0.1396)	[−11.1222]	−0.2415***	(0.0599)	[−4.0291]	0.0719***	(0.0062)	[11.5292]
LA	−1.6432***	(0.0602)	[−27.2763]	−0.2050***	(0.0357)	[−5.7381]	0.0751***	(0.0026)	[28.6534]
MD	−1.1066***	(0.2339)	[−4.7306]	0.1135	(0.1008)	[1.1255]	0.0550***	(0.0088)	[6.2834]
MA	−1.6424***	(0.0771)	[−21.3061]	−0.1757***	(0.0538)	[−3.2668]	0.0742***	(0.0034)	[21.8651]
MI	−1.7657***	(0.0813)	[−21.7133]	−0.1884***	(0.0406)	[−4.6375]	0.0800***	(0.0032)	[25.2349]
MN	−1.6085***	(0.0773)	[−20.7963]	−0.2344***	(0.0521)	[−4.4970]	0.0728***	(0.0027)	[26.9966]
MS	−1.3047***	(0.2959)	[−4.4088]	0.1773	(0.1600)	[1.1086]	0.0570***	(0.0082)	[6.9200]
MO	−1.5382***	(0.0883)	[−17.4271]	−0.2326***	(0.0478)	[−4.8610]	0.0718***	(0.0051)	[14.0987]
NE	−1.4875***	(0.1909)	[−7.7908]	−0.2192***	(0.0746)	[−2.9375]	0.0717***	(0.0063)	[11.3935]
NV	−1.6778***	(0.0862)	[−19.4683]	−0.1872***	(0.0348)	[−5.3846]	0.0763***	(0.0037)	[20.4946]
NH	−1.6586***	(0.0723)	[−22.9526]	−0.1515***	(0.0365)	[−4.1562]	0.0741***	(0.0025)	[30.0037]
NJ	−1.8518***	(0.2428)	[−7.6277]	−0.2395	(0.2427)	[−0.9867]	0.0688***	(0.0060)	[11.3949]
NM	−1.2414***	(0.1640)	[−7.5679]	−0.1188	(0.0803)	[−1.4805]	0.0593***	(0.0066)	[8.9371]
NY	−1.2201***	(0.0468)	[−26.0596]	−0.1482***	(0.0562)	[−2.6358]	0.0482***	(0.0043)	[11.2916]
NC	−1.6575***	(0.0953)	[−17.3914]	−0.1613***	(0.0476)	[−3.3848]	0.0722***	(0.0038)	[18.8471]
OH	−1.8408***	(0.1464)	[−12.5751]	−0.1758**	(0.0750)	[−2.3436]	0.0790***	(0.0048)	[16.3817]
OK	−1.7038***	(0.0544)	[−31.2986]	−0.2463***	(0.0318)	[−7.7497]	0.0767***	(0.0026)	[29.5090]
OR	−0.7953***	(0.2019)	[−3.9392]	0.4395***	(0.1362)	[3.2257]	0.0293***	(0.0069)	[4.2697]
PA	−1.3917***	(0.1279)	[−10.8769]	−0.1845**	(0.0758)	[−2.4348]	0.0716***	(0.0041)	[17.5561]
RI	−1.4924***	(0.0752)	[−19.8418]	−0.1461***	(0.0408)	[−3.5844]	0.0588***	(0.0049)	[12.1036]
SC	−1.2889***	(0.0941)	[−13.7030]	−0.1816***	(0.0513)	[−3.5395]	0.0520***	(0.0069)	[7.5216]
SD	−1.1230***	(0.2939)	[−3.8212]	0.2815**	(0.1388)	[2.0277]	0.0537***	(0.0084)	[6.4280]
TN	−1.5098***	(0.0658)	[−22.9294]	−0.2157***	(0.0524)	[−4.1179]	0.0676***	(0.0020)	[33.1730]
TX	−1.4766***	(0.3041)	[−4.8557]	0.2749	(0.1903)	[1.4442]	0.0660***	(0.0077)	[8.5342]
UT	−1.4381***	(0.1399)	[−10.2768]	−0.1586**	(0.0723)	[−2.1944]	0.0720***	(0.0069)	[10.3640]
VT	−1.5359***	(0.1854)	[−8.2848]	−0.2499***	(0.0848)	[−2.9476]	0.0770***	(0.0081)	[9.5352]
VA	−1.5878***	(0.2504)	[−6.3400]	−0.3147***	(0.1021)	[−3.0837]	0.0767***	(0.0106)	[7.2484]
WA	−1.3476***	(0.1540)	[−8.7488]	−0.2236**	(0.1007)	[−2.2212]	0.0660***	(0.0101)	[6.5118]
WI	−1.3407***	(0.0992)	[−13.5142]	−0.2143***	(0.0698)	[−3.0711]	0.0618***	(0.0053)	[11.6287]

Table 5. Predictability analysis by state. The numbers in parentheses report the standard errors; the t-statistics are given in brackets. ***, ** and * indicate statistical significance at the 0.01, 0.05 and 0.1 levels, respectively. The corresponding critical values are 2.575, 1.96 and 1.645.

Social media platforms can provide us with more qualitative data that can shift the focus to other directions. Such approaches include sentiment analysis, educational purposes, and efforts to measure and raise public awareness. Recent approaches to analyzing aspects of the COVID-19 pandemic using social media data include monitoring the Twitter usage of G7 leaders[58], monitoring self-reported symptoms on Twitter[59], and analyzing the

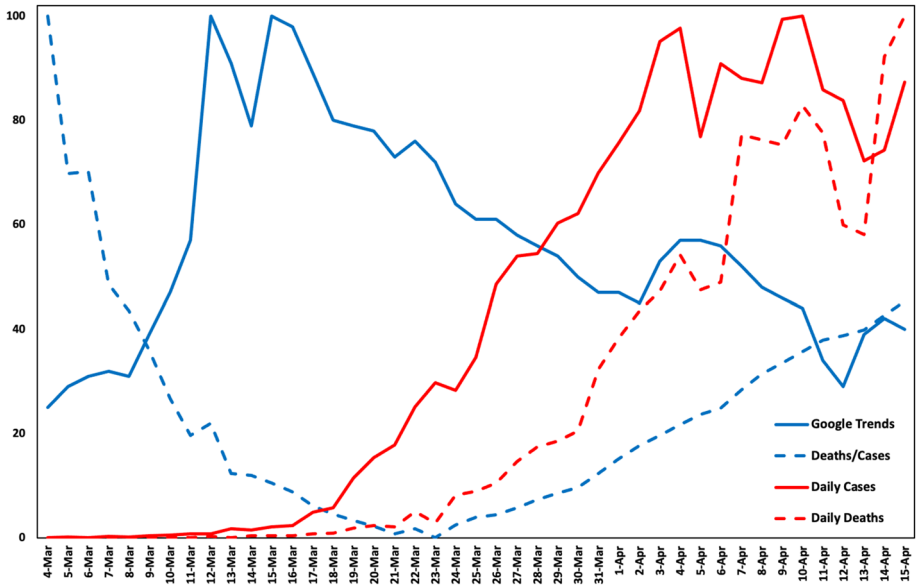


Figure 7. COVID-19 and Google Trends data from March 4th to April 15th in the US (Microsoft Excel).

public perception of the disease through Facebook[60]. Moreover, infodemiology sources have provided valuable input in recruiting online survey participants through Facebook to measure individuals' COVID-19 confidence levels[61] and in assessing the behavioral variations in COVID-19-related online search traffic in more than one search engine[62]. Finally, commentaries that make recommendations on the integration of other social media platforms, such as Facebook, Reddit, and TikTok, for disseminating medical information to inform public health and policy have been published[63].

Google Trends offers a solid foundation for quantitative analysis with respect to the monitoring and predictability of COVID-19, as in the analysis presented in this study, where Google Trends data on the "coronavirus (virus)" topic were used to explore the predictability of COVID-19 in the United States at both national and state level. First, for a preliminary assessment of the relationship between Google Trends and COVID-19 data, Pearson correlation and Kendall rank correlation analyses were performed. Statistically significant correlations were observed for the United States and for several US states, which is in line with previous studies that argue that there is a relationship between Google Trends and COVID-19 data.

The COVID-19 predictability analysis, which used a quantile regression approach, exhibits very promising results and indicates the most important contribution of this study to the international literature: detecting and predicting the early spread of COVID-19 at the regional level. This contribution can be a substantial supplement in further assisting local authorities in taking the appropriate measures to handle the spread of the disease.

Figure 7 illustrates a graph of the COVID-19 deaths/cases ratio, daily COVID-19 deaths, daily COVID-19 cases, and the respective Google Trends normalized data in the United States from March 4th to April 15th, 2020. For purposes of consistency in the graph, the COVID-19-related time series are normalized on a 0–100 scale. As depicted in the graph and confirmed by the predictability analysis, the two variables are not linearly dependent. Instead, they exhibit an inversely proportional relationship, meaning that as COVID-19 progresses, the online interest in the virus decreases.

From a behavioral point of view, this result can be explained as follows. First, online interest starts to increase and reaches a peak as the number of confirmed cases becomes high and as the deaths rates start to show that the pandemic does indeed have severe consequences. However, after a certain period, the interest has an inverse course, which could also indicate that the public is overwhelmed by information overload and decreases its information "intake". The spike in Google queries and the decline in the ratio of COVID-19 deaths/cases could be attributed to the spread of the virus over these days and the "delay" in deaths. Regarding this latter point, this means that cases increase while the total number of deaths has not yet started to considerably increase.

The latter point is in line with previous work on the topic[27] suggesting that although significant correlations between COVID-19 and Google data are observed, the relationship tends to decrease in both strength and significance in regions that have been affected by COVID-19 as we move forward in time because the interest in the virus decreases. This decrease is counterintuitive and occurs before the case and death curves start to exhibit a downward trend, i.e., when a region is being heavily affected, independent of whether or not it has reached its peak. However, it would be interesting for future investigators to explore the relationship from this point onwards since, as shown in Fig. 7, the lines converge, with this convergence being indicative of a future change in the relationship dynamics when deaths peak at a later point and when they start their downward course as well.

The above can partly explain the differences in signs among states in both the Pearson and Kendall rank correlation coefficients, but a more in-depth explanation from a statistical perspective is that the Pearson correlation coefficient is estimated as the average of the deviations of observations from the sample mean. The weights

of observations in the tails of the distribution are equal to the weight of other observations, and therefore, the outliers could affect the estimation of the results, especially in the case of the small sample. In consideration of ties, this study employs a bootstrap bias-corrected approach, but the main conclusions are based on quantile regressions. Unlike linear measures of dependency, quantile regression is considered superior in a sampling situation and more resistant to outliers than linear regressions, the Pearson correlation, or the Kendall rank correlation[64]. Taking into account that the current pandemic is a dynamic process that constantly evolves and has a serious social impact, it is very probable that there now exist—or, at a later stage, could develop—several data anomalies (e.g., due to non-pharmaceutical interventions); therefore, formal statistical tools such as the Pearson and Kendall rank correlations should be carefully interpreted.

This study has limitations. First, data from only one search engine are considered. Although Google Trends is the most popular search engine, some data on the coronavirus topic from other search engines were not included in this analysis. Second, the data at this point are very limited, and the results are based on few observations. Third, the 50 (+ 1) states exhibit diversity in terms of confirmed cases and deaths. Therefore, any conclusions drawn from this analysis refer to each case individually. Despite the known limitations of online search traffic data, the use of infodemiology metrics for informing public health and policy in general and for monitoring outbreaks and epidemics in particular has received wide attention.

To dynamically find the determinants of COVID-19, the predictability analysis in this study provides insights into how online search traffic data can play a considerable role in forming public health policies, especially in times of epidemics and outbreaks, when real-time data are essential. With the COVID-19 pandemic, the world is in uncharted territory socially, economically, and socially. This situation calls for immediate action and open research and data, and the term "multidisciplinary" has never before been more important. To that end, the role of big data in providing "*opportunities for performing modeling studies of viral activity and for guiding individual country healthcare policymakers to enhance preparation for the outbreak*" has been acknowledged[65], and current research on the subject should focus on both exploring the role of other infodemiology variables in the predictability of COVID-19 and combining infodemiology sources with traditional sources to explore the full potential of what online real-time data have to offer for disease surveillance.

Data availability

The COVID-19 and query datasets analyzed during the current study are available on the COVID-19 Tracking Project website[10] and on the "Google Trends" explore page[39], respectively.

Received: 27 April 2020; Accepted: 6 November 2020
Published online: 26 November 2020

References

1. WHO Timeline—COVID-19. *World Health Organization*. https://www.who.int/news-room/detail/08-04-2020-who-timeline---covid-19 (2020).
2. Twitter account. *World Health Organization*. https://twitter.com/WHO/status/1213523866703814656?s=20 (2020).
3. Pneumonia of unknown cause. *World Health Organization*. https://www.who.int/csr/don/05-january-2020-pneumonia-of-unknown-cause-china/en/ (2020).
4. Secon, H., Woodward, A & Mosher, D. A comprehensive timeline of the new coronavirus pandemic, from China's first COVID-19 case to the present. *Business Insider*. https://www.businessinsider.com/coronavirus-pandemic-timeline-history-major-events-2020-3 (2020).
5. Twitter account. *World Health Organization*. https://twitter.com/who/status/1217043229427761152?lang=en (2020).
6. Qin, A. & Wang, V. Wuhan, Center of Coronavirus Outbreak, Is Being Cut Off by Chinese Authorities. *New York Times*. https://www.nytimes.com/2020/01/22/world/asia/china-coronavirus-travel.html (2020).
7. Coronavirus disease named COVID-19. *BBC News*. https://www.bbc.com/news/world-asia-china-51466362 (2020).
8. COVID coronavirus Outbreak: Italy. *Wolrdometer*. https://www.worldometers.info/coronavirus/country/italy/ (2020).
9. COVID coronavirus Outbreak. *Worldometer*. https://www.worldometers.info/coronavirus/ (2020).
10. The COVID Tracking Project. *The Atlantic*. https://covidtracking.com (2020).
11. Eysenbach, G. Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J. Med. Internet Res.* **11**(1), e11 (2009).
12. Eysenbach, G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. *Am. J. Prev. Med.* **40**(5 Suppl 2), S154–S158 (2011).
13. Mavragani, A. Infodemiology and infoveillance: A scoping review. *J. Med. Internet Res.* **22**(4), e16206 (2020).
14. Bernardo, T. M. *et al.* Scoping review on search queries and social media for disease surveillance: A chronology of innovation. *J. Med. Internet Res.* **15**(7), e147 (2013).
15. Eysenbach, G. SARS and population health technology. *J. Med. Internet Res.* **5**(2), e14 (2003).
16. van Lent, L. G., Sungur, H., Kunneman, F. A., van de Velde, B. & Das, E. Too far to care? Measuring public attention and fear for Ebola using twitter. *J. Med. Internet Res.* **19**(6), e193 (2017).
17. Farhadloo, M., Winneg, K., Chan, M. S., Hall, J. K. & Albarracin, D. Associations of topics of discussion on twitter with survey measures of attitudes, knowledge, and behaviors related to Zika: Probabilistic Study in the United States. *JMIR Public Health Surveill.* **4**(1), e16 (2018).
18. Poletto, C., Boëlle, P. & Colizza, V. Risk of MERS importation and onward transmission: A systematic review and analysis of cases reported to WHO. *BMC Infect. Dis.* **16**(1), 448 (2016).
19. Samaras, L., García-Barriocanal, E. & Sicilia, M. A. Comparing Social media and Google to detect and predict severe epidemics. *Sci. Rep.* **10**, 4747 (2020).
20. Mavragani, A. & Ochoa, G. The internet and the anti-vaccine movement: Tracking the 2017 EU measles outbreak. *Big Data Cog. Comp.* **2**(1), 1 (2018).
21. Du, J. *et al.* Public perception analysis of tweets during the 2015 measles outbreak: Comparative study using convolutional neural network models. *J. Med. Internet Res.* **20**(7), e236 (2018).
22. Mavragani, A., Ochoa, G. & Tsagarakis, K. P. Assessing the methods, tools, and statistical approaches in google trends research: Systematic review. *J. Med. Internet Res.* **20**(11), e270 (2018).
23. Google Trends & COVID Advanced Search. *Pubmed*. https://www.ncbi.nlm.nih.gov/pubmed/ (2020).

24. Husnayain, A., Fuad, A. & Su, E. C. Applications of google search trends for risk communication in infectious disease management: A case study of COVID-19 outbreak in Taiwan. *Int. J. Infect Dis.* **95**, 221–223 (2020).
25. Li, C. *et al.* Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. *Euro Surveill.* **25**(10), 2000199 (2020).
26. Effenberger, M. *et al.* Association of the COVID-19 pandemic with internet search volumes: A Google Trends(TM) analysis. *Int. J. Infect Dis.* **95**, 192–197 (2020).
27. Mavragani, A. Tracking COVID-19 in Europe: Infodemiology approach. *JMIR Public Health Surveill.* **6**(2), e18941 (2020).
28. Walker, A., Hopkins, C. & Surda, P. The use of google trends to investigate the loss of smell related searches during COVID-19 outbreak. *Int. Forum Allergy Rhinol.* **10**(7), 839–847 (2020).
29. Hong, Y. R., Lawrence, J., Williams, D. Jr. & Mainous, A. Population-level interest and telehealth capacity of US hospitals in response to COVID-19: Cross-sectional analysis of google search and national hospital survey data. *JMIR Public Health Surveill.* **6**(2), e18961 (2020).
30. Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M. R. & Kalhori, S. N. Predicting COVID-19 incidence through analysis of google trends data in Iran: Data mining and deep learning pilot study. *JMIR Public Health Surveill.* **6**(2), e18828 (2020).
31. Rufai, S.R. & Bunce, C. World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. *J Public Health (Oxf).* fdaa049 (2020).
32. Kouzy, R. *et al.* Coronavirus goes viral: Quantifying the COVID-19 misinformation epidemic on twitter. *Cureus.* **12**(3), e7255 (2020).
33. Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M. & Shah, Z. Top concerns of tweeters during the COVID-19 pandemic: A surveillance study. *J. Med. Internet Res.* **22**(40), e19016 (2020).
34. Dost, B. *et al.* Attitudes of anesthesiology specialists and residents toward patients infected with the novel coronavirus (COVID-19): A national survey study. *Surg. Infect. (Larchmt).* **21**(4), 350–356 (2020).
35. Simcock, R. *et al.* COVID-19: Global radiation oncology's targeted response for pandemic preparedness. *Clin. Transl. Radiat. Oncol.* **22**, 55–68 (2020).
36. Kim, B. Effects of social grooming on incivility in COVID-19. *Cyberpsychol. Behav. Soc. Netw.* **23**(8), 519–525 (2020).
37. Rosenberg, H., Syed, S. & Rezaie, S. The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. *CJEM.* **6**, 1–4 (2020).
38. Chan, A.K.M., Nickson, C.P., Rudolph, J.W., Lee, A. & Joynt, G.M. Social media for rapid knowledge dissemination: Early experience from the COVID-19 pandemic. *Anaesthesia*. (2020)
39. Google Trends Explore. https://trends.google.com/trends/explore. (April 18, 2020).
40. Trends Help. Google Support. https://support.google.com/trends/answer/4365533?hl=en (2020).
41. Mavragani, A. & Ochoa, G. Google trends in infodemiology and infoveillance: Methodology framework. *JMIR Public Health Surveill.* **5**(2), e13439 (2019).
42. PixelMap. *AMCHARTS*. https://pixelmap.amcharts.com (2020).
43. *ChartsBin*. https://chartsbin.com (2020).
44. Phillips, P. C. B. & Perron, P. Testing for a unit root in time series regression. *Biometrica.* **75**(2), 335–346 (1988).
45. Efron, B. & Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* **1**(1), 54–75 (1986).
46. Karlsson, A. Bootstrap methods for bias correction and confidence interval estimation for nonlinear quantile regression of longitudinal data. *J. Stat. Comput. Sim.* **79**(10), 1205–1218 (2009).
47. Guan, W. From the help desk: Bootstrapped standard errors. *Stata J.* **3**(1), 71–80 (2003).
48. Davidson, R. & MacKinnon, J. G. Bootstrap tests: How many bootstraps?. *Econ. Rev.* **19**(1), 55–68 (2000).
49. Koenker, R. & Bassett, G. Regression quantiles. *Econometrica.* **46**(1), 33–50 (1978).
50. Koenker, R. & Hallock, K. F. Quantile regression. *J. Econ. Percepct.* **15**(4), 143–156 (2001).
51. Yu, K., Lu, Z. & Stander, J. Quantile regression: Applications and current research areas. *J. R Stat. Soc. Series D Stat.* **52**(3), 331–350 (2003).
52. Nikitina, L., Paidi, R. & Furuoka, F. Using bootstrapped quantile regression analysis for small sample research in applied linguistics: Some methodological considerations. *PLoS ONE* **14**(1), e0210668 (2019).
53. Chen, F. & Chalhoub-Deville, M. Principles of quantile regression and an application. *Lang. Test.* **31**(1), 63–87 (2014).
54. Firpo, S., Fortin, N. M. & Lemieux, T. Unconditional quantile regressions. *Econometrica.* **77**(3), 953–973 (2009).
55. Salibian-Barrera, M. & Zamar, R. H. Bootrapping robust estimates of regression. *Ann. Stat.* **30**(2), 556–582 (2002).
56. Chernozhukov, V., Hansen, C. & Jansson, M. Finite sample inference for quantile regression models. *J. Econom.* **152**, 93–103 (2009).
57. R Core Team, 2017. R: A language and environment for statistical computing, Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/. R version 3.3.3.
58. Rufai, R. S. & Bunce, C. World leaders' usage of Twitter in response to the COVID-19 pandemic: A content analysis. *J. Public Health.* **42**(3), 510–516 (2020).
59. Sarker, A. *et al.* Self-reported COVID-19 symptoms on Twitter: An analysis and a research resource. *J. Am. Med. Inform. Assoc.* **27**(8), 1310–1315 (2020).
60. Shorey, S., Ang, E., Yamina, A. & Tam, C. Perceptions of public on the COVID-19 outbreak in Singapore: a qualitative content analysis. *J Public Health* (Oxf). fdaa105, (2020).
61. Wang, P. W. *et al.* COVID-19-related information sources and the relationship with confidence in people coping with COVID-19: Facebook survey study in Taiwan. *J. Med. Internet Res.* **22**(6), e20021 (2020).
62. Hou, Z. *et al.* Cross-country comparison of public awareness, rumours, and behavioural responses to the COVID-19 epidemic: An internet surveillance study. *J. Med. Internet Res.* **22**(8), e21143 (2020).
63. Eghtesadi, M. & Florea, A. Facebook, Instagram, Reddit and TikTok: A proposal for health authorities to integrate popular social media platforms in contingency planning amid a global pandemic outbreak. *Can. J. Public Health.* **111**, 389–391 (2020).
64. Gideon, R. A. & Hollister, R. A. A rank correlation coefficient resistant to outliers. *J. Am. Stat. Assoc.* **82**(398), 656–666 (1987).
65. Ting, D. S. W., Carin, L., Dzau, V. & Wong, T. Y. Digital technology and COVID-19. *Nat. Med.* **26**, 459–461 (2020).

Author contributions

A.M. conceived the idea, designed the methodology, performed the data collection, performed the data analysis and interpretation, wrote the paper; K.G. designed the statistical methodology, performed the statistical analysis and interpretation and performed the computational analysis. Both authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-77275-9.

Correspondence and requests for materials should be addressed to A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *Journal of Medical Internet Research,* **2009**;11(1):e11.
2. Eysenbach G. Infodemiology: The epidemiology of (mis)information. *The American Journal of Medicine,* **2002**;113(9):763-765.
3. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. *American Journal of Preventive Medicine*, **2011**;40(5Suppl2):154-158.
4. Kind T, Wheeler KL, Robinson B, Cabana MD. Do the leading children's hospitals have quality Web sites? A description of children's hospital Web sites. *Journal of Medical Internet Research,* **2004**;6(2)e20.
5. Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. *AMIA Annual Symposium Proccedings*, **2006**:244-248.
6. Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping Review on Search Queries and Social Media for Disease Surveillance: A Chronology of Innovation. *Journal of Medical Internet Research*, **2013**;15(7):e147.
7. Social Media Factsheet. Pew Research Center: Internet and Technology. URL: https://www.pewinternet.org/fact-sheet/social-media/.
8. Mavragani A, Tsagarakis KP. Predicting Referendum Results in the Big Data Era. Journal of Big Data, **2019**;6:3.
9. Mavragani A, Tsagarakis KP. YES or NO: Predicting the 2015 GReferendum results using Google Trends. *Technological Forecasting and Social Change*, **2016**;109:1-5.
10. Mavragani A, Ochoa G. Google Trends in Infodemiology and Infoveillance: Methodology Framework. *JMIR Public Health and Surveillance*, **2019**;5(2):e13439.
11. Bian J, Zhao Y, Salloum RG, Guo Y, Wang M, Prosperi M, Zhang H, Du X, Ramirez-Diaz LJ, He Z, Sun Y. Using Social Media Data to Understand the Impact of Promotional Information on Laypeople's Discussions: A Case Study of Lynch Syndrome. *Journal of Medical Internet Research*, **2017**;19(12):e414.
12. Chen T, Dredze M. Vaccine Images on Twitter: Analysis of What Images are Shared. *Journal of Medical Internet Research*, **2018**;20(4):e130.

13. Hswen Y, Naslund JA, Brownstein JS, Hawkins JB. Monitoring Online Discussions About Suicide Among Twitter Users With Schizophrenia: Exploratory Study. *JMIR Mental Health*, **2018**;5(4):e11483.
14. Odlum M, Yoon S, Broadwell P, Brewer R, Kuang D. How Twitter Can Support the HIV/AIDS Response to Achieve the 2030 Eradication Goal: In-Depth Thematic Analysis of World AIDS Day Tweets. *JMIR Public Health and Surveillance*, **2018**;4(4):e10262.
15. Chen S, Xu Q, Buchenberger J, Bagavathi A, Fair G, Shaikh S, Krishnan S. Dynamics of Health Agency Response and Public Engagement in Public Health Emergency: A Case Study of CDC Tweeting Patterns During the 2016 Zika Epidemic. *JMIR Public Health and Surveillance,* **2018**;4(4):e10827.
16. Wakamiya S, Kawai Y, Aramaki E. Twitter-Based Influenza Detection After Flu Peak via Tweets With Indirect Information: Text Mining Study. *JMIR Public Health and Surveillance*, **2018**;4(3):e65.
17. Tufts C, Polsky D, Volpp KG, Groeneveld PW, Ungar L, Merchant RM, Pelullo AP. Characterizing Tweet Volume and Content About Common Health Conditions Across Pennsylvania: Retrospective Analysis. *JMIR Public Health and Surveillance*, **2018**;4(4):e10834.
18. Saha K, Weber I, Birnbaum ML, De Choudhury M. Characterizing Awareness of Schizophrenia Among Facebook Users by Leveraging Facebook Advertisement Estimates. *Journal of Medical Internet Research*, **2017**;19(5):e156.
19. Smith RJ, Crutchley P, Schwartz HA, Ungar L, Shofer F, Padrez KA, Merchant RM. Variations in Facebook Posting Patterns Across Validated Patient Health Conditions: A Prospective Cohort Study. *Journal of Medical Internet Research*, **2017**;19(1):e7.
20. Gittelman S, Lange V, Gotway Crawford CA, Okoro CA, Lieb E, Dhingra SS, Trimarchi E. A New Source of Data for Public Health Surveillance: Facebook Likes. *Journal of Medical Internet Research,* **2015**;17(4):e98.
21. Keller MS, Park HJ, Cunningham ME, Fouladian JE, Chen M, Spiegel BMR. Public Perceptions Regarding Use of Virtual Reality in Health Care: A Social Media Content Analysis Using Facebook. *Journal of Medical Internet Research*, **2017**;19(12):e419.
22. Mejova Y, Weber I, Fernandez-Luque L. Online Health Monitoring using Facebook Advertisement Audience Estimates in the United States: Evaluation Study. *JMIR Public Health and Surveillance,* **2018**;4(1):e30.

23. Cherian R, Westbrook M, Ramo D, Sarkar U. Representations of Codeine Misuse on Instagram: Content Analysis. *JMIR Public Health and Surveillance,* **2018**;4(1):e22.
24. Muralidhara S, Paul MJ. #Healthy Selfies: Exploration of Health Topics on Instagram. *JMIR Public Health and Surveillance,* **2018**;4(2):e10150.
25. Phillips CA, Barz Leahy A, Li Y, Schapira MM, Bailey LC, Merchant RM. Relationship Between State-Level Google Online Search Volume and Cancer Incidence in the United States: Retrospective Study. *Journal of Medical Internet Research*, **2018**; 20(1):e6.
26. Radin M, Sciascia S. Infodemiology of systemic lupus erythematous using Google Trends. *Lupus*, **2017**;26(8):886-889.
27. Seidl S, Schuster B, Rüth M, Biedermann T, Zink A. What Do Germans Want to Know About Skin Cancer? A Nationwide Google Search Analysis From 2013 to 2017. *Journal of Medical Internet Research*, **2018**;20(5):e10327.
28. Wang HW, Chen DR, Yu HW, Chen YM. Forecasting the Incidence of Dementia and Dementia-Related Outpatient Visits with Google Trends: Evidence From Taiwan. *Journal of Medical Internet Research*, **2015**;17(11):e264.
29. Berlinberg EJ, Deiner MS, Porco TC, Acharya NR. Monitoring Interest in Herpes Zoster Vaccination: Analysis of Google Search Data. *JMIR Public Health and Surveillance,* **2018**;4(2):e10180.
30. Mavragani A, Ochoa G. Infoveillance of infectious diseases in USA: STDs, tuberculosis, and hepatitis. *Journal of Big Data*, **2018**;5:30.
31. Mavragani A, Sampri A, Sypsa K, Tsagarakis KP. Integrating Smart Health in the US Health Care System: Infodemiology Study of Asthma Monitoring in the Google Era. *JMIR Public Health and Surveillance,* **2018**;4(1):e24.
32. Zhang Z, Zheng X, Zeng DD, Leischow SJ. Tracking Dabbing Using Search Query Surveillance: A Case Study in the United States. *Journal of Medical Internet Research*, **2016**;18(9):e252.
33. Zheluk A, Gillespie JA, Quinn C. Searching for Truth: Internet Search Patterns as a Method of Investigating Online Responses to a Russian Illicit Drug Policy Debate. *Journal of Medical Internet Research,* **2012**;14(6):e165.
34. Zheluk A, Quinn C, Meylakhs P. Internet search and krokodil in the Russian Federation: an infoveillance study. *Journal of Medical Internet Research*, **2014**;16(9):e212.

35. Domnich A., Arbuzova E.K., Signori A., Amicizia D., Panatto D., Gasparini R. Demand-based web surveillance of sexually transmitted infections in Russia. *International Journal of Public Health*, **2014**;59(5):841-849.
36. Agarwal V, Zhang L, Zhu J, Fang S, Cheng T, Hong C, Shah NH. Impact of Predicting Health Care Utilization Via Web Search Behavior: A Data-Driven Analysis. *Journal of Medical Internet Research*, **2016**;18(9):e251.
37. Yang H, Li S, Sun L, Zhang X, Hou J, Wang Y. Effects of the Ambient Fine Particulate Matter on Public Awareness of Lung Cancer Risk in China: Evidence from the Internet-Based Big Data Platform. *JMIR Public Health and Surveillance,* **2017**;3(4):e64.
38. Cartwright AF, Karunaratne M, Barr-Walker J, Johns NE, Upadhyay UD. Identifying National Availability of Abortion Care and Distance from Major US Cities: Systematic Online Search. *Journal of Medical Internet Research*, **2018**;20(5):e186.
39. Yom-Tov E, Gabrilovich E. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. *Journal of Medical Internet Research*, **2013**;15(6):e124.
40. Seo DW, Jo MW, Sohn CH, Shin SY, Lee J, Yu M, Kim WY, Lim KS, Lee S. Cumulative Query Method for Influenza Surveillance Using Search Engine Data. *Journal of Medical Internet Research*, **2014**;16(12):e289.
41. Woo H, Cho Y, Shim E, Lee JK, Lee CG, Kim SH. Estimating Influenza Outbreaks Using Both Search Engine Query Data and Social Media Data in South Korea. *Journal of Medical Internet Research*, **2016**;18(7):e177.
42. Leung R, Guo H, Pan X. Social Media Users' Perception of Telemedicine and mHealth in China: Exploratory Study. *JMIR mHealth and uHealth,* **2018**;6(9):e181.
43. Li Q, Wang C, Liu R, Wang L, Zeng DD, Leischow SJ. Understanding Users' Vaping Experiences from Social Media: Initial Study Using Sentiment Opinion Summarization Techniques. *Journal of Medical Internet Research*, **2018**;20(8):e252.
44. Sadah SA, Shahbazi M, Wiley MT, Hristidis V. A Study of the Demographics of Web-Based Health-Related Social Media Users. *Journal of Medical Internet Research*, **2015**;17(8):e194.
45. Sadah SA, Shahbazi M, Wiley MT, Hristidis V. Demographic-Based Content Analysis of Web-Based Health-Related Social Media. *Journal of Medical Internet Research*, **2016**;18(6):e148.

46. Abbe A, Falissard B. Stopping Antidepressants and Anxiolytics as Major Concerns Reported in Online Health Communities: A Text Mining Approach. *JMIR Mental Health*, **2017**;4(4):e48.
47. Abdellaoui R, Schück S, Texier N, Burgun A. Filtering Entities to Optimize Identification of Adverse Drug Reaction From Social Media: How Can the Number of Words Between Entities in the Messages Help? *JMIR Public Health and Surveillance,* **2017**;3(2):e36.
48. Abdellaoui R, Foulquia P, Texier N, Faviez C, Burgun A, Schack S. Detection of Cases of Noncompliance to Drug Treatment in Patient Forum Posts: Topic Model Approach. *Journal of Medical Internet Research*, **2018**;20(3):e85.
49. Dyson MP, Newton AS, Shave K, Featherstone RM, Thomson D, Wingert A, Fernandes RM, Hartling L. Social Media for the Dissemination of Cochrane Child Health Evidence: Evaluation Study. *Journal of Medical Internet Research*, **2017**;19(9):e308.
50. Jung Y, Hur C, Jung D, Kim M. Identifying Key Hospital Service Quality Factors in Online Health Communities. *Journal of Medical Internet Research*, **2015**;17(4):e90.
51. Koh S, Gordon AS, Wienberg C, Sood SO, Morley S, Burke DM. Stroke Experiences in Weblogs: A Feasibility Study of Sex Differences. *Journal of Medical Internet Research*, **2014**;16(3):e84.
52. Konheim-Kalkstein YL, Miron-Shatz T, Israel LJ. How Women Evaluate Birth Challenges: Analysis of Web-Based Birth Stories. *JMIR Pediatrics and Parenting*, **2018**;1(2):e12206.
53. Tinschert P, Jakob R, Barata F, Kramer JN, Kowatsch T. The Potential of Mobile Apps for Improving Asthma Self-Management: A Review of Publicly Available and Well-Adopted Asthma Apps. *JMIR mHealth and uHealth*, **2017**;5(8):e113.
54. Athilingam P, Jenkins B. Mobile Phone Apps to Support Heart Failure Self-Care Management: Integrative Review. *JMIR Cardio,* **2018**;2(1):e10057.
55. Hendriks H, Van den Putte B, Gebhardt WA, Moreno MA. Social Drinking on Social Media: Content Analysis of the Social Aspects of Alcohol-Related Posts on Facebook and Instagram. *Journal of Medical Internet Research*, 2018;20(6):e226.
56. Roccetti M, Marfia G, Salomoni P, Prandi C, Zagari RM, Gningaye Kengni FL, Bazzoli F, Montagnani M. Attitudes of Crohn's Disease Patients: Infodemiology Case Study and Sentiment Analysis of Facebook and Twitter Posts. *JMIR Public Health and Surveillance,* **2017**;3(3):e51.

57. Manchaiah V, Ratinaud P, Andersson G. Representation of Tinnitus in the US Newspaper Media and in Facebook Pages: Cross-Sectional Analysis of Secondary Data. *Interactive Journal of Medical Research*, 2018;7(1):e9.
58. Sciascia S, Radin M. What can Google and Wikipedia can tell us about a disease? Big Data trends analysis in Systemic Lupus Erythematosus. *International Journal of Medical Informatics*, **2017**;107:65-69.
59. Sharpe JD, Hopkins RS, Cook RL, Striley CW. Evaluating Google, Twitter, and Wikipedia as Tools for Influenza Surveillance Using Bayesian Change Point Analysis: A Comparative Analysis. *JMIR Public Health and Surveillance,* **2016**;2(2):e161.
60. Timpka T, Spreco A, Dahlström Ö, Eriksson O, Gursky E, Ekberg J, Blomqvist E, Strömgren M, Karlsson D, Eriksson H, Nyce J, Hinkula J, Holm E. Performance of eHealth Data Sources in Local Influenza Surveillance: A 5-Year Open Cohort Study. *Journal of Medical Internet Research*, **2014**;16(4):e116.
61. Wagner M, Lampos V, Yom-Tov E, Pebody R, Cox IJ. Estimating the Population Impact of a New Pediatric Influenza Vaccination Program in England Using Social Media Content. *Journal of Medical Internet Research*, **2017**;19(12):e416.
62. Aslam AA, Tsou MH, Spitzberg BH, An L, Gawron JM, Gupta DK, Peddecord KM, Nagel AC, Allen C, Yang JA, Lindsay S. The reliability of tweets as a supplementary method of seasonal influenza surveillance. *Journal of Medical Internet Research*, **2014**;16(11):e250.
63. Baltrusaitis K, Santillana M, Crawley AW, Chunara R, Smolinski M, Brownstein JS. Determinants of Participants' Follow-Up and Characterization of Representativeness in Flu Near You, A Participatory Disease Surveillance System. *JMIR Public Health and Surveillance*, **2017**;3(2):e18.
64. Broniatowski DA, Dredze M, Paul MJ, Dugas A. Using Social Media to Perform Local Influenza Surveillance in an Inner-City Hospital: A Retrospective Observational Study. *JMIR Public Health and Surveillance*, **2015**;1(1):e5.
65. Chen B, Shao J, Liu K, Cai G, Jiang Z, Huang Y, Gu H, Jiang J. Does Eating Chicken Feet With Pickled Peppers Cause Avian Influenza? Observational Case Study on Chinese Social Media During the Avian Influenza A (H7N9) Outbreak. *JMIR Public Health and Surveillanc*e, **2018**;4(1):e32.
66. Gu H, Chen B, Zhu H, Jiang T, Wang X, Chen L, Jiang Z, Zheng D, Jiang J. Importance of Internet Surveillance in Public Health Emergency Control and

Prevention: Evidence From a Digital Epidemiologic Study During Avian Influenza A H7N9 Outbreaks. *Journal of Medical Internet Research*, **2014**;16(1):e20.

67. Hill S, Mao J, Ungar L, Hennessy S, Leonard CE, Holmes J. Natural supplements for H1N1 influenza: retrospective observational infodemiology study of information and search activity on the Internet. *Journal of Medical Internet Research*, **2011**;13(2):e36.
68. Kagashe I, Yan Z, Suheryani I. Enhancing Seasonal Influenza Surveillance: Topic Analysis of Widely Used Medicinal Drugs Using Twitter Data. *Journal of Medical Internet Research*, **2017**;19(9):e315. PMID: 28899847.
69. Kandula S, Hsu D, Shaman J. Subregional Nowcasts of Seasonal Influenza Using Search Trends. *Journal of Medical Internet Research*, **2017**;19(11):e370.
70. Klembczyk JJ, Jalalpour M, Levin S, Washington RE, Pines JM, Rothman RE, Dugas AF. Google Flu Trends Spatial Variability Validated Against Emergency Department Influenza-Related Visits. *Journal of Medical Internet Research*, **2016**;18(6):e175.
71. Lu FS, Hou S, Baltrusaitis K, Shah M, Leskovec J, Sosic R, Hawkins J, Brownstein J, Conidi G, Gunn J, Gray J, Zink A, Santillana M. Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis. *JMIR Public Health and Surveillance*, **2018**;4(1):e4.
72. Mao C, Wu XY, Fu XH, Di MY, Yu YY, Yuan JQ, Yang ZY, Tang JL. An Internet-Based Epidemiological Investigation of the Outbreak of H7N9 Avian Influenza A in China Since Early 2013. *Journal of Medical Internet Research*, **2014**;16(9):e221.
73. Pervaiz F, Pervaiz M, Abdur Rehman N, Saif U. FluBreaks: Early Epidemic Detection from Google Flu Trends. *Journal of Medical Internet Research*, **2012**;14(5):e125.
74. Poirier C, Lavenu A, Bertaud V, Campillo-Gimenez B, Chazard E, Cuggia M, BouzillÃ G. Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine Learning Methods: Comparison Study. *JMIR Public Health and Surveillance*, **2018**;4(4):e11361.
75. Samaras L, García-Barriocanal E, Sicilia MA. Syndromic Surveillance Models Using Web Data: The Case of Influenza in Greece and Italy Using Google Trends. *JMIR Public Health and Surveillance,* **2017**;3(4):e90.
76. Santos JC, Matos S. Analysing Twitter and web queries for flu trend prediction. *Theoretical Biology and Medical Modelling*, 2014;11(Suppl1):S6.

77. Adrover C, Bodnar T, Huang Z, Telenti A, Salathé M. Identifying Adverse Effects of HIV Drug Treatment and Associated Sentiments Using Twitter. *JMIR Public Health and Surveillance,* **2015**;1(2):e7.
78. Zheluk A, Quinn C, Hercz D, Gillespie JA. Internet search patterns of human immunodeficiency virus and the digital divide in the Russian Federation: infoveillance study. *Journal of Medical Internet Research*, 2013;15(11):e256.
79. Mavragani A, Ochoa G (2018) Forecasting AIDS Prevalence in the United States using Online Search Traffic Data. *Journal of Big Data*, 5:17.
80. Mollema L, Harmsen IA, Broekhuizen E, Clijnk R, De Melker H, Paulussen T, Kok G, Ruiter R, Das E. Disease Detection or Public Opinion Reflection? Content Analysis of Tweets, Other Social Media, and Online Newspapers During the Measles Outbreak in the Netherlands in 2013. *Journal of Medical Internet Research*, **2015**;17(5):e128.
81. Radzikowski J, Stefanidis A, Jacobsen KH, Croitoru A, Crooks A, Delamater PL. The Measles Vaccination Narrative in Twitter: A Quantitative Analysis. *JMIR Public Health and Surveillance*, **2016**;2(1):e1.
82. Du J, Tang L, Xiang Y, Zhi D, Xu J, Song HY, Tao C. Public Perception Analysis of Tweets During the 2015 Measles Outbreak: Comparative Study Using Convolutional Neural Network Models. *Journal of Medical Internet Research*, 2018;20(7):e236.
83. Mavragani A, Ochoa G. The Internet and the Anti-Vaccine Movement: Tracking the 2017 EU Measles Outbreak. *Big Data and Cognitive Computing*, **2018**;2(1).
84. Farhadloo M, Winneg K, Chan MPS, Hall Jamieson K, Albarracin D. Associations of Topics of Discussion on Twitter With Survey Measures of Attitudes, Knowledge, and Behaviors Related to Zika: Probabilistic Study in the United States. *JMIR Public Health and Surveillance*, 2018;4(1):e16.
85. Majumder MS, Santillana M, Mekaru SR, McGinnis DP, Khan K, Brownstein JS. Utilizing Nontraditional Data Sources for Near Real-Time Estimation of Transmission Dynamics During the 2015-2016 Colombian Zika Virus Disease Outbreak. *JMIR Public Health and Surveillance*, 2016;2(1):e30.
86. Miller M, Banerjee T, Muppalla R, Romine W, Sheth A. What Are People Tweeting About Zika? An Exploratory Study Concerning Its Symptoms, Treatment, Transmission, and Prevention. *JMIR Public Health and Surveillance*, **2017**;3(2):e38.

87. Stefanidis A, Vraga E, Lamprianidis G, Radzikowski J, Delamater PL, Jacobsen KH, Pfoser D, Croitoru A, Crooks A. Zika in Twitter: Temporal Variations of Locations, Actors, and Concepts. *JMIR Public Health and Surveillance*, **2017**;3(2):e22.
88. Sanz-Lorente M, Wanden-Berghe C, Castejón-Bolea R, Sanz-Valero J. Web 2.0 Tools in the Prevention of Curable Sexually Transmitted Diseases: Scoping Review. *Journal of Medical Internet Research*, **2018**;20(3):e113.
89. Wongkoblap A, Vadillo MA, Curcin V. Researching Mental Health Disorders in the Era of Social Media: Systematic Review. *Journal of Medical Internet Research*, **2017**;19(6):e228.
90. Athilingam P, Jenkins B. Mobile Phone Apps to Support Heart Failure Self-Care Management: Integrative Review. *JMIR Cardio*, **2018**;2(1):e10057.
91. Gohil S, Vuik S, Darzi A. Sentiment Analysis of Health Care Tweets: Review of the Methods Used. *JMIR Public Health and Surveillance*, **2018**;4(2):e43.
92. Mavragani A, Ochoa G, Tsagarakis KP. Assessing the Methods, Tools, and Statistical Approaches in Google Trends Research: Systematic Review. *Journal of Medical Internet Research*, **2018**;20(11):e270.
93. Hanson CL, Cannon B, Burton S, Giraud-Carrier C. An Exploration of Social Circles and Prescription Drug Abuse Through Twitter. *Journal of Medical Internet Research*, **2013**;15(9):e189.
94. Katsuki T, Mackey TK, Cuomo R. Establishing a Link Between Prescription Drug Abuse and Illicit Online Pharmacies: Analysis of Twitter Data. *Journal of Medical Internet Research*, **2015**;17(12):e280.
95. Peiper NC, Baumgartner PM, Chew RF, Hsieh YP, Bieler GS, Bobashev GV, Siege C, Zarkin GA. Patterns of Twitter Behavior Among Networks of Cannabis Dispensaries in California. *Journal of Medical Internet Research*, **2017**;19(7):e236.
96. Yom-Tov E, Lev-Ran S. Adverse Reactions Associated With Cannabis Consumption as Evident From Search Engine Queries. *JMIR Public Health and Surveillance*, **2017**;3(4):e77.
97. Cavazos-Rehg P, Krauss M, Grucza R, Bierut L. Characterizing the Followers and Tweets of a Marijuana-Focused Twitter Handle. *Journal of Medical Internet Research*, **2014**;16(6):e157.
98. Ricard BJ, Marsch LA, Crosier B, Hassanpour S. Exploring the Utility of Community-Generated Social Media Content for Detecting Depression: An

Analytical Study on Instagram. *Journal of Medical Internet Research,* **2018**;20(12):e11817.

99. Seabrook EM, Kern ML, Fulcher BD, Rickard NS. Predicting Depression From Language-Based Emotion Dynamics: Longitudinal Analysis of Facebook and Twitter Status Updates. *Journal of Medical Internet Research,* **2018**;20(5):e168.
100. Tana JC, Kettunen J, Eirola E, Paakkonen H. Diurnal Variations of Depression-Related Health Information Seeking: Case Study in Finland Using Google Trends Data. *JMIR Mental Health*, **2018**;5(2):e43.
101. DeJohn AD, Schulz EE, Pearson AL, Lachmar EM, Wittenborn AK. Identifying and Understanding Communities Using Twitter to Connect About Depression: Cross-Sectional Study. *JMIR Mental Health*, **2018**;5(4):e61.
102. Jung H, Park HA, Song TM. Ontology-Based Approach to Social Data Sentiment Analysis: Detection of Adolescent Depression Signals. *Journal of Medical Internet Research,* **2017**;19(7):e259.
103. Lachmar EM, Wittenborn AK, Bogen KW, McCauley HL. #MyDepressionLooksLike: Examining Public Discourse About Depression on Twitter. *JMIR Mental Health*, **2017**;4(4):e43.
104. Schlichthorst M, King K, Turnure J, Sukunesan S, Phelps A, Pirkis J. Influencing the Conversation About Masculinity and Suicide: Evaluation of the Man Up Multimedia Campaign Using Twitter Data. *JMIR Mental Health*, **2018**;5(1):e14.
105. Wong PWC, Fu KW, Yau RSP, Ma HHM, Law YW, Chang SS, Yip PSF. Accessing Suicide-Related Information on the Internet: A Retrospective Observational Study of Search Behavior. *Journal of Medical Internet Research*, **2013**;15(1):e3.
106. Braithwaite SR, Giraud-Carrier C, West J, Barnes MD, Hanson CL. Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality. *JMIR Mental Health*, **2016**;3(2):e21.
107. Cheng Q, Li TM, Kwok CL, Zhu T, Yip PS. Assessing Suicide Risk and Emotional Distress in Chinese Social Media: A Text Mining and Machine Learning Study. *Journal of Medical Internet Research*, **2017**;19(7):e243.
108. Lee D, Lee H, Choi M. Examining the Relationship Between Past Orientation and US Suicide Rates: An Analysis Using Big Data-Driven Google Search Queries. *Journal of Medical Internet Research*, **2016**;18(2):e35.

109. Rose SW, Jo CL, Binns S, Buenger M, Emery S, Ribisl KM. Perceptions of Menthol Cigarettes Among Twitter Users: Content and Sentiment Analysis. *Journal of Medical Internet Research*, **2017**;19(2):e56.
110. Ayers JW, Westmaas JL, Leas EC, Benton A, Chen Y, Dredze M, Althouse BM. Leveraging Big Data to Improve Health Awareness Campaigns: A Novel Evaluation of the Great American Smokeout. *JMIR Public Health and Surveillance*, **2016**;2(1):e16.
111. de Viron S, Suggs LS, Brand A, Van Oyen H. Communicating Genetics and Smoking Through Social Media: Are We There Yet? *Journal of Medical Internet Research*, **2013**;15(9):e198.
112. Duke JC, Hansen H, Kim AE, Curry L, Allen J. The Use of Social Media by State Tobacco Control Programs to Promote Smoking Cessation: A Cross-Sectional Study. *Journal of Medical Internet Research*, **2014**;16(7):e169.
113. Myslín M, Zhu SH, Chapman W, Conway M. Using Twitter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products. *Journal of Medical Internet Research*, **2013**;15(8):e174.
114. Rocheleau M, Sadasivam RS, Baquis K, Stahl H, Kinney RL, Pagoto SL, Houston TK. An Observational Study of Social and Emotional Support in Smoking Cessation Twitter Accounts: Content Analysis of Tweets. *Journal of Medical Internet Research*, **2015**;17(1):e18.
115. Lienemann BA, Unger JB, Cruz TB, Chu KH. Methods for Coding Tobacco-Related Twitter Data: A Systematic Review. *Journal of Medical Internet Research*, **2017**;19(3):e91.
116. Staal YC, van de Nobelen S, Havermans A, Talhout R. New Tobacco and Tobacco-Related Products: Early Detection of Product Development, Marketing Strategies, and Consumer Interest. *JMIR Public Health and Surveillance*, **2018**;4(2):e55.
117. Zhan Y, Liu R, Li Q, Leischow SJ, Zeng DD. Identifying Topics for E-Cigarette User-Generated Contents: A Case Study from Multiple Social Media Platforms. *Journal of Medical Internet Research*, **2017**;19(1):e24.
118. Allem JP, Ferrara E, Uppu SP, Cruz TB, Unger JB. E-Cigarette Surveillance with Social Media Data: Social Bots, Emerging Topics, and Trends. *JMIR Public Health and Surveillance*, **2017**;3(4):e98.

119. Chen AT, Zhu SH, Conway M. What Online Communities Can Tell Us About Electronic Cigarettes and Hookah Use: A Study Using Text Mining and Visualization Techniques. *Journal of Medical Internet Research*, **2015**;17(9):e220.
120. Chu KH, Sidhu AK, Valente TW. Electronic Cigarette Marketing Online: a Multi-Site, Multi-Product Comparison. *JMIR Public Health and Surveillance*, **2015**;1(2):e11.
121. Cole-Lewis H, Pugatch J, Sanders A, Varghese A, Posada S, Yun C, Schwarz M, Augustson E. Social Listening: A Content Analysis of E-Cigarette Discussions on Twitter. *Journal of Medical Internet Research*, **2015**;17(10):e243.
122. Cole-Lewis H, Varghese A, Sanders A, Schwarz M, Pugatch J, Augustson E. Assessing Electronic Cigarette-Related Tweets for Sentiment and Content Using Supervised Machine Learning. *Journal of Medical Internet Research*, **2015**;17(8):e208.
123. Harris JK, Moreland-Russell S, Choucair B, Mansour R, Staub M, Simmons K. Tweeting for and Against Public Health Policy: Response to the Chicago Department of Public Health's Electronic Cigarette Twitter Campaign. *Journal of Medical Internet Research*, **2014**;16(10):e238.
124. Kim AE, Hopper T, Simpson S, Nonnemaker J, Lieberman AJ, Hansen H, Guillory J, Porter L. Using Twitter Data to Gain Insights into E-cigarette Marketing and Locations of Use: An Infoveillance Study. *Journal of Medical Internet Research*, **2015**;17(11):e251.
125. Kim A, Miano T, Chew R, Eggers M, Nonnemaker J. Classification of Twitter Users Who Tweet About E-Cigarettes. *JMIR Public Health and Surveillance*, **2017**;3(3):e63.
126. Lazard AJ, Saffer AJ, Wilcox GB, Chung AD, Mackert MS, Bernhardt JM. E-Cigarette Social Media Messages: A Text Mining Analysis of Marketing and Consumer Conversations on Twitter. *JMIR Public Health and Surveillance*, **2016**;2(2):e171.
127. Allem JP, Ramanujam J, Lerman K, Chu KH, Boley Cruz T, Unger JB. Identifying Sentiment of Hookah-Related Posts on Twitter. *JMIR Public Health and Surveillance*, **2017**;3(4):e74.
128. Cawkwell PB, Lee L, Weitzman M, Sherman SE. Tracking Hookah Bars in New York: Utilizing Yelp as a Powerful Public Health Tool. *JMIR Public Health and Surveillance*, **2015**;1(2):e19.

129. Zhang Y, Allem JP, Unger JB, Boley Cruz T. Automated Identification of Hookahs (Waterpipes) on Instagram: An Application in Feature Extraction Using Convolutional Neural Network and Support Vector Machine Classification. *Journal of Medical Internet Research*, **2018**;20(11):e10513.
130. Allem JP, Dharmapuri L, Leventhal AM, Unger JB, Boley Cruz T. Hookah-Related Posts to Twitter From 2017 to 2018: Thematic Analysis. *Journal of Medical Internet Research*, **2018**;20(11):e11669.
131. Liu Y, Mei Q, Hanauer DA, Zheng K, Lee JM. Use of Social Media in the Diabetes Community: An Exploratory Analysis of Diabetes-Related Tweets. *JMIR Diabetes*, **2016**;1(2):e4.
132. Martinez M, Park SB, Maison I, Mody V, Soh LS, Parihar HS. iOS Appstore-Based Phone Apps for Diabetes Management: Potential for Use in Medication Adherence. *JMIR Diabetes*, **2017**;2(2):e12.
133. Oser TK, Oser SM, McGinley EL, Stuckey HL. A Novel Approach to Identifying Barriers and Facilitators in Raising a Child With Type 1 Diabetes: Qualitative Analysis of Caregiver Blogs. *JMIR Diabetes*, **2017**;2(2):e27.
134. Sinnenberg L, Mancheno C, Barg FK, Asch DA, Rivard CL, Horst-Martz E, Buttenheim A, Ungar L, Merchant R. Content Analysis of Metaphors About Hypertension and Diabetes on Twitter: Exploratory Mixed-Methods Study. *JMIR Diabetes*, **2018**;3(4):e11177.
135. Xu X, Litchman ML, Gee PM, Whatcott W, Chacon L, Holmes J, Srinivasan SS. Predicting Prediabetes Through Facebook Postings: Protocol for a Mixed-Methods Study. *JMIR Research Protocols*, **2018**;7(12):e10720.
136. Arnhold M, Quade M, Kirch W. Mobile Applications for Diabetics: A Systematic Review and Expert-Based Usability Evaluation Considering the Special Requirements of Diabetes Patients Age 50 Years or Older. *Journal of Medical Internet Research*, **2014**;16(4):e104.
137. Koschack J, Weibezahl L, Friede T, Himmel W, Makedonski P, Grabowski J. Scientific Versus Experiential Evidence: Discourse Analysis of the Chronic Cerebrospinal Venous Insufficiency Debate in a Multiple Sclerosis Forum. *Journal of Medical Internet Research*, **2015**;17(7):e159.
138. Risson V, Saini D, Bonzani I, Huisman A, Olson M. Patterns of Treatment Switching in Multiple Sclerosis Therapies in US Patients Active on Social Media: Application

of Social Media Content Analysis to Health Outcomes Research. *Journal of Medical Internet Research*, **2016**;18(3):e62.

139. Tapi Nzali MD, Bringay S, Lavergne C, Mollevi C, Opitz T. What Patients Can Tell Us: Topic Analysis for Social Media on Breast Cancer. *JMIR Medical Informatics*, **2017**;5(3):e23.
140. Vasconcellos-Silva PR, Carvalho DBF, Trajano V, de La Rocque LR, Sawada ACMB, Juvanhol LL. Using Google Trends Data to Study Public Interest in Breast Cancer Screening in Brazil: Why Not a Pink February? *JMIR Public Health and Surveillance*, **2017**;3(2):e17.
141. Huesch M, Chetlen A, Segel J, Schetter S. Frequencies of Private Mentions and Sharing of Mammography and Breast Cancer Terms on Facebook: A Pilot Study. *Journal of Medical Internet Research*, **2017**;19(6):e201.
142. Jones J, Pradhan M, Hosseini M, Kulanthaivel A, Hosseini M. Novel Approach to Cluster Patient-Generated Data Into Actionable Topics: Case Study of a Web-Based Breast Cancer Forum. *JMIR Medical Informatics*, **2018**;6(4):e45.
143. Vickey T, Breslin JG. Online Influence and Sentiment of Fitness Tweets: Analysis of Two Million Fitness Tweets. *JMIR Public Health and Surveillance*, **2017**;3(4):e82.
144. Edney S, Bogomolova S, Ryan J, Olds T, Sanders I, Maher C. Creating Engaging Health Promotion Campaigns on Social Media: Observations and Lessons from Fitbit and Garmin. *Journal of Medical Internet Research,* **2018**;20(12):e10911.
145. Madden KM. The Seasonal Periodicity of Healthy Contemplations About Exercise and Weight Loss: Ecological Correlational Study. *JMIR Public Health and Surveillance*, **2017**;3(4):e92.
146. Wang HW, Chen DR. Economic Recession and Obesity-Related Internet Search Behavior in Taiwan: Analysis of Google Trends Data. *JMIR Public Health and Surveillance*, **2018**;4(2):e37.
147. Sugawara Y, Narimatsu H, Tsuya A, Tanaka A, Fukao A. Medical Institutions and Twitter: A Novel Tool for Public Communication in Japan. *JMIR Public Health and Surveillance*, **2016**;2(1):e19.
148. Xu S, Markson C, Costello KL, Xing CY, Demissie K, Llanos AA. Leveraging Social Media to Promote Public Health Knowledge: Example of Cancer Awareness via Twitter *JMIR Public Health and Surveillance*, **2016**;2(1):e17.

149. Surian D, Nguyen DQ, Kennedy G, Johnson M, Coiera E, Dunn AG. Characterizing Twitter Discussions About HPV Vaccines Using Topic Modeling and Community Detection. *Journal of Medical Internet Research,* **2016**;18(8):e232.
150. Dunn AG, Leask J, Zhou X, Mandl KD, Coiera E. Associations Between Exposure to and Expression of Negative Opinions About Human Papillomavirus Vaccines on Social Media: An Observational Study. *Journal of Medical Internet Research,* **2015**;17(6):e144.
151. Lama Y, Chen T, Dredze M, Jamison A, Quinn SC, Broniatowski DA. Discordance Between Human Papillomavirus Twitter Images and Disparities in Human Papillomavirus Risk and Disease in the United States: Mixed-Methods Analysis. *Journal of Medical Internet Research,* **2018**;20(9):e10244.
152. Mahoney LM, Tang T, Ji K, Ulrich-Schad J. The Digital Distribution of Public Health News Surrounding the Human Papillomavirus Vaccination: A Longitudinal Infodemiology Study. *JMIR Public Health and Surveillance*, **2015**;1(1):e2.
153. Massey PM, Leader A, Yom-Tov E, Budenz A, Fisher K, Klassen AC. Applying Multiple Data Collection Tools to Quantify Human Papillomavirus Vaccine Communication on Twitter. *Journal of Medical Internet Research,* **2016**;18(12):e318
154. Nakada H, Yuji K, Tsubokura M, Ohsawa Y, Kami M. Development of a national agreement on human papillomavirus vaccination in Japan: an infodemiology study. *Journal of Medical Internet Research,* **2014**;16(5):e129.
155. Sato A, Aramaki E, Shimamoto Y, Tanaka S, Kawakami K. Blog Posting After Lung Cancer Notification: Content Analysis of Blogs Written by Patients or Their Families. *JMIR Cancer*, **2015**;1(1):e5.
156. Brigo F, Lattanzi S, Bragazzi N, Nardone R, Moccia M, Lavorgna L. Why do people search Wikipedia for information on multiple sclerosis? *Multiple Sclerosis and Related Disorders*, **2018**;20:210-214.
157. Brigo F, Igwe SC, Ausserer H, Nardone R, Tezzon F, Bongiovanni LG, Trinka E. Why do people Google epilepsy? An infodemiological study of online behavior for epilepsy related search terms. *Epilepsy and Behavior*, **2014**;31:67-70.
158. Brigo F, Lattanzi S, Giussani G, Tassi L, Pietrafusa N, Galimberti CA, Nardone R, Bragazzi NL, Mecarelli O. Italian Wikipedia and epilepsy: An infodemiological study of online information-seeking behavior. *Epilepsy and Behavior*, **2018**;81:119-122.

159. Kim SJ, Marsch LA, Hancock JT, Das AK. Scaling Up Research on Drug Abuse and Addiction Through Social Media Big Data. *Journal of Medical Internet Research,* **2017**;19(10):e353.
160. Cherian R, Westbrook M, Ramo D, Sarkar U. Representations of Codeine Misuse on Instagram: Content Analysis. *JMIR Public Health and Surveillance*, **2018**;4(1):e22.
161. Bollegala D, Maskell S, Sloane R, Hajne J, Pirmohamed M. Causality Patterns for Detecting Adverse Drug Reactions from Social Media: Text Mining Approach *JMIR Public Health and Surveillance*, **2018**;4(2):e51.
162. Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N, Jaulent MC, Beyens MN, Burgun A, Bousquet C. Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review. *Journal of Medical Internet Research,* **2015**;17(7):e171.
163. Mackey TK, Liang BA. Global Reach of Direct-to-Consumer Advertising Using Social Media for Illicit Online Drug Sales. *Journal of Medical Internet Research*, **2013**;15(5):e105.
164. Tyrawski J, DeAndrea DC. Pharmaceutical Companies and Their Drugs on Social Media: A Content Analysis of Drug Information on Popular Social Media Sites. *Journal of Medical Internet Research*, **2015**;17(6):e130.
165. Espina K., Estuar Ma.R.J.E. Infodemiology for Syndromic Surveillance of Dengue and Typhoid Fever in the Philippines. *Procedia Computer Science*, **2017**;121:554-561.
166. Livelo E.D., Cheng C. Intelligent dengue infoveillance using gated recurrent neural learning and cross-label frequencies. *Proceedings-2018 IEEE International Conference on Agents*, ICA **2018**, 8459963:2-7.
167. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PLoS One*, ***2010***;5(11):e14118.
168. Liu K, Huang S, Miao ZP, Chen B, Jiang T, Cai G, Jiang Z, Chen Y, Wang Z, Gu H, Chai C, Jiang J. Identifying Potential Norovirus Epidemics in China via Internet Surveillance. *Journal of Medical Internet Research*, **2017**;19(8):e282.
169. Domnich A., Arbuzova E.K., Signori A., Amicizia D., Panatto D., Gasparini R. Demand-based web surveillance of sexually transmitted infections in Russia. *International Journal of Public Health*, **2014**;59(5):841-849.

170. Gabarron E, Serrano JA, Wynn R, Lau AY. Tweet Content Related to Sexually Transmitted Diseases: No Joking Matter. *Journal of Medical Internet Research*, **2014**;16(10):e228.
171. Birnbaum ML, Ernala SK, Rizvi AF, De Choudhury M, Kane JM. A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals. *Journal of Medical Internet Research*, **2017**;19(8):e289.
172. Liu S, Zhu M, Yu DJ, Rasin A, Young SD. Using Real-Time Social Media Technologies to Monitor Levels of Perceived Stress and Emotional State in College Students: A Web-Based Questionnaire Study. *JMIR Mental Health*, **2017**;4(1):e2.
173. Christmann CA, Hoffmann A, Bleser G. Stress Management Apps With Regard to Emotion-Focused Coping and Behavior Change Techniques: A Content Analysis. *JMIR mHealth and uHealth*, **2017**;5(2):e22.
174. Doan S, Ritchart A, Perry N, Chaparro JD, Conway M. How Do You #relax When You're #stressed? A Content Analysis and Infodemiology Study of Stress-Related Tweets. *JMIR Public Health and Surveillance*, **2017**;3(2):e35.
175. Yagahara A, Hanai K, Hasegawa S, Ogasawara K. Relationships Among Tweets Related to Radiation: Visualization Using Co-Occurring Networks. *JMIR Public Health and Surveillance*, **2018**;4(1):e26.
176. Nishimoto N, Ota M, Yagahara A, Ogasawara K. Estimating the Duration of Public Concern After the Fukushima Dai-ichi Nuclear Power Station Accident from the Occurrence of Radiation Exposure-Related Terms on Twitter: A Retrospective Data Analysis. *JMIR Public Health Surveillance*, **2016**;2(2):e168.
177. Davis MA, Zheng K, Liu Y, Levy H. Public Response to Obamacare on Twitter. *Journal of Medical Internet Research*, **2017**;19(5):e167.
178. Wong CA, Sap M, Schwartz A, Town R, Baker T, Ungar L, Merchant RM. Twitter Sentiment Predicts Affordable Care Act Marketplace Enrollment. *Journal of Medical Internet Research*, **2015**;17(2):e51.
179. Pretorius KA, Mackert M, Wilcox GB. Sudden Infant Death Syndrome and Safe Sleep on Twitter: Analysis of Influences and Themes to Guide Health Promotion Efforts. *JMIR Pediatrics and Parenting*, **2018**;1(2):e10435.
180. Meaney S, Cussen L, Greene RA, O'Donoghue K. Reaction on Twitter to a Cluster of Perinatal Deaths: A Mixed Method Study. *JMIR Public Health and Surveillance*, **2016**;2(2):e36.

181. Nakhasi A, Shen AX, Passarella RJ, Appel LJ, Anderson CA. Online Social Networks That Connect Users to Physical Activity Partners: A Review and Descriptive Analysis. *Journal of Medical Internet Research*, **2014**;16(6):e153.
182. Nguyen QC, Li D, Meng HW, Kath S, Nsoesie E, Li F, Wen M. Building a National Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity. *JMIR Public Health and Surveillance*, **2016**;2(2):e158.
183. Zhang N, Campo S, Janz KF, Eckler P, Yang J, Snetselaar LG, Signorini A. Electronic word of mouth on twitter about physical activity in the United States: exploratory infodemiology study. *Journal of Medical Internet Research*, **2013**;15(11):e261.
184. Adawi M, Bragazzi NL, Watad A, Sharif K, Amital H, Mahroum N. Discrepancies Between Classic and Digital Epidemiology in Searching for the Mayaro Virus: Preliminary Qualitative and Quantitative Analysis of Google Trends. *JMIR Public Health and Surveillance*, **2017**;3(4):e93.
185. Anderson LS, Bell HG, Gilbert M, Davidson JE, Winter C, Barratt MJ, Win B, Painter JL, Menone C, Sayegh J, Dasgupta N. Using Social Listening Data to Monitor Misuse and Nonmedical Use of Bupropion: A Content Analysis. *JMIR Public Health and Surveillance*, **2017**;3(1):e6.
186. Balls-Berry J, Sinicrope P, Valdez Soto M, Brockman T, Bock M, Patten C. Linking Podcasts With Social Media to Promote Community Health and Medical Research: Feasibility Study. *JMIR Formative Research*, **2018**;2(2):e10025.
187. Bousquet C, Dahamna B, Guillemin-Lanne S, Darmoni SJ, Faviez C, Huot C, Katsahian S, Leroux V, Pereira S, Richard C, Schück S, Souvignet J, Lillo-Le Louët A, Texier N. The Adverse Drug Reactions from Patient Reports in Social Media Project: Five Major Challenges to Overcome to Operationalize Analysis and Efficiently Support Pharmacovigilance Process. *JMIR Research Protocols*, **2017**;6(9):e179.
188. Carrotte ER, Prichard I, Lim MSC. "Fitspiration" on Social Media: A Content Analysis of Gendered Images. *Journal of Medical Internet Research*, **2017**;19(3):e95.
189. Colditz JB, Chu KH, Emery SL, Larkin CR, James AE, Welling J, Primack BA. Toward Real-Time Infoveillance of Twitter Health Messages. *The American Journal of Public Health*, **2018**;108(8):1009-1014.
190. García-Díaz J.A., Apolinario-Arzube O., Medina-Moreira J., Salavarria-Melo J.O., Lagos-Ortiz K., Luna-Aveiga H., Valencia-García R. Opinion mining for measuring

the social perception of infectious diseases. An infodemiology approach. *Communications in Computer and Information Science*, **2018**;883:229-239.

191. Genes N, Chary M, Chason K. Analysis of Twitter Users' Sharing of Official New York Storm Response Messages. *Medicine 2.0*, **2014**;3(1):e1.
192. Gianfredi V, Bragazzi NL, Mahamid M, Bisharat B, Mahroum N, Amital H, Adawi M. Monitoring public interest toward pertussis outbreaks: an extensive Google Trends-based analysis. *Public Health*, **2018**;165:9-15.
193. Jung H, Park HA, Song TM. Ontology-Based Approach to Social Data Sentiment Analysis: Detection of Adolescent Depression Signals. *Journal of Medical Internet Research*, **2017**;19(7):e259.
194. Kamiński, M.; Łoniewski, I.; Misera, A.; Marlicz, W. Heartburn-Related Internet Searches and Trends of Interest across Six Western Countries: A Four-Year Retrospective Analysis Using Google Ads Keyword Planner. *International Journal of Environmental Research and Public Health,* **2019**, *16*, 4591.
195. Kürzinger ML, Schück S, Texier N, Abdellaoui R, Faviez C, Pouget J, Zhang L, Tcherny-Lessenot S, Lin S, Juhaeri J. Web-Based Signal Detection Using Medical Forums Data in France: Comparative Analysis. *Journal of Medical Internet Research*, **2018**;20(11):e10466.
196. Madden KM, Feldman B. Weekly, Seasonal, and Geographic Patterns in Health Contemplations About Sundown Syndrome: An Ecological Correlational Study. *JMIR Aging*, **2019**;2(1):e13302.
197. Martinez-Millana A, Fernandez-Llatas C, Basagoiti Bilbao I, Traver Salcedo M, Traver Salcedo V. Evaluating the Social Media Performance of Hospitals in Spain: A Longitudinal and Comparative Study. *Journal of Medical Internet Research*, **2017**;19(5):e181.
198. Martins-Filho PRS, Mendes MLT, Reinheimer DM, do Nascimento-JÃºnior EM, Vaez AC, Santos VS, Santos HP Jr. Femicide trends in Brazil: relationship between public interest and mortality rates. *Archives of Womens Mental Health*, **2018** Oct;21(5):579-582.
199. Matsuda S, Aoki K, Tomizawa S, Sone M, Tanaka R, Kuriki H, Takahashi Y. Analysis of Patient Narratives in Disease Blogs on the Internet: An Exploratory Study of Social Pharmacovigilance. *JMIR Public Health and Surveillance*, **2017**;3(1):e10.

200. Mowery D, Smith H, Cheney T, Stoddard G, Coppersmith G, Bryan C, Conway M. Understanding Depressive Symptoms and Psychosocial Stressors on Twitter: A Corpus-Based Study. *Journal of Medical Internet Research*, **2017**;19(2):e48.
201. Mukhija D, Venkatraman A, Nagpal SJS. Effectivity of Awareness Months in Increasing Internet Search Activity for Top Malignancies Among Women. *JMIR Public Health and Surveillance,* **2017**;3(3):e55.
202. Noll-Hussong M. Whiplash Syndrome Reloaded: Digital Echoes of Whiplash Syndrome in the European Internet Search Engine Context. *JMIR Public Health and Surveillance,* **2017**;3(1):e15.
203. Park SH, Hong SH. Identification of Primary Medication Concerns Regarding Thyroid Hormone Replacement Therapy From Online Patient Medication Reviews: Text Mining of Social Network Data. *Journal of Medical Internet Research*, **2018**;20(10):e11085.
204. Rabarison KM, Croston MA, Englar NK, Bish CL, Flynn SM, Johnson CC. Measuring Audience Engagement for Public Health Twitter Chats: Insights From #LiveFitNOLA. *JMIR Public Health and Surveillance*, **2017**;3(2):e34.
205. Shi J, Salmon CT. Identifying Opinion Leaders to Promote Organ Donation on Social Media: Network Study. *Journal of Medical Internet Research*, **2018**;20(1):e7.
206. Tafti A, Badger J, LaRose E, Shirzadi E, Mahnke A, Mayer J, Ye Z, Page D, Peissig P. Adverse Drug Event Discovery Using Biomedical Literature: A Big Data Neural Network Adventure. *JMIR Medical Informatics*, **2017**;5(4):e51.
207. Tougas ME, Chambers CT, Corkum P, Robillard JM, Gruzd A, Howard V, Kampen A, Boerner KE, Hundert AS. Social Media Content About Children's Pain and Sleep: Content and Network Analysis. *JMIR Pediatrics and Parenting*, **2018**;1(2):e11193.
208. Winchester DE, Baxter D, Markham MJ, Beyth RJ. Quality of Social Media and Web-Based Information Regarding Inappropriate Nuclear Cardiac Stress Testing and the Choosing Wisely Campaign: A Cross-Sectional Study. *Interactive Journal of Medical Research*, **2017**;6(1):e6.
209. Wood LN, Jamnagerwalla J, Markowitz MA, Thum DJ, McCarty P, Medendorp AR, Raz S, Kim JH. Public Awareness of Uterine Power Morcellation Through US Food and Drug Administration Communications: Analysis of Google Trends Search Term Patterns. *JMIR Public Health and Surveillance*, **2018**;4(2):e47.
210. Google Trends. URL: https://trends.google.com/trends/

211. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of google trends in health care research: a systematic review. *PLoS One*, 2014;9(10):e109583
212. Hyndman RJ, Athanaspoulos G. Forecasting Principles and Practice. *Otexts*, 2014 (www.otexts.com/fpp).
213. Pearl J. Causal inference in statistics: An overview. *Statistics Surveys*, 2009;3:96–146
214. Moreno-Fernandez MM, Matute H. Biased Sampling and Causal Estimation of Health-Related Information: Laboratory-Based Experimental Research. *Journal of Medical Internet Research,* 2020;22(7):e17502.
215. Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology*, **2016**;45:6
216. VanderWeele TJ, Hernan MA. Causal effects and natural laws: Towards a conceptualization of causal counterfactuals for non- manipulable exposures, with application to the effects of race and sex. In: Berzuini C, Dawid A, Bernardinelli L (eds). *Causality: Statistical Perspective and Applications. Hoboken, NJ: John Wiley,* **2012**:101–13.
217. Lopez-Paz D, Muandet K, Scholkopf B, Tolstikhin I. Towards a Learning Theory of Cause-Effect Inference. *Proceedings of the* 32nd *International Conference on Machine Learning*, Lille, France, **2015**. JMLR: W&CP volume 37.
218. Hoyer PO, Janzing D, Mooij J, Peters J, Scholkopf B. Nonlinear causal discovery with additive noise models. Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, **2008**.
219. Shimizu S, Hoyer PO, Hyvarinen A, Kerminen AJ. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, **2006**;7:2003–2030.
220. Sun X, Janzing D, Scholkopf B.. Distinguishing between cause and effect via kernel-based complexity measures for conditional probability densities. *Neurocomputing*, 2008;1248–1256.
221. Marx A, Vreeken J. Telling cause from effect by local and global regression. *Knowledge and Information Systems*, 2019;60:1277–1305.

222. Greenland S. For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates. *European Journal of Epidemiology*, **2017**;32:3–20.
223. Rothman KJ, Greenland S. Causation and Causal Inference in Epidemiology. *American Journal of Public Health,* **2005**;95(1):144-150.
224. Budhathoki K, Vreeken J. Causal Inference by Compression. IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 12-15 Dec, 2016.
225. Hill AB. The Environment and Disease: Association or Causation? Journal of the Royal Society of Medicine, 2015;108(1):32-37.
226. Bousquet J, O'Hehir RE, Anto JM, D'Amato G, Mösges R, Hellings PW, Van Eerd M, Sheikh A. Assessment of thunderstorm-induced asthma using Google Trends. *Journal of Allergy and Clinical Immunology*, **2017**;140(3):891-893.
227. Sahanic S, Boehm A, Pizzini A, Sonnweber T, Aichner M, Weiss G, Loeffler-Ragg J, Tancevski I. Assessing self-medication for obstructive airway disease during COVID-19 using Google Trends. *European Respiratory Journal,* **2020**;56(5):2002851.
228. Patel JC, Khurana P, Sharma YK, Kumar B, Ragumani S. Chronic lifestyle diseases display seasonal sensitive comorbid trend in human population evidence from Google Trends. *PLoS One*, **2018**;13(12):e0207359.
229. Sousa-Pinto B, Heffler E, Antó A *et al.* Anomalous asthma and chronic obstructive pulmonary disease Google Trends patterns during the COVID-19 pandemic. *Clinical and Translational Allergy,* **2020**;10:47.
230. Wisniewski JA, McLaughlin AP, Stenger PJ, Patrie J, Brown MA, El-Dahr JM, PlattspMills TAE, Byrd NJ, Heymann PW. A comparison of seasonal trends in asthma exacerbations among children from geographic regions with different climates. *Allergy and Asthma Proceedings*, **2016**;37(6):475-481.
231. Cohen HA, Blau H, Hoshen M, Batat E, Balicer RD. Seasonality of asthma: a retrospective population study. *Pediatrics*, **2014;**133(4):e923-32.
232. Szefler SJ, Raphiou I, Zeiger RS, Stempel D, Kral K, Pascoe S. Seasonal variation in asthma exacerbations in the AUSTRI and VESTRI studies, *ERJ Open Research*, **2019**;5(2):001533-2018.
233. Fleming DM, Cross KW, Sunderland R, Ross AM. Comparison of the seasonal patterns of asthma identified in general practitioner episodes, hospital admissions, and deaths. *BMJ Thorax*, **2000**;55(8):662-5.

234. Santangelo OE, Provenzano S, Piazza D, Giordano D, Calamusa G, Firenze A. Digital epidemiology: assessment of measles infection through Google Trends mechanism in Italy. *Annali di Igiene*, **2019**;31:385-391.
235. Santangelo OE, Provenzano S, Grigis D, Giordano D, Armetta F, Firenze A. Can Google Trends and Wikipedia help traditional surveillance? A pilot study on measles. *Acta Biomedica*, **2020**;91(4):e2020190.
236. Rampally V, Mondal H, Mondal S. Global search trends on common vaccine-related information in English on the Internet. *Journal of Family Medicine and Primary Care*, **2020**;9(2):698-705.
237. Zhou X, Ye J, Feng Y. Tuberculosis surveillance by analyzing Google trends. *IEEE Transactions on Biomedical Engineering,* **2011**;58(8).
238. Kostkova P, Fowler D, Wiseman S, Weinberg JR. Major Infection Events Over 5 Years: How Is Media Coverage Influencing Online Information Needs of Health Care Professionals and the Public? *Journal of Medical Internet Research*, **2013**;15(7):e107.
239. Frauenfeld L, Nann D, Sulyok Z, Feng YS, Sulyok M. Forecasting tuberculosis using diabetes-related google trends data. *Pathogens and Global Health*, **2020**;114(5):236-241.
240. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis, *Science*, **2014**, 343(6176):1203-1205.
241. Adelhoefer S, Henry TS, Blankstein R, Graham G, Blaha MJ, Dzaye O. Declining interest in clinical imaging during the COVID-19 pandemic: An analysis of Google Trends data. *Clinical Imaging*, **2020**;73:20-22.
242. Ahmad I, Flanagan R, Staller K. Increased Internet Search Interest for GI Symptoms May Predict COVID-19 Cases in US Hotspots. *Clinal Gastroenterology and Hepatolology*, **2020**;18(12):2833-2834.e3.
243. Arshad Ali S, Bin Arif T, Maab H, Baloch M, Manazir S, Jawed F, Ochani RK. Global Interest in Telehealth During COVID-19 Pandemic: An Analysis of Google Trends™. *Cureus*, **2020**;12(9):e10487.
244. Asseo K, Fierro F, Slavutsky Y, Frasnelli J, Niv MY. Tracking COVID-19 using taste and smell loss Google searches is not a reliable strategy. *Scientific Reports*, **2020**;10(1):20527.
245. Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, R Niakan Kalhori S. Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran:

Data Mining and Deep Learning Pilot Study. *JMIR Public Health and Surveillance*. **2020**;6(2):e18828.

246. Azzam DB, Cypen SG, Tao JP. Oculofacial plastic surgery-related online search trends including the impact of the COVID-19 pandemic. *Orbit the International Journal on Orbital Disorders, Oculoplastic and Lacrimal Surgery,* **2021**;40(1):44-50.
247. Badell-Grau RA, Cuff JP, Kelly BP, Waller-Evans H, Lloyd-Evans E. Investigating the Prevalence of Reactive Online Searching in the COVID-19 Pandemic: Infoveillance Study. *Journal of Medical Internet Research*, **2020**;22(10):e19791.
248. Bento AI, Nguyen T, Wing C, Lozano-Rojas F, Ahn YY, Simon K. Evidence from internet search data shows information-seeking responses to news of local COVID-19 cases. *Proceedings of the National Academy of Sciences*, **2020**;117(21):11220-11222.
249. Bettencourt-Silva JH, Mulligan N, Jochim C, Yadav N, Sedlazek W, Lopez V, Gleize M. Exploring the Social Drivers of Health During a Pandemic: Leveraging Knowledge Graphs and Population Trends in COVID-19. *Studies in Health Technology and Informatics*, **2020**;275:6-11
250. Boserup B, McKenney M, Elkbuli A. The impact of the COVID-19 pandemic on emergency department visits and patient safety in the United States. *The American Journal of Emergency Medicine*, **2020**;38(9):1732-1736
251. Brodeur A, Clark AE, Fleche S, Powdthavee N. COVID-19, lockdowns and well-being: Evidence from Google Trends. *Journal of Public Economics*, **2021**;193:104346.
252. Burnett D, Eapen V, Lin PI. Time Trends of the Public's Attention Toward Suicide During the COVID-19 Pandemic: Retrospective, Longitudinal Time-Series Study. *JMIR Public Health and Surveillance*, **2020**;6(4):e24694.
253. Cherry G, Rocke J, Chu M, Liu J, Lechner M, Lund VJ, Kumar BN. Loss of smell and taste: a new marker of COVID-19? Tracking reduced sense of smell during the coronavirus pandemic using search trends. *Expert Review of Anti Infective Therapy*, **2020**;18(11):1165-1170.
254. Choi M, Tessler H, Kao G. Arts and crafts as an educational strategy and coping mechanism for Republic of Korea and United States parents during the COVID-19 pandemic. *International Review of Education*, **2020**;23:1-21.

255. Ciofani JL, Han D, Allahwala UK, Asrress KN, Bhindi R. Internet search volume for chest pain during the COVID-19 pandemic. *The American Heart Journal*, **2021**;231:157-159.
256. Cousins HC, Cousins CC, Harris A, Pasquale LR. Regional Infoveillance of COVID-19 Case Rates: Analysis of Search-Engine Query Patterns. *Journal of Medical Internet Research*, **2020**;22(7):e19483.
257. Du H, Yang J, King RB, Yang L, Chi P. COVID-19 Increases Online Searches for Emotional and Health-Related Terms. *Applied Psychology: Health and Well-Being,* **2020**;12(4):1039-1053.
258. Effenberger M, Kronbichler A, Shin JI, Mayer G, Tilg H, Perco P. Association of the COVID-19 pandemic with Internet Search Volumes: A Google Trends Analysis. *International Journal of Infectious Diseases*, **2020**;95:192-197.
259. Englund TR, Kinlaw AC, Sheikh SZ. Rise and Fall: Hydroxychloroquine and COVID-19 Global Trends: Interest, Political Influence, and Potential Implications. *ACR Open Rheumatology*, **2020**;2(12):760-766.
260. Gazendam A, Nucci N, Ekhtiari S, Gohal C, Zhu M, Payne A, Bhandari M. Trials and tribulations: so many potential treatments, so few answers. *International Orthopaedics*, **2020**;44(8):1467-1471.
261. Ghosh A, E-Roub F, Krishnan NC, Choudhury S, Basu A. Can google trends search inform us about the population response and public health impact of abrupt change in alcohol policy? - a case study from india during the covid-19 pandemic. *International Journal of Drug Policy,* **2020**;87:102984.
262. Greiner B, Ottwell R, Vassar M, Hartwell M. Public Interest in Preventive Measures of Coronavirus Disease 2019 Associated with Timely Issuance of Statewide Stay-at-Home Orders. *Disaster Medicine and Public Health Preparedness*, **2020**;5:1-4.
263. Gupta AK, Quinlan EM. Changing trends in surgical hair restoration: Use of Google Trends and the ISHRS practice census survey. *Journal of Cosmetic Dermatology*, **2020**;19(11):2974-2981.
264. Halford EA, Lake AM, Gould MS. Google searches for suicide and suicide risk factors in the early stages of the COVID-19 pandemic. *PLoS One*, **2020**;15(7):e0236777.
265. Hamulka J, Jeruszka-Bielak M, Górnicka M, Drywień ME, Zielinska-Pukos MA. Dietary Supplements during COVID-19 Outbreak. Results of Google Trends Analysis Supported by PLifeCOVID-19 Online Studies. *Nutrients*, **2020**;13(1):54.

266. Hartwell M, Greiner B, Kilburn Z, Ottwell R. Association of Public Interest in Preventive Measures and Increased COVID-19 Cases After the Expiration of Stay-at-Home Orders: A Cross-Sectional Study. *Disaster Medicine and Public Health Preparedness*, **2020**;10:1-5.
267. Higgins TS, Wu AW, Sharma D, Illing EA, Rubel K, Ting JY; Snot Force Alliance. Correlations of Online Search Engine Trends With Coronavirus Disease (COVID-19) Incidence: Infodemiology Study. *JMIR Public Health and Surveillance*, **2020**;6(2):e19702.
268. Hoerger M, Alonzi S, Perry LM, Voss HM, Easwar S, Gerhart JI. Impact of the COVID-19 pandemic on mental health: Real-time surveillance using Google Trends. *Psychological Trauma: Theory, Research, Practice, and Policy*, **2020**;12(6):567-568.
269. Hong YR, Lawrence J, Williams D Jr, Mainous III A. Population-Level Interest and Telehealth Capacity of US Hospitals in Response to COVID-19: Cross-Sectional Analysis of Google Search and National Hospital Survey Data. *JMIR Public Health and Surveillance*, **2020**;6(2):e18961.
270. Hou Z, Du F, Zhou X, Jiang H, Martin S, Larson H, Lin L. Cross-Country Comparison of Public Awareness, Rumors, and Behavioral Responses to the COVID-19 Epidemic: Infodemiology Study. *Journal of Medical Internet Research*, **2020**;22(8):e21143.
271. Hu D, Lou X, Xu Z, Meng N, Xie Q, Zhang M, Zou Y, Liu J, Sun G, Wang F. More effective strategies are required to strengthen public awareness of COVID-19: Evidence from Google Trends. *Journal of Global Health*, **2020**;10(1):011003.
272. Husain I, Briggs B, Lefebvre C, Cline DM, Stopyra JP, O'Brien MC, Vaithi R, Gilmore S, Countryman C. Fluctuation of Public Interest in COVID-19 in the United States: Retrospective Analysis of Google Trends Search Data. *JMIR Public Health and Surveillance,* **2020**;6(3):e19969.
273. Husnayain A, Shim E, Fuad A, Su EC. Understanding the Community Risk Perceptions of the COVID-19 Outbreak in South Korea: Infodemiology Study. *Journal of Medical Internet Research*, **2020**;22(9):e19788.
274. Husnayain A, Fuad A, Su EC. Applications of Google Search Trends for risk communication in infectious disease management: A case study of the COVID-19 outbreak in Taiwan. *International Journal of Infectious Diseases*, **2020**;95:221-223.

275. Jimenez AJ, Estevez-Reboredo RM, Santed MA, Ramos V. COVID-19 Symptom-Related Google Searches and Local COVID-19 Incidence in Spain: Correlational Study. *Journal of Medical Internet Research*, **2020**;22(12):e23518.
276. Kardeş S, Kuzu AS, Pakhchanian H, Raiker R, Karagülle M. Population-level interest in anti-rheumatic drugs in the COVID-19 era: insights from Google Trends. *Clinical Rheumatology*, **2020**;31:1–9.
277. Kardeş S, Kuzu AS, Raiker R, Pakhchanian H, Karagülle M. Public interest in rheumatic diseases and rheumatologist in the United States during the COVID-19 pandemic: evidence from Google Trends. *Rheumatology International*, **2021**;41(2):329-334.
278. Knipe D, Evans H, Marchant A, Gunnell D, John A. Mapping population mental health concerns related to COVID-19 and the consequences of physical distancing: a Google trends analysis. *Wellcome Open Research*, **2020**;5:82.
279. Kurian SJ, Bhatti AUR, Alvi MA, Ting HH, Storlie C, Wilson PM, Shah ND, Liu H, Bydon M. Correlations Between COVID-19 Cases and Google Trends Data in the United States: A State-by-State Analysis. *Mayo Clinic Proceedings,* **2020**;95(11):2370-2381
280. Kutlu Ö. Analysis of dermatologic conditions in Turkey and Italy by using Google Trends analysis in the era of the COVID-19 pandemic. *Dermatologic Therapy*, **2020**;33(6):e13949.
281. Landy DC, Chalmers BP, Utset-Ward TJ, Ast MP. Public Interest in Knee Replacement Fell During the Onset of the COVID-19 Pandemic: A Google Trends Analysis. *HSS Journal*, **2020**;16(Suppl 1):1-5.
282. Li C, Chen LJ, Chen X, Zhang M, Pang CP, Chen H. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. *Euro Surveillance*, **2020**;25(10):2000199.
283. Lim JL, Ong CY, Xie B, Low LL. Estimating Information Seeking-Behaviour of Public in Malaysia During COVID-19 by Using Google Trends. *Malaysian Journal of Medical Sciences*, **2020**;27(5):202-204.
284. Lin YH, Chiang TW, Lin YL. Increased Internet Searches for Insomnia as an Indicator of Global Mental Health During the COVID-19 Pandemic: Multinational Longitudinal Study. *Journal of Medical Internet Research*, **2020**;22(9):e22181.
285. Lippi G, Mattiuzzi C, Cervellin G. Google search volume predicts the emergence of COVID-19 outbreaks. Acta Biomed. 2020 Sep 7;91(3):e2020006.

286. Mavragani A. Tracking COVID-19 in Europe: Infodemiology Approach. *JMIR Public Health and Surveillance*, **2020**;6(2):e18941.
287. Mavragani A, Gkillas K. COVID-19 predictability in the United States using Google Trends time series. Scientific Reports, **2020**;10(1):20693.
288. Mayasari NR, Ho DKN, Lundy DJ, Skalny AV, Tinkov AA, Teng IC, Wu MC, Faradina A, Mohammed AZM, Park JM, Ngu YJ, Aliné S, Shofia NM, Chang JS. Impacts of the COVID-19 Pandemic on Food Security and Diet-Related Lifestyle Behaviors: An Analytical Study of Google Trends-Based Query Volumes. *Nutrients*, **2020**;12(10):3103.
289. Muselli M, Cofini V, Desideri G, Necozione S. Coronavirus (Covid-19) pandemic: How may communication strategies influence our behaviours? *International Journal of Disaster Risk Reduction*, **2021**;53:101982.
290. Niburski K, Niburski O. Impact of Trump's Promotion of Unproven COVID-19 Treatments and Subsequent Internet Trends: Observational Study. *Journal of Medical Internet Research*, **2020**;22(11):e20044.
291. Niu B, Liang R, Zhang S, Zhang H, Qu X, Su Q, Zheng L, Chen Q. Epidemic analysis of COVID-19 in Italy based on spatiotemporal geographic information and Google Trends. *Transboundary and Emergency Diseases*, **2020**. In print.
292. Nsoesie EO, Cesare N, Müller M, Ozonoff A. COVID-19 Misinformation Spread in Eight Countries: Exponential Growth Modeling Study. *Journal of Medical Internet Research*, **2020**;22(12):e24425.
293. Paguio JA, Yao JS, Dee EC. Silver lining of COVID-19: Heightened global interest in pneumococcal and influenza vaccines, an infodemiology study. *Vaccine,* **2020**;38(34):5430-5435.
294. Pang R, Wei Z, Liu W, Chen Z, Cheng X, Zhang H, Li G, Liu L. Influence of the pandemic dissemination of COVID-19 on facial rejuvenation: A survey of Twitter. *Journal of Cosmetic Dermatology*, **2020**;19(11):2778-2784.
295. Panuganti BA, Jafari A, MacDonald B, DeConde AS. Predicting COVID-19 Incidence Using Anosmia and Other COVID-19 Symptomatology: Preliminary Analysis Using Google and Twitter. *Otolaryngology-Head and Neck Surgery*, **2020**;163(3):491-497.
296. Peng Y, Li C, Rong Y, Chen X, Chen H. Retrospective analysis of the accuracy of predicting the alert level of COVID-19 in 202 countries using Google Trends and machine learning. *Journal of Global Health*, **2020**;10(2):020511.

297. Pier MM, Pasick LJ, Benito DA, Alnouri G, Sataloff RT. Otolaryngology- related Google Search trends during the COVID-19 pandemic. *The American Journal of Otolaryngology*, **2020**;41(6):102615.
298. Prasanth S, Singh U, Kumar A, Tikkiwal VA, Chong PHJ. Forecasting spread of COVID-19 using google trends: A hybrid GWO-deep learning approach. *Chaos Solitons & Fractals*, **2021**;142:110336.
299. Rajan A, Sharaf R, Brown RS, Sharaiha RZ, Lebwohl B, Mahadev S. Association of Search Query Interest in Gastrointestinal Symptoms With COVID-19 Diagnosis in the United States: Infodemiology Study. *JMIR Public Health and Surveillance*, **2020**;6(3):e19354.
300. Rokhmah D, Ali K, Putri SMD, Khoiron K. Increase in public interest concerning alternative medicine during the COVID-19 pandemic in Indonesia: a Google Trends study. *F1000Research*, **2020**,;9:1201.
301. Rovetta A, Bhagavathula AS. Global Infodemiology of COVID-19: Analysis of Google Web Searches and Instagram Hashtags. *Journal of Medical Internet Research*, **2020**;22(8):e20673.
302. Rovetta A, Bhagavathula AS. COVID-19-Related Web Search Behaviors and Infodemic Attitudes in Italy: Infodemiological Study. *JMIR Public Health and Surveillance*, **2020**;6(2):e19374.
303. Rovetta A, Castaldo L. The Impact of COVID-19 on Italian Web Users: A Quantitative Analysis of Regional Hygiene Interest and Emotional Response. *Cureus*, **2020**;12(9):e10719.
304. Schnoell J, Besser G, Jank BJ, Bartosik TJ, Parzefall T, Riss D, Mueller CA, Liu DT. The association between COVID-19 cases and deaths and web-based public inquiries. *Infectious Diseases*, **2021**;53(3):176-183.
305. Senecal C, Gulati R, Lerman A. Google Trends Insights Into Reduced Acute Coronary Syndrome Admissions During the COVID-19 Pandemic: Infodemiology Study. *JMIR Cardio*, **2020**;4(1):e20426.
306. Singh S, Sharma P, Balhara YPS. The impact of nationwide alcohol ban during the COVID-19 lockdown on alcohol use-related internet searches and behaviour in India: An infodemiology study. *Drug and Alcohol Reviews*, **2020**;13187. In print.
307. Sinyor M, Spittal MJ, Niederkrotenthaler T. Changes in Suicide and Resilience-related Google Searches during the Early Stages of the COVID-19 Pandemic. *Canadian Journal of Psychiatry*, **2020**;65(10):741-743.

308. Sousa-Pinto B, Heffler E, Antó A, Czarlewski W, Bedbrook A, Gemicioglu B, Canonica GW, Antó JM, Fonseca JA, Bousquet J. Anomalous asthma and chronic obstructive pulmonary disease Google Trends patterns during the COVID-19 pandemic. *Clinical and Translational Allergy*, **2020**;10(1):47.
309. Sousa-Pinto B, Anto A, Czarlewski W, Anto JM, Fonseca JA, Bousquet J. Assessment of the Impact of Media Coverage on COVID-19-Related Google Trends Data: Infodemiology Study. *Journal of Medical Internet Research,* **2020**;22(8):e19611
310. Strzelecki A, Azevedo A, Albuquerque A. Correlation between the Spread of COVID-19 and the Interest in Personal Protective Measures in Poland and Portugal. *Healthcare (Basel)*, **2020**;8(3):203.
311. Subhash AK, Maldonado DR, Kajikawa TM, Chen SL, Stavrakis A, Photopoulos C. Public Interest in Sports Medicine and Surgery (Anterior Cruciate Ligament, Meniscus, Rotator Cuff) Topics Declined Following the COVID-19 Outbreak. *Arthroscopy, Sports Medicine, and Rehabililation*, **2021**;3(1):e149-e154.
312. Sulyok M, Ferenci T, Walker M. Google Trends Data and COVID-19 in Europe: correlations and model enhancement are European wide. *Transboundary and Emergency Disorders*, **2020**;13887. In print.
313. Sycinska-Dziarnowska M, Paradowska-Stankiewicz I. Dental Challenges and the Needs of the Population during the Covid-19 Pandemic Period. Real-Time Surveillance Using Google Trends. *International Journal of Environmental Research and Public Health*, **2020**;17(23):8999
314. Szmuda T, Ali S, Hetzger TV, Rosvall P, Słoniewski P. Are online searches for the novel coronavirus (COVID-19) related to media or epidemiology? A cross-sectional study. *International Journal of Infectious Disorders*, **2020**;97:386-390.
315. Tijerina JD, Cohen SA, Parham MJ, Debbaut C, Cohen L, Stevanovic M, Lefebvre R. Public Interest in Elective Orthopedic Surgery Following Recommendations During COVID-19: A Google Trends Analysis. *Cureus*, **2020**;12(12):e12123.
316. Uvais NA. Association Between the COVID-19 Outbreak and Mental Health in India: A Google Trends Study. *Primary Care Companion for CNS Disorders*, **2020**;22(6):20br02778.
317. Venkatesh U, Gandhi PA. Prediction of COVID-19 Outbreaks Using Google Trends in India: A Retrospective Analysis. *Healthcare Informatics Research*, **2020**;26(3):175-184.

318. Walker A, Hopkins C, Surda P. Use of Google Trends to investigate loss-of-smell-related searches during the COVID-19 outbreak. *International Forum of Allergy and Rhinology*, **2020**;10(7):839-847.
319. Walker MD, Sulyok M. Online behavioural patterns for Coronavirus disease 2019 (COVID-19) in the United Kingdom. *Epidemiology and Infection*, **2020**;148:e110.
320. Xie T, Tan T, Li J. An Extensive Search Trends-Based Analysis of Public Attention on Social Media in the Early Outbreak of COVID-19 in China. *Risk Management Healthcare Policy*, **2020**;13:1353-1364.
321. Younis J, Freitag H, Ruthberg JS, Romanes JP, Nielsen C, Mehta N. Social Media as an Early Proxy for Social Distancing Indicated by the COVID-19 Reproduction Number: Observational Study. *JMIR Public Health and Surveillance*, **2020**;6(4):e21340.
322. Yuan X, Xu J, Hussain S, Wang H, Gao N, Zhang L. Trends and Prediction in Daily New Cases and Deaths of COVID-19 in the United States: An Internet Search-Interest Based Model. *Exploratory Research and Hypothesis in Medicine,* **2020**;5(2):1-6.
323. Zattoni F, Gül M, Soligo M, Morlacco A, Motterle G, Collavino J, Barneschi AC, Moschini M, Moro FD. The impact of COVID-19 pandemic on pornography habits: a global analysis of Google Trends. *International Journal of Impotence Research,* **2020**;28:1–8.
324. Salathé M. Digital epidemiology: what is it, and where is it going? *Life Sciences, Society and Policy,* **2018***;*14(1).
325. Eckmanns T, Füller H, Roberts SL. Digital epidemiology and global health security; an interdisciplinary conversation. *Life Sciences, Society and Policy*, **2019**;19;15(1):2.
326. Choi BC, Pak AW. Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. *Clinical and Investigating Medicine*, **2006**;29(6):351-64.
327. Smith Jervelund S, Villadsen SF. Evidence in public health: An integrated, multidisciplinary concept. *Scandinavian Journal of Public Health*, **2022**;50(7):1012-1017.
328. RAYNER G. Multidisciplinary public health: Leading from the front? *Public Health,* **2007**;121(6):449-454.
329. Kivits J, Ricci L, Minary L. Interdisciplinary research in public health: the 'why' and the 'how'. *Journal of Epidemiology and Community Health*, **2019**;73(12):1061-1062.

330. Hinchman A, Ali D, Goodwin BW, Gillie M, Boudreaux J, Laborde Y. Global Health Is Local Health: A Multidisciplinary Perspective of COVID-19. *Ochsner Journal*, **2020**;20(2):123-133.
331. FAQs about Google Trends data. URL: https://support.google.com/trends/answer/4365533?hl=en
332. Google News Lab. What is Google Trends data -and what does it mean? URL: https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8
333. Rangarajan P, Mody SK, Marathe M. Forecasting dengue and influenza incidences using a sparse representation of Google trends, electronic health records, and time series data. *PLoS Computational Biology*, **2019**;15(11): e1007518.
334. Lampros V, Miller AC, Crossan S, Stefansen C. Advances in nowcasting influenza-like illness rates using query logs. *Scientific Reports*, **2015**;5:12760.
335. Reich NG, McGowan CJ, Yamana TK, Tushar A, Ray EL, Osthus D, et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. PLoS Computational Biology, **2019**;5(11):e1007486.
336. Nikolopoulos K, Punia S, Schäfers A, Tsinopoulos C, Vasilakis C. Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions, *European Journal of Operational Research*, **2021**;290:99–115.
337. Henning KJ. Overview of Syndromic Surveillance What is Syndromic Surveillance? *Centers for Disease Control and Prevention: Morbidity and Mortality Weekly Report (MMWR)*, **2004**;53(Suppl):5-11.
338. Yan SJ, Chughtaia AA, Macintyrea CR. Utility and potential of rapid epidemic intelligence from internet-based sources. *International Journal of Infectious Diseases*, **2017**;63:77–87.
339. Davidson MW, Haim DA, Rdin JM. Using Networks to Combine 'Big Data' and Traditional Surveillance to Improve Influenza Predictions. *Scientific Reports*, **2015**;5:8154

Appendix

This section consists of the following appendices of the publications:

1. Appendix 1: Journal metrics and citations
2. Multimedia Appendix for [A]: Publication details and categorization
3. Multimedia Appendix for [B]: Category selection
4. Multimedia Appendix 1 for [C]: State data tables
5. Multimedia Appendix 2 for [C]: 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) by US State

Appendix 1

Table S1 provides an overview of the metrics of the Journals that the papers are published in, as well as an overview of the citations that the papers have received up to April 15, 2025.

Table S1. Metrics of the journals and citations of the published papers in the Thesis (2018-2025).

Title	Journal	Publisher	Year	Impact Factor	CiteScore (Scopus)	Quartile	Percentile (Scopus)	Citations GS [19]
Assessing the Methods, Tools, and Statistical Approaches in Google Trends Research [1]	Journal of Medical internet Research [9]	JMIR Publications	2018	5.8	11.4 [14]	Q1	**95th percentile** in *Health informatics* [14]	385
Google Trends in Infodemiology and Infoveillance: Methodology Framework [2]	JMIR Public Health and Surveillance [10]	JMIR Publications	2019	3.5	6.3 [15]	Q1	**97th percentile** in *Public Health, Environmental and Occupational Health* [15]	441
Integrating Smart Health in the US Health Care System: Infodemiology Study of Asthma Monitoring in the Google Era [3]	JMIR Public Health and Surveillance [10]	JMIR Publications	2018	3.5	6.3 [15]	Q1	**97th percentile** in *Public Health, Environmental and Occupational Health* [15]	29
The Internet and the Anti-Vaccine Movement: Tracking the 2017 EU Measles Outbreak [4]	Big Data and Cognitive Computing [11]	MDPI	2018	3.7	9.2 [16]	Q2	**76th percentile** in *Computer Science Applications* [16]	67
Forecasting AIDS prevalence in the United States using online search traffic data [5]	Journal of Big Data [12]	Springer Nature	2018	8.6	21.1 [17]	Q1	**97th percentile** in *Computer networks and communications* [17]	36
Infoveillance of infectious diseases in USA: STDs, tuberculosis, and hepatitis [6]	Journal of Big Data [12]	Springer Nature	2018	8.6	21.1 [17]	Q1	**97th percentile** in *Computer networks and communications* [17]	24
Tracking COVID-19 in Europe: infodemiology approach [7]	JMIR Public Health and Surveillance [10]	JMIR Publications	2020	3.5	6.3 [15]	Q1	**97th percentile** in *Public Health, Environmental and Occupational Health* [15]	201
COVID-19 Predictability in the United States using Google Trends time series [8]	Scientific Reports [13]	Nature Publishing Group	2020	3.8	6.6 [18]	Q1	**92nd percentile** in *Multidisciplinary* [18]	153

References

1. Mavragani A, Ochoa G, Tsagarakis KP. Assessing the Methods, Tools, and Statistical Approaches in Google Trends Research: Systematic Review; J Med Internet Res 2018;20(11):e270
2. Mavragani A, Ochoa G. Google Trends in Infodemiology and Infoveillance: Methodology Framework; JMIR Public Health Surveill 2019;5(2):e13439
3. Mavragani A, Sampri A, Sypsa K, Tsagarakis KP. Integrating Smart Health in the US Health Care System: Infodemiology Study of Asthma Monitoring in the Google Era; JMIR Public Health Surveill 2018;4(1):e24
4. Mavragani A, Ochoa G. The Internet and the Anti-Vaccine Movement: Tracking the 2017 EU Measles Outbreak. Big Data Cogn. Comput. 2018;2(2)
5. Mavragani A, Ochoa G. Forecasting AIDS prevalence in the United States using online search traffic data. J Big Data 2018;5(17)
6. Mavragani A, Ochoa G. Infoveillance of infectious diseases in USA: STDs, tuberculosis, and hepatitis. J Big Data 2018;5(30)
7. Mavragani A. Tracking COVID-19 in Europe: Infodemiology Approach; JMIR Public Health Surveill 2020; 6(2): e18941
8. Mavragani A, Gkillas K. COVID-19 predictability in the United States using Google Trends time series. Sci Rep 2020;10(20693)
9. Journal of Medical Internet Research website [https://www.jmir.org]
10. JMIR Public Health & Surveillance website [https://publichealth.jmir.org]
11. Big Data & Cognitive Computing website [https://www.mdpi.com/journal/BDCC]
12. Journal of Big Data website [https://journalofbigdata.springeropen.com]
13. Scientific Reports website [https://www.nature.com/srep/]
14. Journal of Medical Internet Research – CiteScore and Rank [https://www.scopus.com/sourceid/23709]
15. JMIR Public Health & Surveillance – CiteScore and Rank [https://www.scopus.com/sourceid/21101030810]
16. Big Data & Cognitive Computing – CiteScore and Rank [https://www.scopus.com/sourceid/21101020112]
17. Journal of Big Data – CiteScore and Rank [https://www.scopus.com/sourceid/21100791292]
18. Scientific Reports – CiteScore and Rank [https://www.scopus.com/sourceid/21100200805]
19. Google Scholar - Amaryllis Mavragani [https://scholar.google.gr/citations?user=MUoMSPgAAAAJ&hl=en]

Multimedia Appendix 1. Publication details and categorization.

	Authors	Period	Region	Keywords	V	S	C	F	M	St	Languages
1	Alicino et al 2015	2013-2015	Guinea, Sierra Leone, Liberia, Nigeria, Mali, Senegal, USA, Spain, UK, Italy	Ebola	✔		✔		✔		English, African languages (no data)
2	Arora et al 2016	2004-2013	UK	Suicide	✔		✔				English
3	Bakker et al 2016	2004-2015	Worldwide (36 Countries)	Chicken Pox, Varicella Zoster Virus, Vaccination	✔	✔	✔	✔	✔		Multiple Languages
4	Barnes et al 2015	2008-2013	USA	Sleep, Moral Awareness	✔				✔		English
5	Bentley & Ormerod, 2009	2005, 2009	Worldwide	Bird Flu, Swine Flu	✔				✔		English
6	Borron et al 2016	2009-2015	USA	Loperamide	✔						English
7	Bragazzi, 2013	2004-2012	Italy	Multiple Sclerosis	✔		✔		✔		Italian
8	Bragazzi et al 2016	2004-2010	USA	Silicosis	✔		✔				English
9	Bragazzi et al 2016	2004-2015	Worldwide	Vaccination	✔						English
10	Bragazzi et al 2016	2004-2015	Italy	West Nile Virus	✔		✔				Italian
11	Bragazzi et al 2016	2004-2015	Worldwide	Epilepsy	✔		✔				English
12	Bragazzi et al 2016	2004-2015	Worldwide	Silicosis	✔		✔			✔	English
13	Bragazzi et al 2016	2004-2016	Worldwide, USA	Vasculitis, Autoimmune Diseases, Celebrity	✔		✔			✔	English
14	Bragazzi, 2014	2004-2012	Italy	Non-Suicidal Self Injury (NSSI)	✔		✔				Italian
15	Braun & Harreus, 2013	2005-2012	Germany	Otolaryngology, Sinusitis	✔	✔					German
16	Brigo & Trinka, 2015	2004-2014	Worldwide	Epilepsy	✔						English
17	Brigo et al 2014	2004-2013	Worldwide	Epilepsy	✔						English
18	Brigo et al 2014	2004-2013	Worldwide	Multiple Sclerosis, Epilepsy, Dementia	✔						English
19	Van Campen et al 2014	2004-2013	Netherlands, USA, UK	Epilepsy, Seizures	✔					✔	English, Dutch

	Authors	Period	Region	Keywords	V	S	C	F	M	St	Languages
20	Carneiro & Mylonakis, 2009	2004-2009	Worldwide, USA	Influenza, Flu, West Nile virus, Bird Flu, Respiratory Syncytial Virus (RSV), Bird Flu	✔						English
21	Cavazos-Regh et al 2015	2011	USA	Tobacco	✔		✔				English
22	Cha & Stow, 2015	2004-2014	USA	Toledo Water Crisis, Algae	✔						English
23	Chaves et al 2015	2004-2012	Worldwide	Tele-health, Newborn Hearing Screening							Portuguese
24	Cho et al 2013	2007-2012	South Korea	Influenza (Flu, New Flu, Swine Flu, New Influenza, Fever, Tamiflu)	✔		✔				Korean
25	Crowson et al 2016	2008-2015	USA	Otitis Externa, Ototopicals, Ciprodex, Cortisporin, Ofloxacin	✔	✔	✔			✔	English
26	Davis et al 2015	2005-2015	Worldwide	Interstitial Cystitis, Painful Bladder Syndrome	✔						English
27	Fazeli et al 2014	2004-2014	USA	Breast Cancer, Dense Breast	✔						English
28	Deiner et al 2016	2012-2014	USA, Australia	Pink Eye, Eye Allergy, Flu, Eye Drops, Eye Diseases	✔	✔	✔				English
29	DeVilbiss & Lee, 2014	2004-2014	USA	Autism Awareness, Autism, Asperger's, ADHD	✔						English
30	Domnich et al 2015	2011-2015	Italy	Influenza, Fever, Cough, Tachipirina, Paracetamol	✔		✔	✔	✔		Italian
31	El-Sheikha, 2015	2006-2013	Worldwide, 44 countries	Varicose Vein Syndrome	✔	✔			✔	✔	English, 32 languages
32	Fenichel et al 2013	2008-2010	USA	Pandemic Influenza, Swine Flu	✔				✔		English
33	Fond et al 2015	2005-2014	Worldwide	Suicide, Depression, Bipolar	✔						English
34	Foroughi et al 2016	2004-2015	Australia, Canada, New Zealand, UK, USA	Cancer	✔		✔				English
35	Gafson & Giovannoni, 2014	2008-2012	Worldwide	Chronic Cerebrospinal Venous Insufficiency	✔						English

	Authors	Period	Region	Keywords	V	S	C	F	M	St	Languages
36	Gahr et al 2015	2004-2013	Germany	Antidepressants, Prescriptions	✔		✔		✔		German
37	Gamma et al 2016	2004-2016	Switzerland, Germany, Austria	Drugs, Methamphetamine Crime	✔		✔				German (same in English)
38	Garrison et al 2015	2004-2012	USA, Australia	Leg cramps	✔	✔			✔		English
39	Gollust et al 2016	2013-2014	USA	Affordable Care Act, Health Insurance, Obamacare	✔		✔		✔		English
40	Guernier et al 2016	2011-2013	Australia	Veterinary Diseases, Tick Paralysis	✔		✔				English
41	Haney et al 2014	2004-2013	USA	Radiology Residency, Radiology Salary	✔				✔		English
42	Harorli & Harorli, 2014	2004-2014	Worldwide	Oral problems	✔						English
43	Harsha et al 2014	2004-2012	USA	Varicose Vein Syndrome	✔	✔			✔	✔	English
44	Harsha et al 2015	2006-2013	USA	Interventional Radiology, Fellowships	✔	✔			✔		English
45	Hassid et al 2016	2008-2011	USA	Gastrointestinal Symptoms	✔	✔	✔				English
46	Hossain et al 2016	2014	Guinea, Liberia, Sierra Leone, USA, UK	Ebola, Flu	✔						English
47	Huang et al 2013	2009-2011	China	Smoking, Smoking Ban, Electronic Cigarette	✔						Chinese
48	Huesch et al 2014	2012-2013	USA	Public Hospitals, Quality, Ratings	✔						English
49	Ingram & Plante, 2013	2004-2012	USA, Australia, UK, Canada, Germany	Restless Legs Syndrome	✔	✔					English
50	Ingram et al 2015	2006-2012	USA, Australia	Breathing Sleep Disorder	✔	✔					English
51	Jha et al 2015	2004-2015	USA	Oral Bisphosphonate, Prescriptions, Hip Fractures, Fosamax	✔						English
52	Johnson et al 2014	2005-2011	USA	Sexually Transmitted Infections	✔		✔				English
53	Kadry et al 2011	2010	USA	Physician Rating							English
54	Kang et al 2013	2008-2011	China	Influenza, ILI, Flu, H1N1	✔		✔				Chinese

	Authors	Period	Region	Keywords	V	S	C	F	M	St	Languages
55	Kang et al 2015	2008-2013	USA, UK, Australia	Allergic Rhinitis, Allergic Rhinitis, Pollen count, Claritin, Zyrtec	✔	✔	✔				English
56	Koburger et al 2015	2009-2010	Germany, Austria, Hungary, Netherlands, Slovenia	Suicide, Robert Enke	✔		✔				German, Hungarian, Dutch, Slovenian
57	Kostkova et al 2013	2006-2010	UK	Infectious diseases, Clostridium difficile, MRSA, Tuberculosis, Meningitis, Norovirus, Influenza	✔						English
58	Lawson McLean et al 2016	2004-2014	Worldwide, Germany	Neurosurgery	✔						English
59	Leffler et al 2010	2004-2008	USA, UK, Canada, Australia	Ophthalmology	✔	✔			✔		English
60	Ling & Lee, 2016	2004-2015	Canada	Health Campaigns, HIV, AIDS, Stroke, Colorectal Cancer, Marijuana use	✔		✔				English
61	Linkov et al 2014	2004-2012	Worldwide, USA	Bariatric Surgery	✔				✔		English
62	Liu et al 2016	2004-2016	USA, Australia	Ankle Swelling	✔	✔			✔		English
63	Luckett et al 2016	2015	Worldwide	Chronic Breathlessness							English
64	Majumder et al 2016	2015-2016	Colombia	Zika Virus	✔				✔		Spanish
65	Mattin et al 2014	2007-2013	France, Greece, Italy, Portugal, Spain	Canine Leishmaniosis	✔						French, Greek, Italian, Portuguese, Spanish
66	Mavragani et al 2016	2004-2014	UK, Worldwide	Drugs, Prescriptions, Diclofenac, Estradiol, Macrolide Antibiotics	✔		✔				English
67	Murray et al 2016	2010-2013	Ireland	Mouth Cancer, Oral Cancer	✔					✔	English
68	Myers et al 2016	2004-2015	USA	Psychogenic Non-Epileptic Seizures	✔						English
69	Noar et al 2013	2006-2011	USA	Pancreatic Cancer, Public Figure	✔				✔	✔	English
70	Nuti et al 2014	2004-2014	Worldwide	Review	-	-	-	-	-	-	-

	Authors	Period	Region	Keywords	V	S	C	F	M	St	Languages
71	Pandey et al 2014	2004-2013	USA	Heart Transplant, Ventricular Assist Devices, Breast Cancer, Pulmonary Embolism, Bipolar Disorder, Sjogren Syndrome, Multiple Sclerosis	✔						English
72	Parker et al 2016	2010-2014	USA	Premature Deaths, Alcohol, Drugs, Suicide	✔			✔	✔		English
73	Phelan et al 2014	2009-2012	USA, UK, Australia, Ireland	Metal-on-Metal Hip	✔				✔	✔	English
74	Phelan et al 2016	2010-2015	USA	Anatomy, Education	✔	✔	✔		✔		English
75	Plante & Ingram, 2014	2004-2013	USA, Australia, Germany, UK, Canada, Sweden, Switzerland	Tinnitus Symptomatology	✔	✔					English, German, Swedish, French, Italian
76	Poletto et al 2016	2013-2015	Worldwide	Middle East Respiratory Syndrome (MERS)	✔		✔				English
77	Pollett et al 2015	2009-2014	USA	Pertussis	✔		✔	✔	✔		English
78	Rohart et al 2016	2009-2013	Australia	Disease Surveillance	✔		✔	✔	✔	✔	English
79	Rosenkrantz & Prabhu, 2016	2004-2014	USA	Imaging-Based Cancer Screening, Breast Cancer, Lung Cancer, Colon Cancer, Prostate Cancer	✔						English
80	Rossignol et al 2013	2004-2012	France, Germany, Italy, USA, China, Australia, Brazil, South Africa	Urinary Tract Infection, Cystitis	✔	✔					English, French, German, Italian, Chinese, Portuguese
81	Scatà et al 2016	2015-2016	56 countries in South America, Europe, Oceania	Epidemics, Zika Virus	✔				✔		N/A. The term is the same
82	Scheres et al 2016	2009-2015	Netherlands, Worldwide	Thrombosis, Venous Thrombosis	✔						English, Dutch
83	Shin et al 2016	2015-2016	Korea	MERS	✔		✔				Korean
84	Schootman et al 2015	2004-2014	50 US States and DC, Puerto Rico, US Virgin Islands, Guam, American Samoa, Palau	Cancer Screening	✔		✔				English

	Authors	Period	Region	Keywords	V	S	C	F	M	St	Languages
85	Schuster et al 2010	2004-2009	USA	Statins, Lipitor, Simvastatin	✔		✔		✔		English
86	Seifter et al 2010	2004-2009	USA	Lyme Disease	✔	✔					English
87	Sentana-Lledo et al 2016	2004-2014	USA	Bed bugs	✔	✔	✔				English
88	Simmering et al 2014	2004-2014	USA	Drugs, Prescriptions, Antibiotics	✔		✔				English
89	Skeldon et al 2015	2004-2007	USA	Drugs, Prostatic Hyperplasia, Avodart, Flomax	✔						English
90	Solano et al 2016	2008-2012	Italy	Suicide	✔		✔	✔			Italian
91	Stein et al 2013	2007-2010	USA, UK, Canada, India	Laser Eye Surgery	✔				✔		English
92	Takada, 2012	2004-2011	Japan	Fireflies, Beetles	✔	✔					Japanese
93	Telfer & Woodburn, 2015	2004-2014	UK, USA, Canada, Australia	Foot pain, Ankle pain, Heel pain	✔	✔			✔	✔	English
94	Troelstra et al 2016	2004-2013	Netherlands, Belgium	Tobacco Control	✔				✔	✔	Dutch
95	Toosi & Kalia, 2015	2004-2013	Canada, USA, Australia	Tanning	✔	✔					English
96	Wang et al 2015	2009-2011	Taiwan	Dementia, Alzheimer's Disease, Neurology	✔		✔	✔			Chinese
97	Warren & Wen, 2016	2004-2015	USA	Measles, MMR, Vaccine	✔						English
98	Willson et al 2015	2011-2012	USA	Aeroallergens, Allergies, Pollen	✔	✔	✔		✔		English
99	Willson et al 2015	2011-2014	USA	Pollen, Mountain Cedar	✔				✔		English
100	Yang et al 2015	2009-2015	USA	Influenza, Epidemic	✔				✔		English
101	Zhang et al 2015	2004-2014	USA, Canada, UK, Australia, China	Tobacco, Lung Cancer	✔	✔	✔			✔	English, Chinese
102	Zhang et al 2016	2004-2015	USA	Drugs, Dabbing, Cannabis Smoking	✔		✔	✔		✔	English
103	Zheluk et al 2014	2009-2013	Russia	Drugs, Krokodil, Desomorphine, Codeine	✔						Russian
104	Zhou et al 2011	2004-2009	USA	Tuberculosis	✔			✔	✔		English

References

1. Alicino C, Bragazzi NL, Faccio V, Amicizia D, Panatto D, Gasparini R, et al. Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes. Infect Dis Poverty 2015 Dec 10;4:54 [FREE Full text] [doi: 10.1186/s40249-015-0090-9] [Medline: 26654247]
2. Arora VS, Stuckler D, McKee M. Tracking search engine queries for suicide in the United Kingdom, 2004-2013. Public Health 2016 Aug;137:147-153. [doi: 10.1016/j.puhe.2015.10.015] [Medline: 26976489]
3. Bakker KM, Martinez-Bakker ME, Helm B, Stevenson TJ. Digital epidemiology reveals global childhood disease seasonality and the effects of immunization. PNAS 2016;113(24):6689.
4. Barnes CM, Gunia BC, Wagner DT. Sleep and moral awareness. J Sleep Res 2015 Apr;24(2):181-188 [FREE Full text] [doi: 10.1111/jsr.12231] [Medline: 25159702]
5. Bentley RA, Ormerod P. Social versus independent interest in 'bird flu' and 'swine flu'. PLoS Curr 2009 Sep 3;1:RRN1036. [doi: 10.1371/currents.RRN1036]
6. Borron SW, Watts SH, Tull J, Baeza S, Diebold S, Barrow A. Intentional Misuse and Abuse of Loperamide: A New Look at a Drug with "Low Abuse Potential". J Emerg Med 2017 Jul;53(1):73-84. [doi: 10.1016/j.jemermed.2017.03.018] [Medline: 28501383]
7. Bragazzi NL. Infodemiology and infoveillance of multiple sclerosis in Italy. Mult Scler Int 2013;2013:924029 [FREE Full text] [doi: 10.1155/2013/924029] [Medline: 24027636]
8. Bragazzi NL, Dini G, Toletone A, Brigo F, Durando P. Infodemiological data concerning silicosis in the USA in the period 2004-2010 correlating with real-world statistical data. Data Brief 2017 Feb;10:457-464 [FREE Full text] [doi: 10.1016/j.dib.2016.11.021] [Medline: 28054008]
9. Bragazzi NL, Barberis I, Rosselli R, Gianfredi V, Nucci D, Moretti M, et al. How often people google for vaccination: Qualitative and quantitative insights from a systematic search of the web-based activities using Google Trends. Hum Vaccin Immunother 2017 Feb;13(2):464-469. [doi: 10.1080/21645515.2017.1264742] [Medline: 27983896]
10. Bragazzi N, Bacigaluppi S, Robba C, Siri A, Canepa G, Brigo F. Infodemiological data of West-Nile virus disease in Italy in the study period 2004-2015. Data Brief 2016:839-845 [FREE Full text]
11. Bragazzi NL, Bacigaluppi S, Robba C, Nardone R, Trinka E, Brigo F. Infodemiology of status epilepticus: A systematic validation of the Google Trends-based search queries. Epilepsy Behav 2016 Feb;55:120-123. [doi: 10.1016/j.yebeh.2015.12.017] [Medline: 26773681]
12. Bragazzi NL, Dini G, Toletone A, Brigo F, Durando P. Leveraging Big Data for Exploring Occupational Diseases-Related Interest at the Level of Scientific Community, Media Coverage and Novel Data Streams: The Example of Silicosis as a Pilot Study. PLoS One 2016;11(11):e0166051 [FREE Full text] [doi: 10.1371/journal.pone.0166051] [Medline: 27806115]
13. Bragazzi NL, Watad A, Brigo F, Adawi M, Amital H, Shoenfeld Y. Public health awareness of autoimmune diseases after the death of a celebrity. Clin Rheumatol 2016 Dec 20:1911-1917. [doi: 10.1007/s10067-016-3513-5] [Medline: 28000011]
14. Bragazzi NL. A Google Trends-based approach for monitoring NSSI. Psychol Res Behav Manag 2013 Dec;7:1-8 [FREE Full text] [doi: 10.2147/PRBM.S44084] [Medline: 24376364]
15. Braun T, Harréus U. Medical nowcasting using Google Trends: application in otolaryngology. Eur Arch Otorhinolaryngol 2013 Jul;270(7):2157-2160. [doi: 10.1007/s00405-013-2532-y] [Medline: 23632877]
16. Brigo F, Trinka E. Google search behavior for status epilepticus. Epilepsy Behav 2015 Aug;49:146-149. [doi: 10.1016/j.yebeh.2015.02.029] [Medline: 25873438]
17. Brigo F, Igwe SC, Ausserer H, Nardone R, Tezzon F, Bongiovanni LG, et al. Why do people Google epilepsy? An infodemiological study of online behavior for epilepsy-related search terms. Epilepsy Behav 2014 Feb;31:67-70. [doi: 10.1016/j.yebeh.2013.11.020] [Medline: 24361764]
18. Brigo F, Lochner P, Tezzon F, Nardone R. Web search behavior for multiple sclerosis: An infodemiological study. Multiple Sclerosis and Related Disorders 2014 Jul;3(4):440-443. [doi: 10.1016/j.msard.2014.02.005]

19. van CJS, van DE, Otte WM, Joels M, Jansen FE, Braun KPJ. Does Saint Nicholas provoke seizures? Hints from Google Trends. Epilepsy Behav 2014 Mar;32:132-134. [doi: 10.1016/j.yebeh.2014.01.019] [Medline: 24548849]
20. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 2009 Nov 15;49(10):1557-1564 [FREE Full text] [doi: 10.1086/630200] [Medline: 19845471]
21. Cavazos-Rehg PA, Krauss MJ, Spitznagel EL, Lowery A, Grucza RA, Chaloupka FJ, et al. Monitoring of non-cigarette tobacco use using Google Trends. Tob Control 2015 May;24(3):249-255 [FREE Full text] [doi: 10.1136/tobaccocontrol-2013-051276] [Medline: 24500269]
22. Cha Y, Stow CA. Mining web-based data to assess public response to environmental events. Environ Pollut 2015 Mar;198:97-99. [doi: 10.1016/j.envpol.2014.12.027] [Medline: 25577650]
23. Chaves JN, Libardi AL, Agostinho-Pesse RS, Morettin M, Alvarenga KDF. Tele-health: assessment of websites on newborn hearing screening in Portuguese Language. Codas 2015 Dec;27(6):526-533 [FREE Full text] [doi: 10.1590/2317-1782/20152014169] [Medline: 26691616]
24. Cho S, Sohn CH, Jo MW, Shin S, Lee JH, Ryoo SM, et al. Correlation between national influenza surveillance data and google trends in South Korea. PLoS One 2013 Dec;8(12):e81422 [FREE Full text] [doi: 10.1371/journal.pone.0081422] [Medline: 24339927]
25. Crowson MG, Schulz K, Tucci DL. National Utilization and Forecasting of Ototopical Antibiotics. Otology & Neurotology 2016;37(8):1049-1054. [doi: 10.1097/MAO.0000000000001115]
26. Davis NF, Gnanappiragasam S, Thornhill JA. Interstitial cystitis/painful bladder syndrome: the influence of modern diagnostic criteria on epidemiology and on Internet search activity by the public. Transl Androl Urol 2015 Oct;4(5):506-511 [FREE Full text] [doi: 10.3978/j.issn.2223-4683.2015.06.08] [Medline: 26816850]
27. Fazeli DS, Carlos RC, Hall KS, Dalton VK. Novel data sources for women's health research: mapping breast screening online information seeking through Google trends. Acad Radiol 2014 Sep;21(9):1172-1176 [FREE Full text] [doi: 10.1016/j.acra.2014.05.005] [Medline: 24998689]
28. Deiner MS, Lietman TM, McLeod SD, Chodosh J, Porco TC. Surveillance Tools Emerging From Search Engines and Social Media Data for Determining Eye Disease Patterns. JAMA Ophthalmol 2016 Sep 01;134(9):1024-1030 [FREE Full text] [doi: 10.1001/jamaophthalmol.2016.2267] [Medline: 27416554]
29. DeVilbiss E, Lee B. Brief Report: Trends in U.S. National Autism Awareness from 2004 to 2014: The Impact of National Autism Awareness Month. Journal of Autism and Developmental Disorders 2014;44(12):3271-3273. [doi: 10.1007/s10803-014-2160-4] [Medline: 24915931]
30. Domnich A, Panatto D, Signori A, Lai PL, Gasparini R, Amicizia D. Age-related differences in the accuracy of web query-based predictions of influenza-like illness. PLoS One 2015;10(5):e0127754 [FREE Full text] [doi: 10.1371/journal.pone.0127754] [Medline: 26011418]
31. El-Sheikha J. Global search demand for varicose vein information on the internet. Phlebology 2015 Sep;30(8):533-540. [doi: 10.1177/0268355514542681] [Medline: 24993972]
32. Fenichel EP, Kuminoff NV, Chowell G. Skip the trip: air travelers' behavioral responses to pandemic influenza. PLoS One 2013 Mar;8(3):e58249 [FREE Full text] [doi: 10.1371/journal.pone.0058249] [Medline: 23526970]
33. Fond G, Gaman A, Brunel L, Haffen E, Llorca P. Google Trends ® : Ready for real-time suicide prevention or just a Zeta-Jones effect? An exploratory study. Psychiatry Research 2015 Aug;228(3):913-917. [doi: 10.1016/j.psychres.2015.04.022]
34. Foroughi F, Lam AK, Lim MS, Saremi N, Ahmadvand A. “Googling” for Cancer: An Infodemiological Assessment of Online Search Interests in Australia, Canada, New Zealand, the United Kingdom, and the United States. JMIR Cancer 2016 May 04;2(1):e5 [FREE Full text] [doi: 10.2196/cancer.5212] [Medline: 28410185]
35. Gafson AR, Giovannoni G. CCSVI-A. A call to clinicans and scientists to vocalise in an Internet age. Mult Scler Relat Disord 2014 Mar;3(2):143-146. [doi: 10.1016/j.msard.2013.10.005] [Medline: 25878001]

36. Gahr M, Uzelac Z, Zeiss R, Connemann BJ, Lang D, Schönfeldt-Lecuona C. Linking Annual Prescription Volume of Antidepressants to Corresponding Web Search Query Data: A Possible Proxy for Medical Prescription Behavior? J Clin Psychopharmacol 2015 Dec;35(6):681-685. [doi: 10.1097/JCP.0000000000000397] [Medline: 26355849]
37. Gamma A, Schleifer R, Weinmann W, Buadze A, Liebrenz M. Could Google Trends Be Used to Predict Methamphetamine-Related Crime? An Analysis of Search Volume Data in Switzerland, Germany, and Austria. PLoS ONE 2016 Nov 30;11(11):e0166566. [doi: 10.1371/journal.pone.0166566]
38. Garrison SR, Dormuth CR, Morrow RL, Carney GA, Khan KM. Seasonal effects on the occurrence of nocturnal leg cramps: a prospective cohort study. CMAJ 2015 Mar 03;187(4):248-253 [FREE Full text] [doi: 10.1503/cmaj.140497] [Medline: 25623650]
39. Gollust SE, Qin X, Wilcock AD, Baum LM, Barry CL, Niederdeppe J, et al. Search and You Shall Find: Geographic Characteristics Associated With Google Searches During the Affordable Care Act's First Enrollment Period. Med Care Res Rev 2017 Dec;74(6):723-735. [doi: 10.1177/1077558716660944] [Medline: 27457426]
40. Guernier V, Milinovich GJ, Bezerra SMA, Haworth M, Coleman G, Soares MRJ. Use of big data in the surveillance of veterinary diseases: early detection of tick paralysis in companion animals. Parasit Vectors 2016 Dec 23;9(1):303 [FREE Full text] [doi: 10.1186/s13071-016-1590-6] [Medline: 27215214]
41. Haney NM, Kinsella SD, Morey JM. United States medical school graduate interest in radiology residency programs as depicted by online search tools. J Am Coll Radiol 2014 Feb;11(2):193-197. [doi: 10.1016/j.jacr.2013.06.023] [Medline: 24120904]
42. Harorli OT, Harorli H. Evaluation of internet search trends of some common oral problems, 2004 to 2014. Community Dental Health 2014;31(3):188-192. [doi: 10.1922/CDH_3330Harorl?05]
43. Harsha AK, Schmitt JE, Stavropoulos SW. Know your market: use of online query tools to quantify trends in patient information-seeking behavior for varicose vein treatment. J Vasc Interv Radiol 2014 Jan;25(1):53-57. [doi: 10.1016/j.jvir.2013.09.015] [Medline: 24286941]
44. Harsha AK, Schmitt JE, Stavropoulos SW. Match day: online search trends reflect growing interest in IR training. J Vasc Interv Radiol 2015 Jan;26(1):95-100. [doi: 10.1016/j.jvir.2014.09.011] [Medline: 25541447]
45. Hassid BG, Day LW, Awad MA, Sewell JL, Osterberg EC, Breyer BN. Using Search Engine Query Data to Explore the Epidemiology of Common Gastrointestinal Symptoms. Dig Dis Sci 2017 Dec;62(3):588-592. [doi: 10.1007/s10620-016-4384-y] [Medline: 27878646]
46. Hossain L, Kam D, Kong F, Wigand RT, Bossomaier T. Social media in Ebola outbreak. Epidemiol Infect 2016 Jul;144(10):2136-2143. [doi: 10.1017/S095026881600039X] [Medline: 26939535]
47. Huang J, Zheng R, Emery S. Assessing the impact of the national smoking ban in indoor public places in china: evidence from quit smoking related online searches. PLoS One 2013 Jun;8(6):e65577 [FREE Full text] [doi: 10.1371/journal.pone.0065577] [Medline: 23776504]
48. Huesch M, Chetlen A, Segel J, Schetter S. Frequencies of Private Mentions and Sharing of Mammography and Breast Cancer Terms on Facebook: A Pilot Study. J Med Internet Res 2017 Jun 09;19(6):e201 [FREE Full text] [doi: 10.2196/jmir.7508] [Medline: 28600279]
49. Ingram DG, Plante DT. Seasonal trends in restless legs symptomatology: evidence from Internet search query data. Sleep Med 2013 Dec;14(12):1364-1368. [doi: 10.1016/j.sleep.2013.06.016] [Medline: 24152798]
50. Ingram DG, Matthews CK, Plante DT. Seasonal trends in sleep-disordered breathing: evidence from Internet search engine query data. Sleep Breath 2015 Mar;19(1):79-84. [doi: 10.1007/s11325-014-0965-1] [Medline: 24595717]
51. Jha S, Wang Z, Laucis N, Bhattacharyya T. Trends in Media Reports, Oral Bisphosphonate Prescriptions, and Hip Fractures 1996-2012: An Ecological Analysis. J Bone Miner Res 2015 Dec;30(12):2179-2187 [FREE Full text] [doi: 10.1002/jbmr.2565] [Medline: 26018247]

52. Johnson AK, Mehta SD. A comparison of Internet search trends and sexually transmitted infection rates using Google trends. Sex Transm Dis 2014 Jan;41(1):61-63. [doi: 10.1097/OLQ.0000000000000065] [Medline: 24326584]
53. Kadry B, Chu LF, Kadry B, Gammas D, Macario A. Analysis of 4999 online physician ratings indicates that most patients give physicians a favorable rating. J Med Internet Res 2011 Nov;13(4):e95 [FREE Full text] [doi: 10.2196/jmir.1960] [Medline: 22088924]
54. Kang M, Zhong H, He J, Rutherford S, Yang F. Using Google Trends for influenza surveillance in South China. PLoS One 2013;8(1):e55205 [FREE Full text] [doi: 10.1371/journal.pone.0055205] [Medline: 23372837]
55. Kang M, Song W, Choi S, Kim H, Ha H, Kim S, et al. Google unveils a glimpse of allergic rhinitis in the real world. Allergy 2015 Jan;70(1):124-128. [doi: 10.1111/all.12528] [Medline: 25280183]
56. Koburger N, Mergl R, Rummel-Kluge C, Ibelshäuser A, Meise U, Postuvan V, et al. Celebrity suicide on the railway network: Can one case trigger international effects? J Affect Disord 2015 Oct 01;185:38-46. [doi: 10.1016/j.jad.2015.06.037] [Medline: 26143403]
57. Kostkova P, Fowler D, Wiseman S, Weinberg JR. Major infection events over 5 years: how is media coverage influencing online information needs of health care professionals and the public? J Med Internet Res 2013 Jul 15;15(7):e107 [FREE Full text] [doi: 10.2196/jmir.2146] [Medline: 23856364]
58. Lawson MAC, Lawson MA, Kalff R, Walter J. Google Search Queries About Neurosurgical Topics: Are They a Suitable Guide for Neurosurgeons? World Neurosurg 2016 Jun;90:179-185. [doi: 10.1016/j.wneu.2016.02.045] [Medline: 26898496]
59. Leffler CT, Davenport B, Chan D. Frequency and seasonal variation of ophthalmology-related internet searches. Can J Ophthalmol 2010 Jun;45(3):274-279. [doi: 10.3129/i10-022] [Medline: 20436544]
60. Ling R, Lee J. Disease Monitoring and Health Campaign Evaluation Using Google Search Activities for HIV and AIDS, Stroke, Colorectal Cancer, and Marijuana Use in Canada: A Retrospective Observational Study. JMIR Public Health Surveill 2016 Oct 12;2(2):e156 [FREE Full text] [doi: 10.2196/publichealth.6504] [Medline: 27733330]
61. Linkov F, Bovbjerg DH, Freese KE, Ramanathan R, Eid GM, Gourash W. Bariatric surgery interest around the world: what Google Trends can teach us. Surg Obes Relat Dis 2014 May;10(3):533-538. [doi: 10.1016/j.soard.2013.10.007] [Medline: 24794184]
62. Liu F, Allan GM, Korownyk C, Kolber M, Flook N, Sternberg H, et al. Seasonality of Ankle Swelling: Population Symptom Reporting Using Google Trends. Ann Fam Med 2016 Dec;14(4):356-358 [FREE Full text] [doi: 10.1370/afm.1953] [Medline: 27401424]
63. Luckett T, Disler R, Hosie A, Johnson M, Davidson P, Currow D, et al. Content and quality of websites supporting self-management of chronic breathlessness in advanced illness: a systematic review. NPJ Prim Care Respir Med 2016 Dec 26;26:16025 [FREE Full text] [doi: 10.1038/npjpcrm.2016.25] [Medline: 27225898]
64. Majumder MS, Santillana M, Mekaru SR, McGinnis DP, Khan K, Brownstein JS. 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 2016 Jun 01;2(1):e30 [FREE Full text] [doi: 10.2196/publichealth.5814] [Medline: 27251981]
65. Mattin MJ, Solano-Gallego L, Dhollander S, Afonso A, Brodbelt DC. The frequency and distribution of canine leishmaniosis diagnosed by veterinary practitioners in Europe. Vet J 2014 Jun;200(3):410-419. [doi: 10.1016/j.tvjl.2014.03.033] [Medline: 24767097]
66. Mavragani A, Sypsa K, Sampri A, Tsagarakis K. Quantifying the UK Online Interest in Substances of the EU Watchlist for Water Monitoring: Diclofenac, Estradiol, and the Macrolide Antibiotics. Water 2016 Nov 18;8(11):542. [doi: 10.3390/w8110542]
67. Murray G, O'Rourke C, Hogan J, Fenton JE. Detecting internet search activity for mouth cancer in Ireland. Br J Oral Maxillofac Surg 2016 Feb;54(2):163-165. [doi: 10.1016/j.bjoms.2015.12.005] [Medline: 26774361]

68. Myers L, Jones J, Boesten N, Lancman M. Psychogenic non-epileptic seizures (PNES) on the Internet: Online representation of the disorder and frequency of search terms. Seizure 2016 Aug 01;40:114-122 [FREE Full text]
69. Noar S, Ribisl K, Althouse B, Willoughby J, Ayers J. Using digital surveillance to examine the impact of public figure pancreatic cancer announcements on media and search query outcomes. Journal of the National Cancer Institute - Monographs 2013:188-194.
70. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of google trends in health care research: a systematic review. PLoS One 2014 Oct;9(10):e109583 [FREE Full text] [doi: 10.1371/journal.pone.0109583] [Medline: 25337815]
71. Pandey A, Abdullah K, Drazner MH. Impact of Vice President Cheney on public interest in left ventricular assist devices and heart transplantation. Am J Cardiol 2014 May 01;113(9):1529-1531. [doi: 10.1016/j.amjcard.2014.02.007] [Medline: 24630787]
72. Parker J, Cuthbertson C, Loveridge S, Skidmore M, Dyar W. Forecasting state-level premature deaths from alcohol, drugs, and suicides using Google Trends data. J Affect Disord 2017 Dec 15;213:9-15. [doi: 10.1016/j.jad.2016.10.038] [Medline: 28171770]
73. Phelan N, Kelly JC, Moore DP, Kenny P. The effect of the metal-on-metal hip controversy on Internet search activity. Eur J Orthop Surg Traumatol 2014 Oct;24(7):1203-1210. [doi: 10.1007/s00590-013-1399-3] [Medline: 24390041]
74. Phelan N, Davy S, O'Keeffe GW, Barry DS. Googling in anatomy education: Can google trends inform educators of national online search patterns of anatomical syllabi? Anat Sci Educ 2017 Mar;10(2):152-159. [doi: 10.1002/ase.1641] [Medline: 27547967]
75. Plante DT, Ingram DG. Seasonal trends in tinnitus symptomatology: evidence from Internet search engine query data. Eur Arch Otorhinolaryngol 2015 Oct;272(10):2807-2813. [doi: 10.1007/s00405-014-3287-9] [Medline: 25234771]
76. Poletto C, Boëlle P, Colizza V. Risk of MERS importation and onward transmission: a systematic review and analysis of cases reported to WHO. BMC Infect Dis 2016 Aug 25;16(1):448 [FREE Full text] [doi: 10.1186/s12879-016-1787-5] [Medline: 27562369]
77. Pollett S, Wood N, Boscardin WJ, Bengtsson H, Schwarcz S, Harriman K, et al. Validating the Use of Google Trends to Enhance Pertussis Surveillance in California. PLoS Curr 2015 Oct 19:1-10. [doi: 10.1371/currents.outbreaks.7119696b3e7523faa4543faac87c56c2]
78. Rohart F, Milinovich GJ, Avril SMR, Lê CK, Tong S, Hu W. Disease surveillance based on Internet-based linear models: an Australian case study of previously unmodeled infection diseases. Sci Rep 2016 Dec 20;6:38522 [FREE Full text] [doi: 10.1038/srep38522] [Medline: 27994231]
79. Rosenkrantz AB, Prabhu V. Public Interest in Imaging-Based Cancer Screening Examinations in the United States: Analysis Using a Web-Based Search Tool. AJR Am J Roentgenol 2016 Jan;206(1):113-118. [doi: 10.2214/AJR.15.14840] [Medline: 26700342]
80. Rossignol L, Pelat C, Lambert B, Flahault A, Chartier-Kastler E, Hanslik T. A method to assess seasonality of urinary tract infections based on medication sales and google trends. PLoS One 2013;8(10):e76020 [FREE Full text] [doi: 10.1371/journal.pone.0076020] [Medline: 24204587]
81. Scatà M, Di SA, Liò P, La CA. The Impact of Heterogeneity and Awareness in Modeling Epidemic Spreading on Multiplex Networks. Sci Rep 2016 Dec 16;6:37105 [FREE Full text] [doi: 10.1038/srep37105] [Medline: 27848978]
82. Scheres LJJ, Lijfering WM, Middeldorp S, Cannegieter SC. Influence of World Thrombosis Day on digital information seeking on venous thrombosis: a Google Trends study. J Thromb Haemost 2016 Dec;14(12):2325-2328. [doi: 10.1111/jth.13529] [Medline: 27735128]
83. Shin S, Seo D, An J, Kwak H, Kim S, Gwack J, et al. High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. Sci Rep 2016 Sep 06;6:32920 [FREE Full text] [doi: 10.1038/srep32920] [Medline: 27595921]
84. Schootman M, Toor A, Cavazos-Rehg P, Jeffe DB, McQueen A, Eberth J, et al. The utility of Google Trends data to examine interest in cancer screening. BMJ Open 2015 Jun 08;5(6):e006678 [FREE Full text] [doi: 10.1136/bmjopen-2014-006678] [Medline: 26056120]

85. Schuster N, Rogers M, McMahon JL. Using search engine query data to track pharmaceutical utilization: a study of statins. The American journal of managed care 2010;16(8):215-219.
86. Seifter A, Schwarzwalder A, Geis K, Aucott J. The utility of "Google Trends" for epidemiological research: Lyme disease as an example. Geospat Health 2010 May;4(2):135-137. [doi: 10.4081/gh.2010.195] [Medline: 20503183]
87. Sentana-Lledo D, Barbu CM, Ngo MN, Wu Y, Sethuraman K, Levy MZ. Seasons, Searches, and Intentions: What The Internet Can Tell Us About The Bed Bug (Hemiptera: Cimicidae) Epidemic. J Med Entomol 2016 Jan;53(1):116-121. [doi: 10.1093/jme/tjv158] [Medline: 26474879]
88. Simmering JE, Polgreen LA, Polgreen PM. Web search query volume as a measure of pharmaceutical utilization and changes in prescribing patterns. Res Social Adm Pharm 2014;10(6):896-903. [doi: 10.1016/j.sapharm.2014.01.003] [Medline: 24603135]
89. Skeldon SC, Kozhimannil KB, Majumdar SR, Law MR. The effect of competing direct-to-consumer advertising campaigns on the use of drugs for benign prostatic hyperplasia: time series analysis. J Gen Intern Med 2015 Apr;30(4):514-520 [FREE Full text] [doi: 10.1007/s11606-014-3063-y] [Medline: 25338730]
90. Solano P, Ustulin M, Pizzorno E, Vichi M, Pompili M, Serafini G, et al. A Google-based approach for monitoring suicide risk. Psychiatry Res 2016 Dec 30;246:581-586. [doi: 10.1016/j.psychres.2016.10.030] [Medline: 27837725]
91. Stein JD, Childers DM, Nan B, Mian SI. Gauging interest of the general public in laser-assisted in situ keratomileusis eye surgery. Cornea 2013 Jul;32(7):1015-1018 [FREE Full text] [doi: 10.1097/ICO.0b013e318283c85a] [Medline: 23538615]
92. Takada K. Japanese Interest in "Hotaru" (Fireflies) and "Kabuto-Mushi" (Japanese Rhinoceros Beetles) Corresponds with Seasonality in Visible Abundance. Insects 2012 Apr 10;3(2):424-431 [FREE Full text] [doi: 10.3390/insects3020424] [Medline: 26466535]
93. Telfer S, Woodburn J. Let me Google that for you: a time series analysis of seasonality in internet search trends for terms related to foot and ankle pain. J Foot Ankle Res 2015 Jul;8:27 [FREE Full text] [doi: 10.1186/s13047-015-0074-9] [Medline: 26146521]
94. Troelstra SA, Bosdriesz JR, de BMR, Kunst AE. Effect of Tobacco Control Policies on Information Seeking for Smoking Cessation in the Netherlands: A Google Trends Study. PLoS One 2016 Feb;11(2):e0148489 [FREE Full text] [doi: 10.1371/journal.pone.0148489] [Medline: 26849567]
95. Toosi B, Kalia S. Seasonal and Geographic Patterns in Tanning Using Real-Time Data From Google Trends. JAMA Dermatol 2016 Feb;152(2):215-217. [doi: 10.1001/jamadermatol.2015.3008] [Medline: 26719968]
96. Wang H, Chen D, Yu H, Chen Y. Forecasting the Incidence of Dementia and Dementia-Related Outpatient Visits With Google Trends: Evidence From Taiwan. J Med Internet Res 2015 Nov 19;17(11):e264 [FREE Full text] [doi: 10.2196/jmir.4516] [Medline: 26586281]
97. Warren KE, Wen LS. Measles, social media and surveillance in Baltimore City. J Public Health (Oxf) 2017 Sep 01;39(3):e73-e78. [doi: 10.1093/pubmed/fdw076] [Medline: 27521926]
98. Willson TJ, Lospinoso J, Weitzel E, McMains K. Correlating Regional Aeroallergen Effects on Internet Search Activity. Otolaryngol Head Neck Surg 2014 Dec 12;152(2):228-232. [doi: 10.1177/0194599814560149] [Medline: 25505261]
99. Willson TJ, Shams A, Lospinoso J, Weitzel E, McMains K. Searching for Cedar: Geographic Variation in Single Aeroallergen Shows Dose Response in Internet Search Activity. Otolaryngol Head Neck Surg 2015 Nov 02;153(5):770-774. [doi: 10.1177/0194599815601650] [Medline: 26340925]
100. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. PNAS 2015;112(47):14473.
101. Zhang Z, Zheng X, Zeng DD, Leischow SJ. Information seeking regarding tobacco and lung cancer: effects of seasonality. PLoS One 2015 Mar;10(3):e0117938 [FREE Full text] [doi: 10.1371/journal.pone.0117938] [Medline: 25781020]

102. Zhang Z, Zheng X, Zeng DD, Leischow SJ. Tracking Dabbing Using Search Query Surveillance: A Case Study in the United States. J Med Internet Res 2016 Sep 16;18(9):e252 [FREE Full text] [doi: 10.2196/jmir.5802] [Medline: 27637361]
103. Zheluk A, Quinn C, Meylakhs P. Internet search and krokodil in the Russian Federation: an infoveillance study. J Med Internet Res 2014 Sep 18;16(9):e212 [FREE Full text] [doi: 10.2196/jmir.3203] [Medline: 25236385]
104. Zhou X, Ye J, Feng Y. Tuberculosis surveillance by analyzing Google trends. IEEE Trans Biomed Eng 2011 Aug;58(8):2247-2254. [doi: 10.1109/TBME.2011.2132132] [Medline: 21435969]

Multimedia Appendix

Google Trends in Infodemiology and Infoveillance: Methodology Framework

Amaryllis Mavragani and Gabriela Ochoa

The available categories in Google Trends are listed in Table A1. The available subcategories (2nd level of categorization) of "Health" and all available subcategories of these subcategories (3rd and 4th level) are listed in Table A2.

Table A1. Available Category Selection in Google Trends

Arts & Entertainment	Health	People & Society
Autos & Vehicles	Hobbies & Leisure	Pets & Animals
Beauty & Fitness	Home & Garden	Property
Books & Literature	Internet & Telecom	Reference
Business & Industrial	Jobs & Education	Science
Computers & Electronics	Law & Government	Shopping
Finance	News	Sports
Food & Drink	Online Communities	Travel
Games		

Table A2. All levels of Available Categories and Subcategories of the "Health" Category

2nd level	3rd level	4th level
Ageing & Geriatrics ⇨	Alzheimer's Disease	
Alternative & Natural Medicine ⇨	Acupuncture & Chinese Medicine	
	Cleansing & Detoxification	
Health Conditions ⇨	AIDS & HIV	
	Allergies	
	Arthritis	
	Cancer	
	Cold & Flu	
	Diabetes	
	Ear, Noise & Throat	
	Eating Disorders	
	Endocrine Conditions ⇨	Diabetes Thyroid Conditions
	Genetic Disorders	
	GERD & Digestive Disorders	
	Heart and Hypertension	
	Infectious Diseases ⇨	Cold & Flu Parasites & Parasitic Diseases Sexually Transmitted Diseases Vaccines & Immunization
	Injury	
	Neurological Disorders ⇨	Alzheimer's Disease
	Obesity	
	Pain Management ⇩	Headaches & Migraines
	Respiratory Conditions ⇩	Asthma
	Skin Conditions	
	Sleep Disorders	
Health Education & Medical Training		
Health Foundations & Med. Research		
Health News ⇩	Health Policy	
Medical Devices & Equipment ⇩	Assistive Technology	
Medical Facilities & Services ⇩	Doctor's Offices	
	Hospitals & Treatment Centers	
	Medical Procedures	
	Physical Therapy	
Medical Literature & Resources ⇩	Medical Photos & Illustrations	
Men's Health ⇩	Erectile Dysfunction	
Mental Health ⇩	Anxiety & Stress	
	Depression	
	Learning & Development Disabilities ⇩	ADD & ADHD
Nursing ⇩	Assisted Living & Long Term Care	
Nutrition ⇩	Special & Restricted Diets ⇩	Cholesterol Issues
	Vitamins & Supplements	
Oral & Dental Care		
Pediatrics		
Pharmacy ⇩	Drugs & Medications	
Public Health ⇩	Health Policy	
	Occupational Health & Safety	
	Poisons & Overdoses	
	Vaccines & Immunisations	
Reproductive Health ⇩	Birth Control	
	Erectile Dysfunction	
	Infertility	
	OBGYN ⇩	Pregnancy & Maternity
	Sex Education & Counseling	
	Sexual Enhancement	
	Sexually Transmitted Diseases ⇩	AIDS & HIV
Substance Abuse ⇩	Drug & Alcohol Testing	
	Drug & Alcohol Treatment	
	Smoking & Smoking Cessation	
	Steroids & Performance-Enhancing Drugs	
Vision Care ⇩	Eye Glasses & Contacts	
Women's Health ⇩	OBGYN ⇩	Pregnancy & Maternity

Multimedia Appendix 1: State data tables

Table A1 consists of the States by declining interest in the term 'Asthma' from 2004 to 2015, Table A2 consists of the normalized values for the online interest by State for each year from 2004 to 2015, and Tables A3 and A4 present the smoothing parameters and coefficients for the Holt-Winters forecasting by State, respectively.

Table A1. Online Interest for 'Asthma' by State in USA from January 2004 to December 2015

State	Score	State	Score	State	Score	State	Score
Delaware	100	South Carolina	89	Massachusetts	81	Arizona	78
West Virginia	97	Indiana	88	Oklahoma	81	Wisconsin	77
North Carolina	95	Colorado	87	New Jersey	81	Texas	76
Kentucky	95	Alabama	87	Montana	80	Hawaii	73
Maine	94	Georgia	86	Washington	80	Utah	72
Tennessee	93	South Dakota	86	New York	80	Iowa	72
Connecticut	92	New Mexico	85	Kansas	79	Louisiana	71
Maryland	92	Vermont	85	Alaska	79	Florida	69
Mississippi	92	Missouri	85	New Hampshire	78	Nevada	67
Pennsylvania	91	Minnesota	84	Ohio	78	California	67
Nebraska	91	Arkansas	84	North Dakota	78	Virginia	61
Idaho	89	District of Columbia	83	Wyoming	78	Oregon	55
Rhode Island	89	Michigan	83	Illinois	78		

Table A2. Online Interest for 'Asthma' by State by Year from 2004 to 2015

State	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Alabama	38	58	76	50	52	73	80	83	90	77	82	88
Alaska	67	73	100	52	44	77	70	88	86	86	79	93
Arizona	48	52	58	39	47	67	64	72	83	66	70	78
Arkansas	45	76	80	51	54	63	74	85	82	80	78	79
California	40	45	53	35	36	44	37	61	65	56	57	63
Colorado	48	54	62	52	53	73	75	83	87	73	72	81
Connecticut	54	63	76	48	59	81	75	93	91	78	87	90
Delaware	78	79	95	100	100	100	82	93	95	100	100	100
DC	52	52	68	43	45	68	64	78	82	71	74	78
Florida	37	41	49	34	40	57	55	65	68	56	61	65
Georgia	48	54	67	50	51	73	68	83	89	73	71	79
Hawaii	58	62	52	43	48	63	67	71	80	62	64	66
Idaho	81	73	65	52	58	89	77	95	97	80	87	85
Illinois	52	56	67	42	49	69	64	69	75	64	66	72
Indiana	59	69	82	45	51	77	71	81	88	70	76	78
Iowa	55	62	68	43	47	75	73	73	75	61	60	67
Kansas	51	57	68	47	48	64	77	79	85	69	67	62
Kentucky	67	72	79	52	64	94	80	96	98	83	83	94
Louisiana	64	49	63	40	44	58	57	69	71	59	61	63
Maine	58	83	88	66	56	83	83	99	100	89	93	87
Maryland	53	58	78	52	55	85	82	82	85	74	77	91
Massachusetts	57	55	69	45	54	73	67	77	79	68	71	80
Michigan	44	59	72	49	53	73	68	79	76	68	68	77
Minnesota	51	62	74	47	52	75	78	78	81	69	72	78
Mississippi	54	83	68	41	45	65	64	84	92	75	79	99
Missouri	56	59	71	46	52	75	73	80	80	68	69	81
Montana	49	62	81	56	47	81	67	95	89	81	77	81
Nebraska	78	73	79	49	55	87	80	87	98	79	85	86
Nevada	39	42	51	35	44	64	59	60	65	65	61	68
New Hampshire	60	49	78	41	50	69	68	79	79	67	67	82
New Jersey	51	56	66	41	47	67	63	76	78	69	71	75
New Mexico	75	69	64	47	61	79	95	76	86	77	74	85
New York	49	57	65	46	49	69	64	70	77	66	65	72
North Carolina	55	63	79	51	62	85	82	91	95	82	80	82
North Dakota	91	91	94	54	69	99	79	86	85	84	73	74
Ohio	48	54	68	44	46	68	68	73	78	66	67	72
Oklahoma	46	64	63	46	48	62	71	77	87	72	71	74
Oregon	48	50	68	43	50	71	69	74	79	46	34	38
Pennsylvania	52	61	74	48	55	75	75	88	91	78	77	84
Rhode Island	64	49	80	52	61	72	84	94	93	85	84	89
South Carolina	61	54	73	46	51	82	70	77	89	80	78	79
South Dakota	100	97	89	57	64	73	71	88	96	70	80	98
Tennessee	48	58	70	45	48	76	76	87	96	78	81	88
Texas	45	52	64	41	47	64	60	70	73	63	66	70
Utah	43	52	66	37	42	53	59	70	69	65	66	67
Vermont	63	95	74	67	44	70	55	93	85	91	76	92
Virginia	40	48	63	40	35	37	33	76	82	67	55	53
Washington	46	52	65	46	51	70	66	78	77	70	68	74
West Virginia	86	81	67	75	69	95	100	100	98	94	81	93
Wisconsin	57	61	71	43	50	67	71	74	73	64	63	68
Wyoming	85	100	88	50	55	52	84	91	86	90	78	80

Table A3. Smoothing Parameters for the Holt-Winters' Forecasting by State

State	alpha (α)	beta (β*)	gamma (γ)
Alabama	0.1050	0.0162	0.7013
Alaska	0.0491	0.0877	0.4980
Arizona	0.1491	0.0381	0.4310
Arkansas	0.0426	0.0882	0.6637
California	0.2817	0.0000	0.6552
Colorado	0.0685	0.0000	0.3896
Connecticut	0.1059	0.0000	0.4359
Delaware	0.1665	0.0208	0.4038
District of Columbia	0.0923	0.0000	0.5157
Florida	0.1447	0.0195	0.6686
Georgia	0.2049	0.0321	0.4613
Hawaii	0.0146	0.6730	0.6628
Idaho	0.0090	1.0000	0.5640
Illinois	0.1829	0.0274	0.4946
Indiana	0.0525	0.0334	0.5058
Iowa	0.1160	0.0540	0.5406
Kansas	0.0701	0.0122	0.4871
Kentucky	0.0604	0.0494	0.4764
Louisiana	0.0893	0.1310	0.4797
Maine	0.0274	0.1580	0.3866
Maryland	0.0263	0.0000	0.4569
Massachusetts	0.0841	0.0554	0.5767
Michigan	0.1009	0.0239	0.6195
Minnesota	0.1291	0.0088	0.5211
Mississippi	0.1009	0.0000	0.6027
Missouri	0.0972	0.0000	0.5170
Montana	0.0954	0.0783	0.3634
Nebraska	0.0566	0.1721	0.2803
Nevada	0.0321	0.0490	0.6138
New Hampshire	0.0902	0.0775	0.4581
New Jersey	0.1657	0.0260	0.4642
New Mexico	0.1470	0.0072	0.3644
New York	0.1818	0.0245	0.5085
North Carolina	0.1209	0.0098	0.4294
North Dakota	0.0459	0.0485	0.4518
Ohio	0.0997	0.1281	0.3487
Oklahoma	0.2460	0.0389	0.5948
Oregon	0.1868	0.0000	0.5835
Pennsylvania	0.1784	0.0215	0.5000
Rhode Island	0.1205	0.0124	0.5135
South Carolina	0.0805	0.0851	0.2857
South Dakota	0.1044	0.0957	0.4531
Tennessee	0.0779	0.0372	0.5314
Texas	0.1822	0.0000	0.4703
Utah	0.0448	0.0430	0.4546
Vermont	0.0945	0.0330	0.4606
Virginia	0.5253	0.0000	0.8483
Washington	0.1281	0.0209	0.4646
West Virginia	0.0755	0.0453	0.4634
Wisconsin	0.1103	0.0241	0.6439
Wyoming	0.0826	0.1182	0.4289

Table A4. Coefficients for the Holt-Winters' Forecastings by State

	a	b	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12
AL	27.42	-0.03	6.89	8.01	7.80	8.62	3.85	5.88	-0.34	2.91	5.36	12.26	2.44	4.88
AK	40.03	0.25	18.27	5.19	1.36	12.3	22.14	6.02	5.59	5.90	8.89	8.78	13.97	3.01
AZ	69.21	0.15	10.48	19.51	9.27	14.7	0.40	-3.72	-10.33	-0.01	-2.12	8.94	5.75	6.45
AR	32.11	0.08	11.38	10.28	8.23	11.0	2.22	-1.70	2.75	-1.77	14.8	7.41	8.68	1.56
CA	67.85	0.11	4.71	1.43	4.29	4.19	2.15	-10.63	-13.96	-12.67	-5.53	-1.71	-0.40	-3.70
CO	54.69	-0.08	-2.42	4.61	8.37	9.81	4.61	-0.35	-1.59	1.27	3.21	5.99	0.83	-2.45
CT	58.75	0.01	-2.08	0.88	0.08	7.65	0.51	-3.06	-8.06	-7.41	3.88	1.24	-1.54	-4.23
DC	58.46	-0.09	-2.35	-3.71	0.45	4.03	7.67	-6.60	-6.42	-3.51	1.25	2.11	-4.05	2.25
DE	21.62	-0.15	3.10	6.64	7.49	4.81	6.08	3.14	4.57	3.89	7.41	9.57	4.92	8.63
FL	70.72	0.06	-3.60	-0.12	-0.05	-2.19	-6.04	-8.45	-8.75	-10.65	-4.73	-0.48	-1.97	-7.95
GA	71.98	0.10	-5.87	-3.70	6.21	11.0	-4.32	-10.44	-9.74	-6.82	-1.93	-0.18	-1.72	-5.00
HI	15.87	0.06	4.97	9.97	10.18	9.70	2.19	6.44	2.52	0.43	8.84	7.21	2.48	5.21
ID	28.41	0.01	2.77	1.29	5.85	5.20	3.58	2.32	3.66	6.95	8.46	9.85	4.85	2.67
IL	55.00	0.00	0.57	1.09	2.85	7.97	4.34	-2.93	-2.40	-2.49	4.16	10.01	5.10	-1.56
IN	60.80	-0.02	-2.92	-5.85	0.56	6.84	-1.90	-9.62	-13.63	-6.00	-0.92	2.18	-1.71	-6.39
IA	35.73	0.06	-8.69	-6.07	1.02	-1.72	-2.74	-8.36	-8.67	-7.87	-5.42	-0.24	-6.21	-6.19
KS	29.79	-0.13	2.03	5.58	6.96	7.19	4.50	4.38	-0.43	6.63	6.57	10.74	0.60	-7.72
KY	39.66	0.04	5.51	3.14	4.30	5.31	6.17	2.29	4.65	2.76	9.29	6.25	6.65	0.89
LA	38.63	-0.05	8.32	9.87	11.57	6.96	2.77	0.58	-4.60	3.40	1.11	10.80	8.72	-0.06
ME	52.14	0.12	7.53	5.14	17.53	11.5	5.48	1.57	-5.89	-1.72	9.17	12.08	13.91	7.92
MD	64.07	0.01	-4.38	-2.99	-1.80	7.94	1.47	-10.15	-5.29	-8.14	2.58	5.83	2.85	-3.81
MA	45.67	0.05	2.61	3.99	5.96	6.61	8.90	-1.26	-3.03	-1.90	7.41	6.26	6.35	3.95
MI	64.56	0.04	-8.75	-3.33	0.59	2.26	-1.66	-6.55	-13.19	-11.45	-2.61	-0.14	-4.01	-4.58
MN	56.21	-0.19	4.86	8.46	4.26	11.8	8.08	-1.28	-3.49	-1.41	7.16	11.42	7.37	5.21
MS	43.92	-0.22	-0.27	2.54	11.47	14.9	3.34	-5.71	0.86	-1.29	1.73	15.53	4.36	-0.29
MO	61.87	0.01	0.02	-1.09	-3.40	4.83	-0.57	-8.90	-9.70	-7.99	-1.50	2.06	-1.30	-3.40
MT	41.50	-0.04	0.46	7.20	10.72	4.97	6.30	6.63	2.03	-1.68	10.6	5.85	2.58	1.58
NE	41.78	0.18	-2.37	2.35	3.91	9.18	3.38	-1.05	2.72	3.74	3.91	6.07	-0.02	-2.02
NV	35.65	0.08	13.62	21.59	19.62	14.7	11.32	5.52	-2.01	2.55	3.11	5.61	8.26	2.12
NH	39.74	0.20	9.90	3.84	8.49	8.16	11.15	-5.53	0.35	-0.09	7.76	18.00	6.88	2.66
NJ	73.25	0.10	-3.38	-4.18	-1.22	7.08	3.99	-9.22	-10.01	-16.13	-4.15	5.06	-1.75	-3.62
NM	51.18	0.07	-0.53	1.35	1.93	1.52	-5.30	-6.87	-11.80	-8.40	-6.07	3.24	1.32	-9.66
NY	57.21	0.00	3.08	3.44	5.12	7.45	7.69	-2.12	-7.03	-9.90	0.23	7.17	6.13	4.04
NC	74.44	0.04	-5.21	-2.60	1.13	5.15	-5.03	-9.10	-9.61	-10.47	-2.49	0.94	0.77	-5.73
ND	8.29	-0.05	6.22	0.05	4.50	5.17	3.01	0.71	3.06	6.29	1.92	2.27	7.73	4.79
OH	50.09	0.05	-0.61	1.45	4.15	10.0	2.10	-2.71	-2.57	-2.66	2.57	6.42	3.27	-1.41
OK	47.43	0.13	7.67	9.41	7.08	9.96	2.45	3.14	3.14	7.37	12.1	11.51	7.66	2.27
OR	41.00	-0.23	-1.71	0.42	6.73	3.92	11.57	11.82	-0.56	-0.65	6.15	2.78	0.88	-3.18
PA	66.30	0.10	-4.80	-2.94	-1.10	4.99	1.12	-6.37	-8.97	-9.78	-0.76	5.71	-0.21	-6.30
RI	52.70	0.08	-14.61	-5.99	-2.57	4.28	-3.41	-13.80	-10.98	-16.87	-6.07	1.10	-9.84	-10.82
SC	41.38	-0.08	2.31	5.67	6.36	8.23	2.60	-3.13	-5.09	-4.33	1.37	5.68	3.29	0.66
SD	13.80	-0.03	2.91	3.66	9.35	7.16	4.73	4.23	1.56	2.18	2.26	3.29	3.51	-0.33
TN	65.42	0.08	-11.31	-6.03	-6.32	8.95	-7.26	-10.38	-13.09	-6.63	-0.82	-6.65	-6.78	-9.28
TX	73.10	-0.10	4.56	2.22	4.96	13.6	1.65	-9.83	-10.45	-6.94	2.25	6.22	4.11	-2.04
UT	56.92	0.05	-3.00	-5.14	3.93	-0.90	-4.96	-6.57	-11.50	-7.32	-3.27	1.85	-6.75	-8.76
VT	29.90	-0.04	12.23	10.60	0.83	11.5	6.20	0.49	-1.66	-3.09	5.08	11.63	20.28	5.04
VA	77.71	-0.32	-2.12	-0.68	5.15	12.8	7.04	-3.16	2.38	-1.00	7.60	3.32	-0.06	-0.78
WA	72.27	0.03	-3.31	0.71	4.07	5.81	2.07	-3.52	-8.45	-11.27	-5.87	-2.82	-6.01	-11.25
WV	35.61	0.08	3.74	8.13	14.29	13.0	5.31	4.06	5.54	2.53	7.07	11.44	9.41	6.47
WI	37.06	-0.09	3.57	6.21	2.00	13.5	3.93	0.47	0.40	-3.24	2.58	14.99	7.85	2.24
WY	13.31	0.07	9.10	5.16	7.87	6.86	3.00	3.92	3.04	6.00	1.38	8.21	5.36	4.46

Multimedia Appendix 2: 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) by US State.

Figures B1 to B51 depict the changes in online interest in the term “asthma” from 2004 to 2015 and forecasts from 2016 to 2020 in each US State (and DC) in alphabetical order.

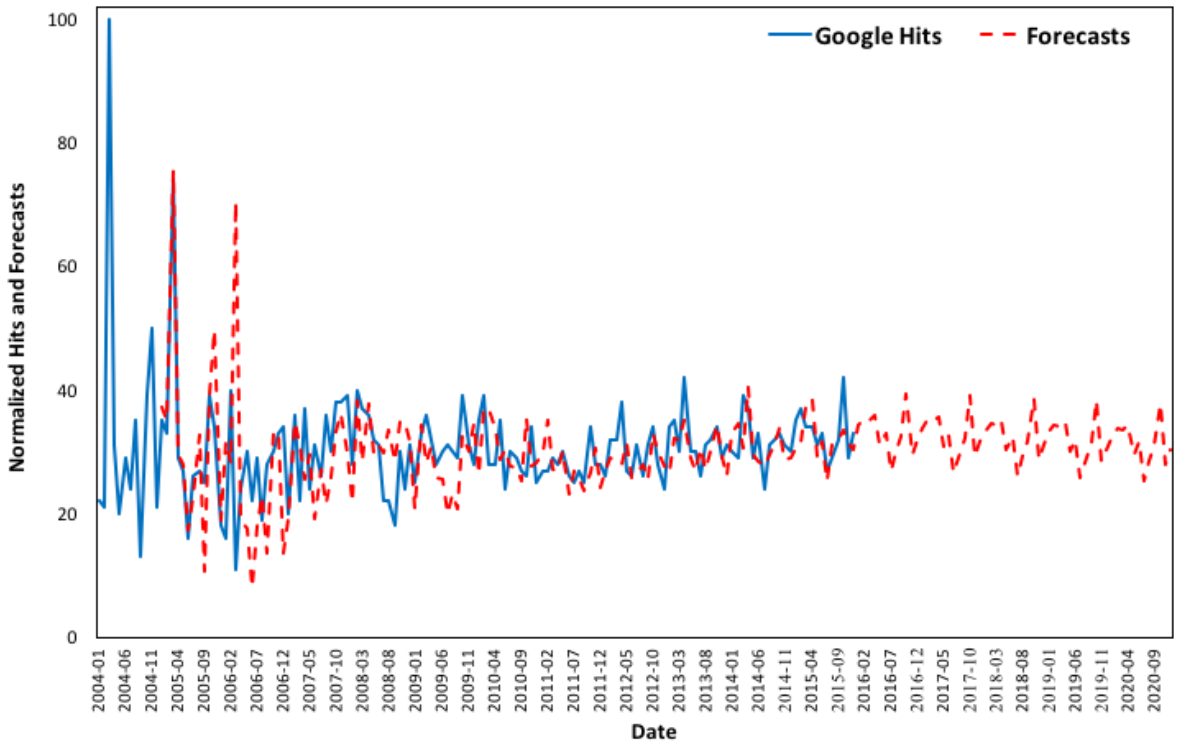


Figure B1. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Alabama.

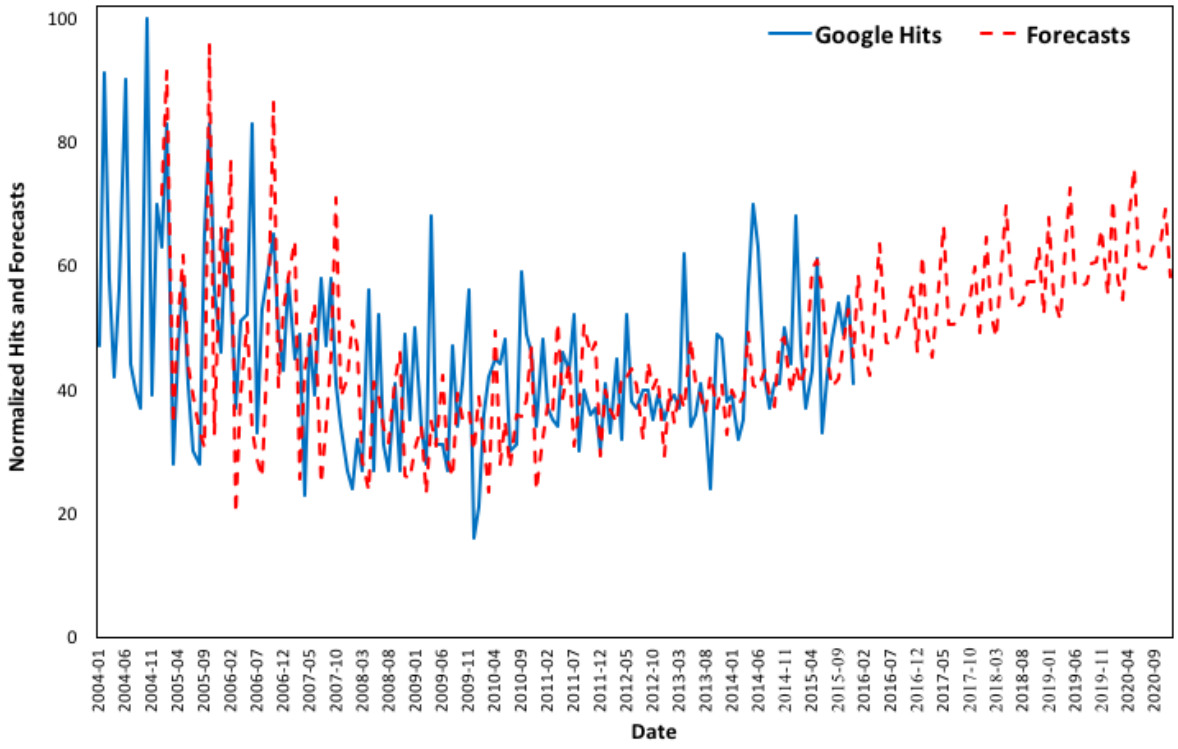


Figure B2. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Alaska.

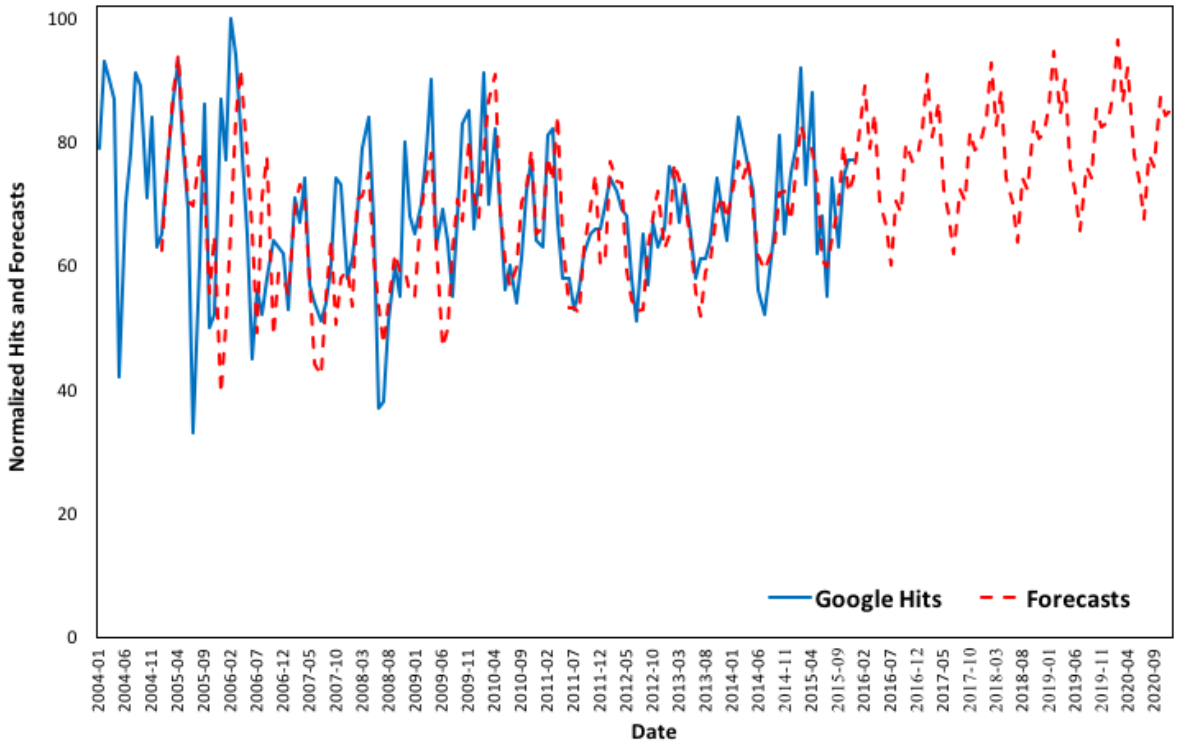


Figure B3. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Arizona.

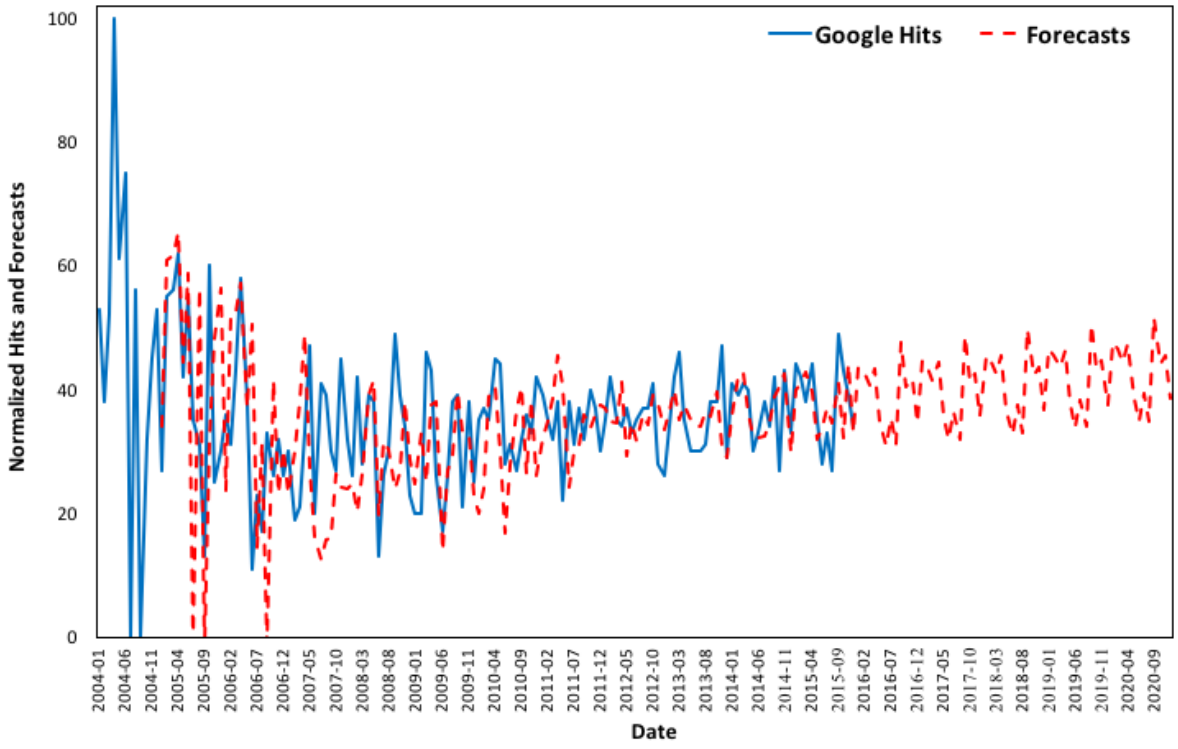


Figure B4. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Arkansas.

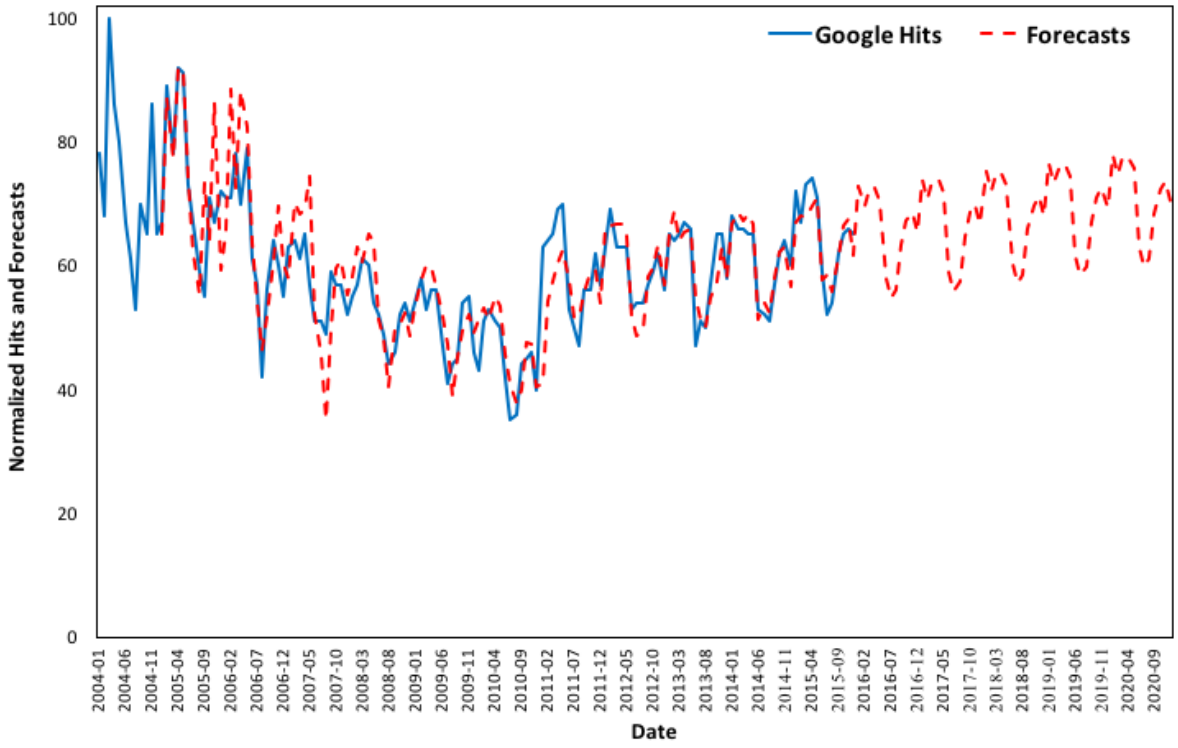


Figure B5. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in California.

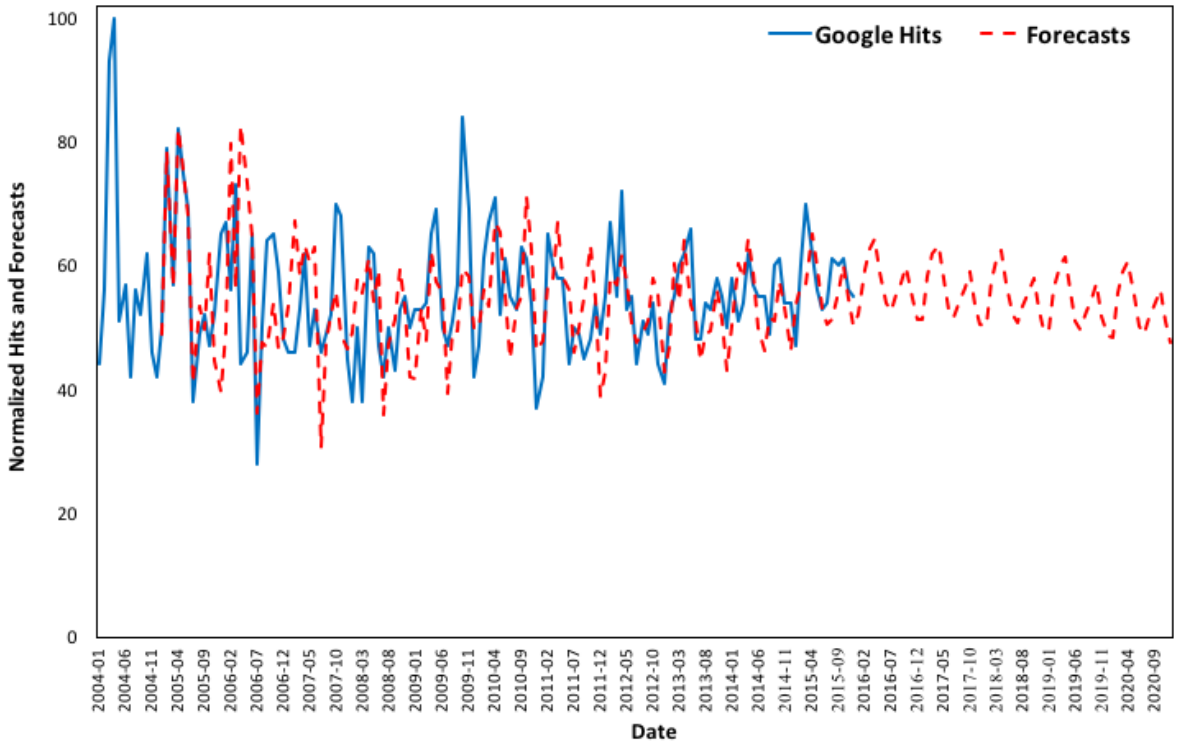


Figure B6. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Colorado.

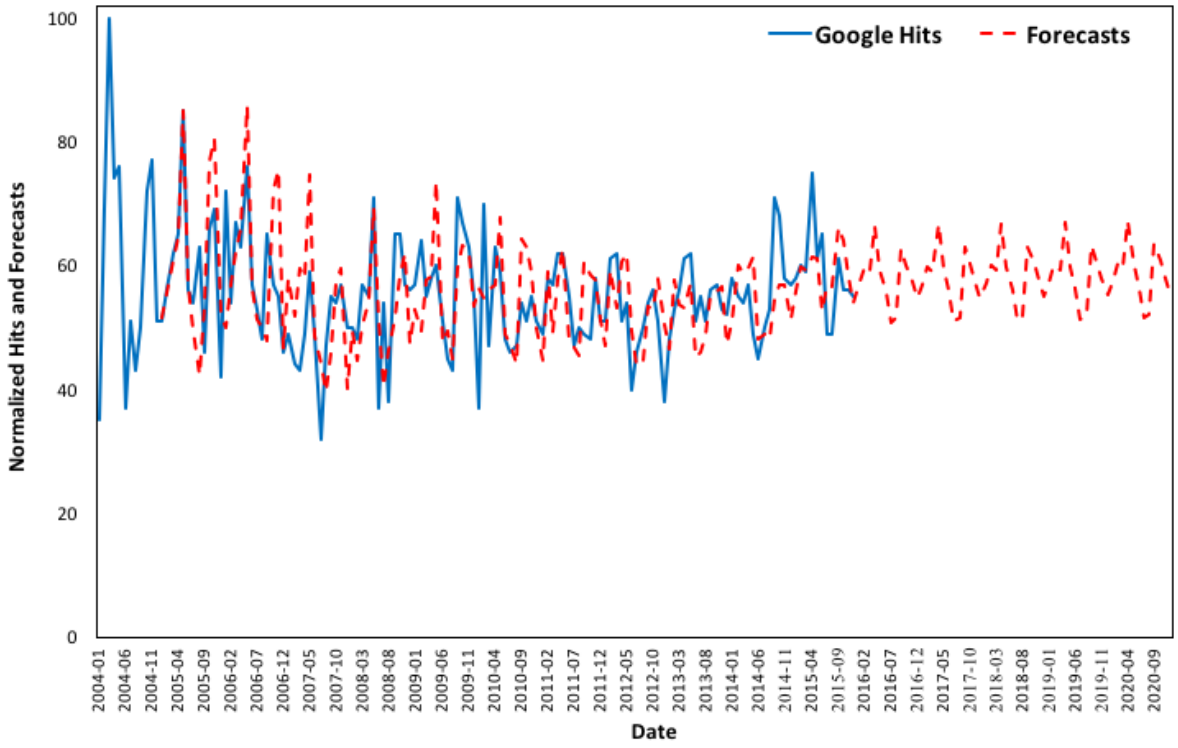


Figure B7. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Connecticut.

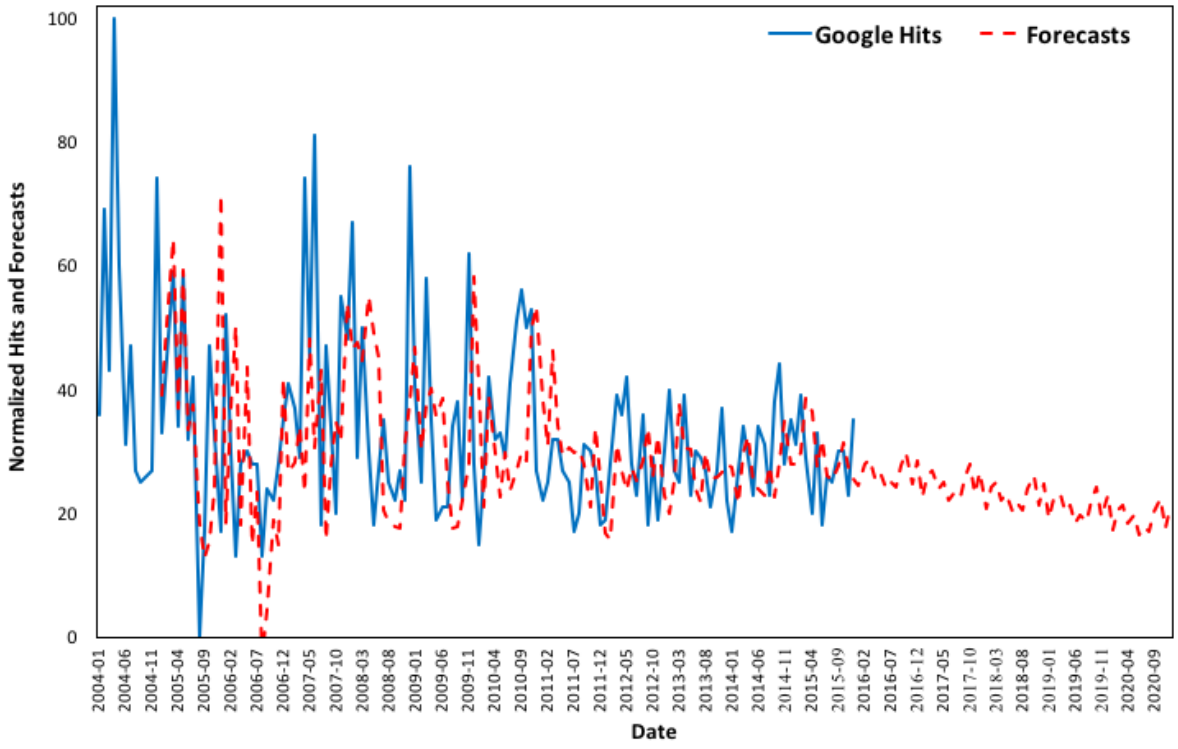


Figure B8. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Delaware.

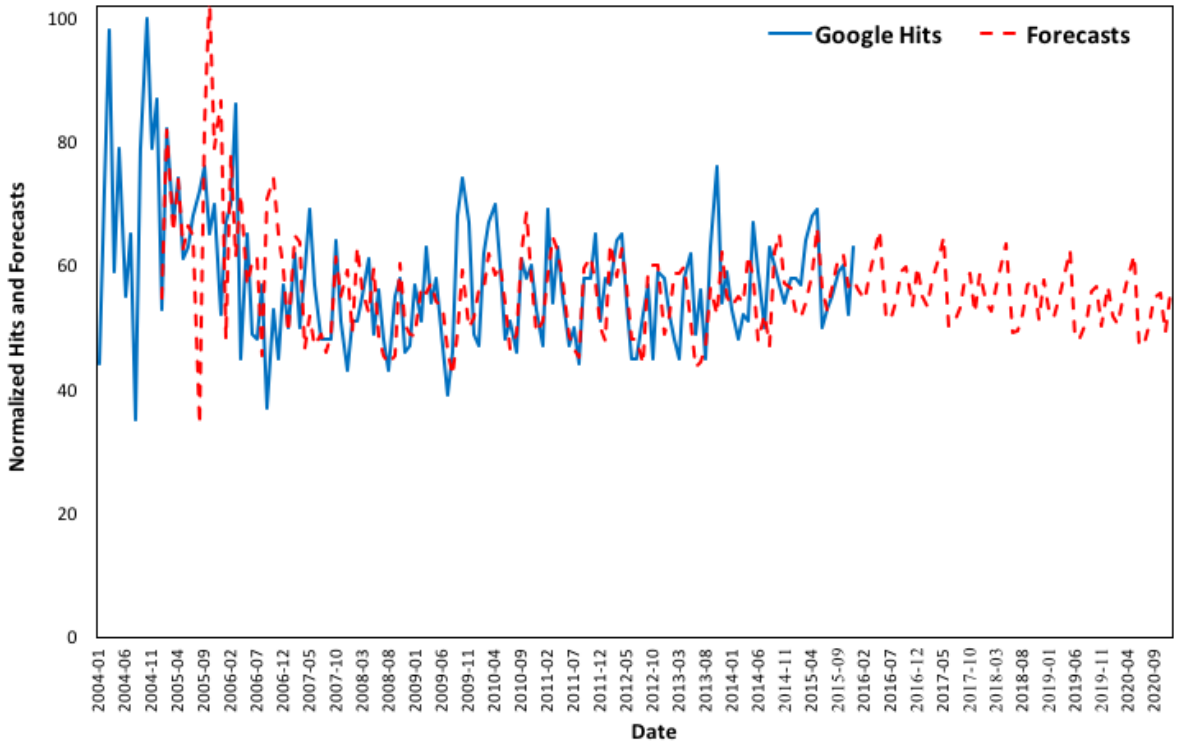


Figure B9. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in DC.

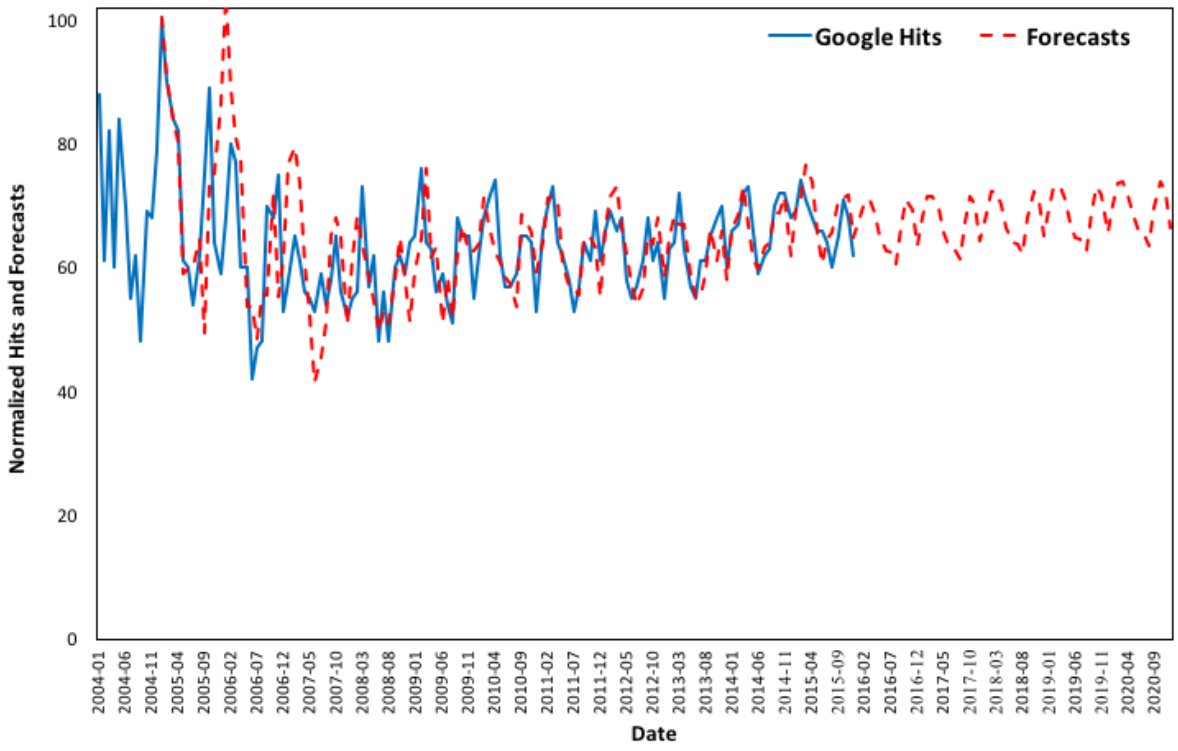


Figure B10. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Florida.

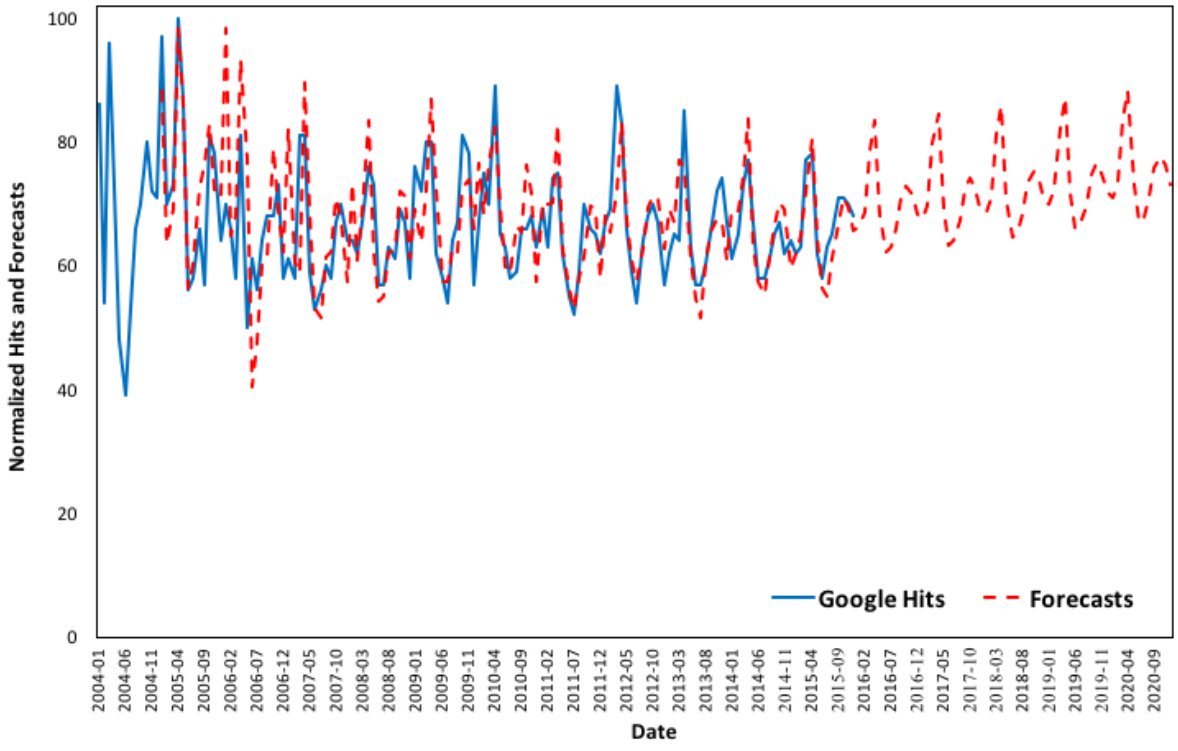


Figure B11. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Georgia.

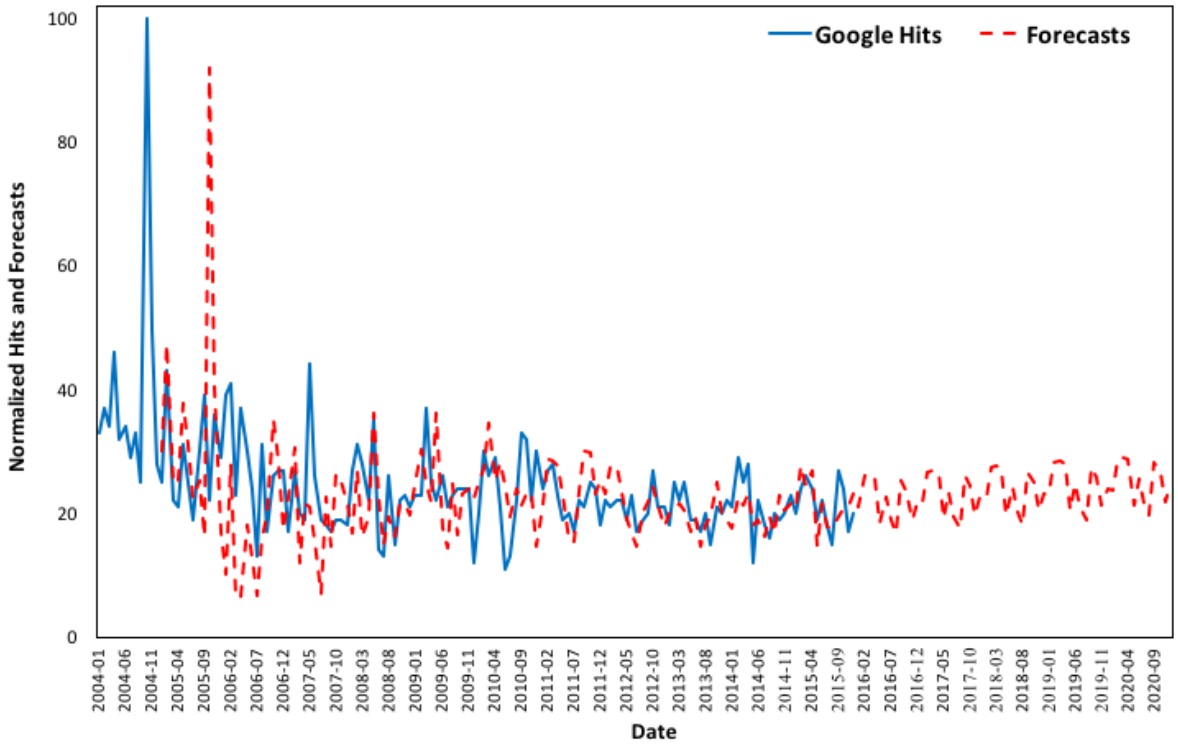


Figure B12. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Hawaii.

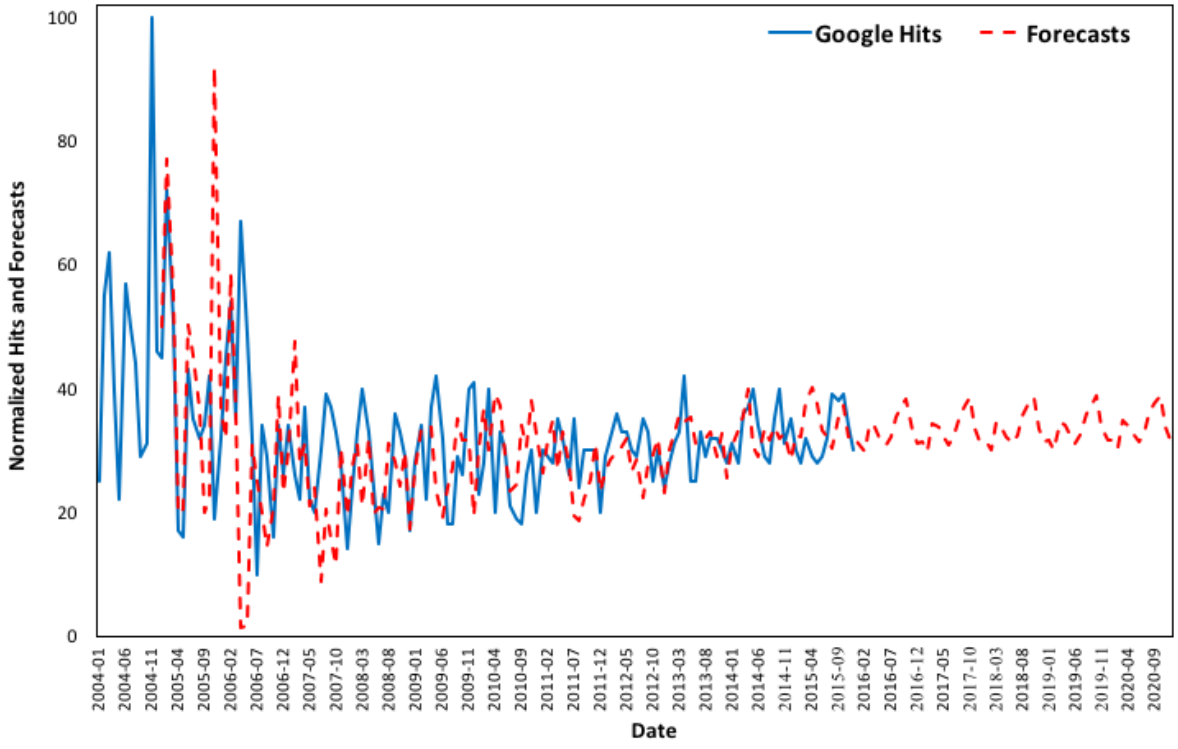


Figure B13. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Idaho.

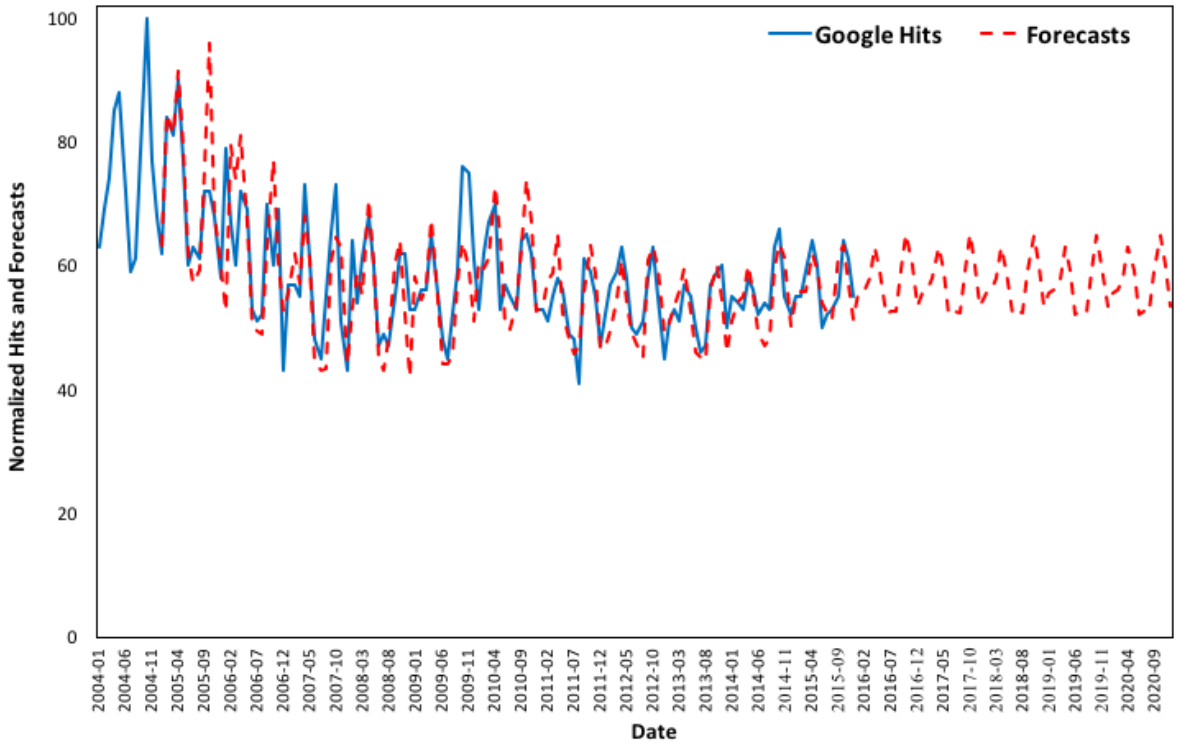


Figure B14. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Illinois.

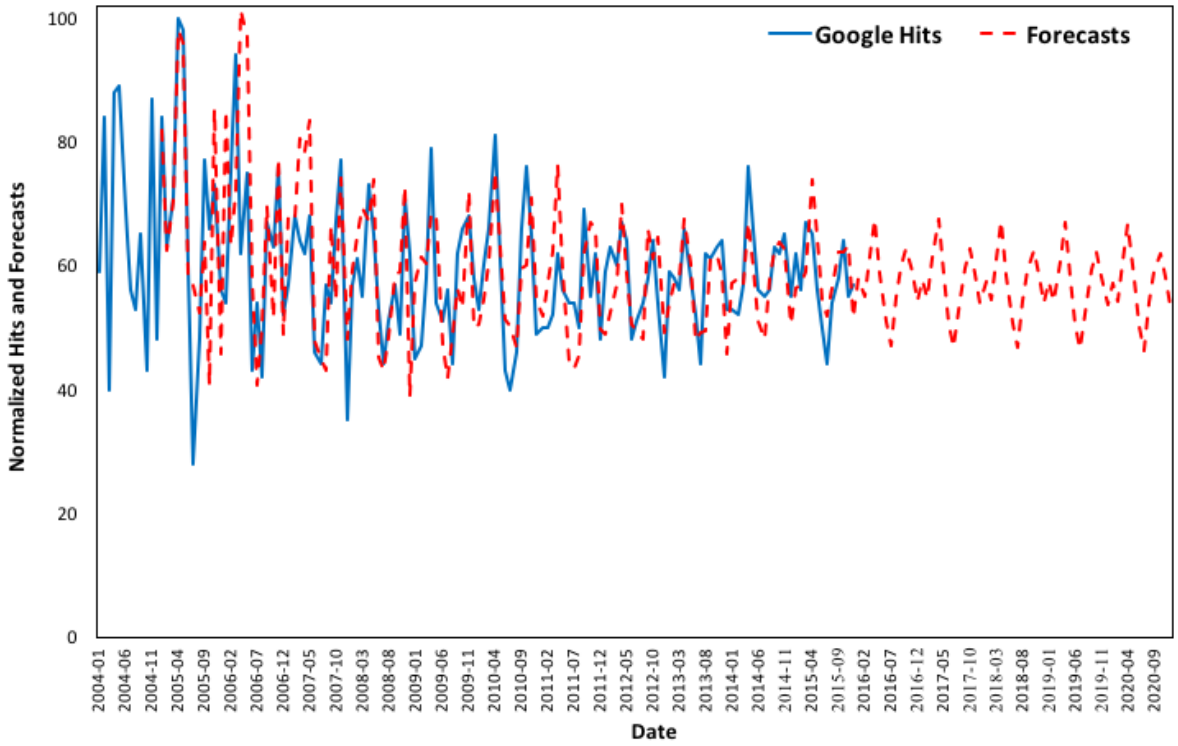


Figure B15. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Indiana.

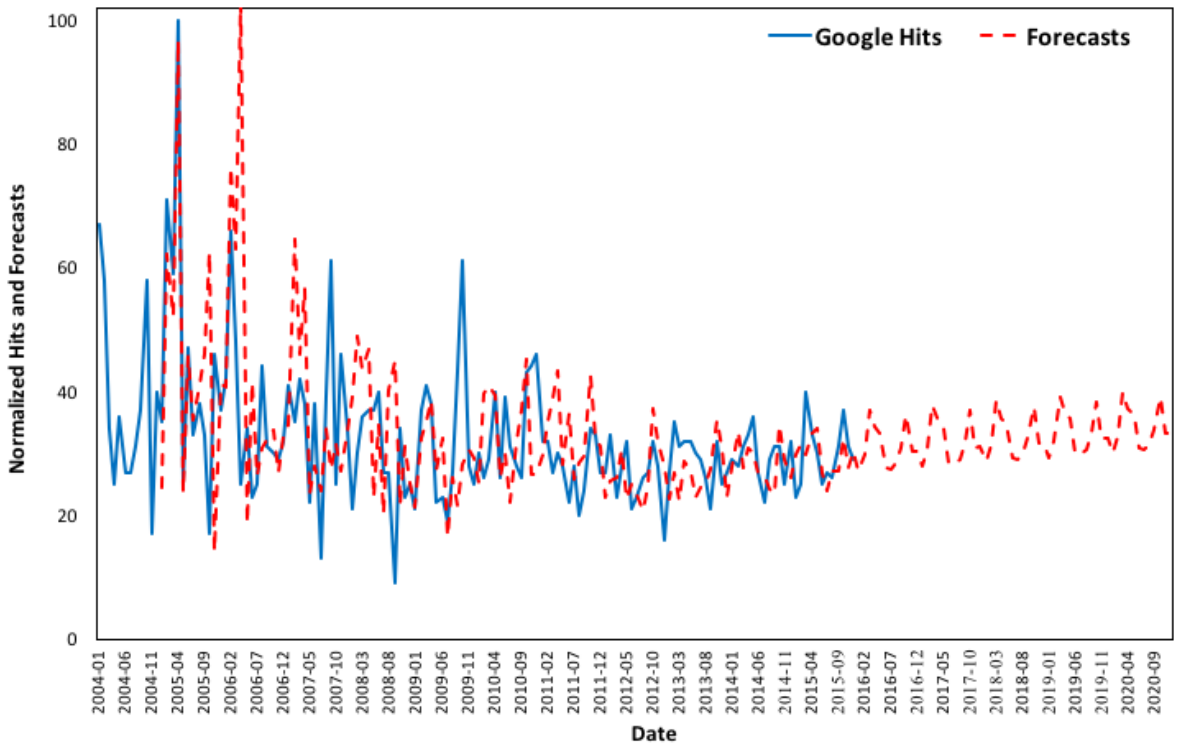


Figure B16. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Iowa.

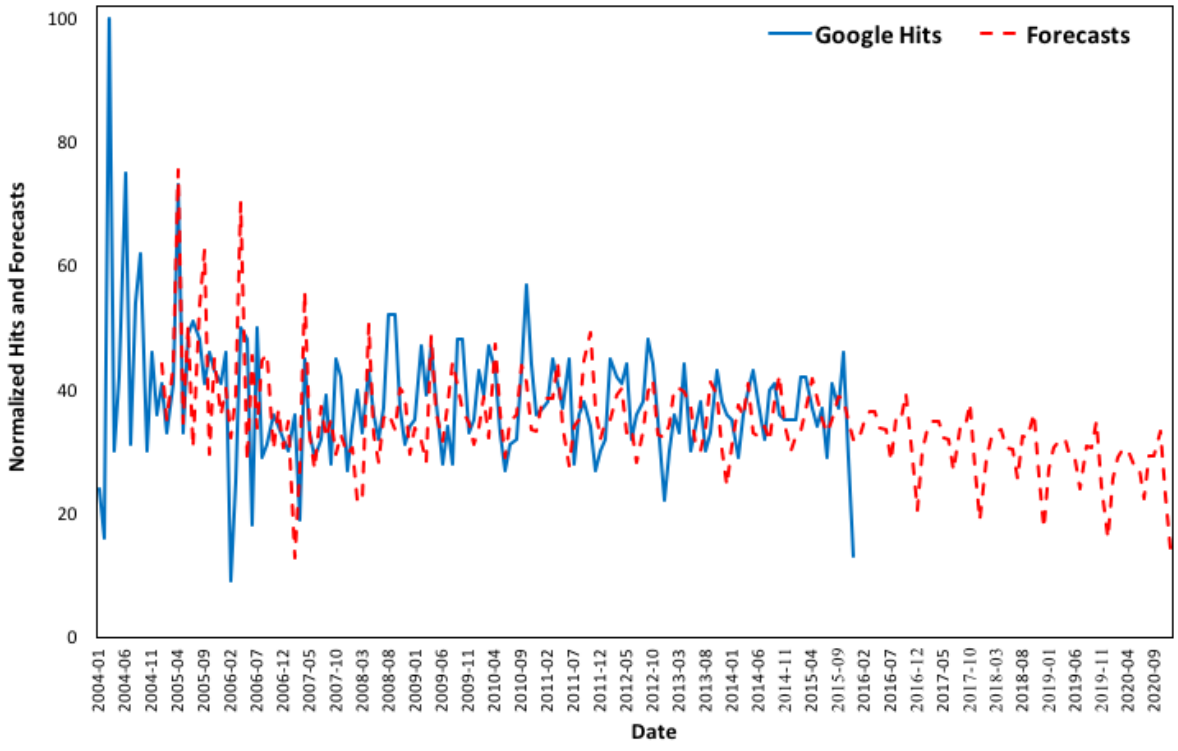


Figure B17. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Kansas.

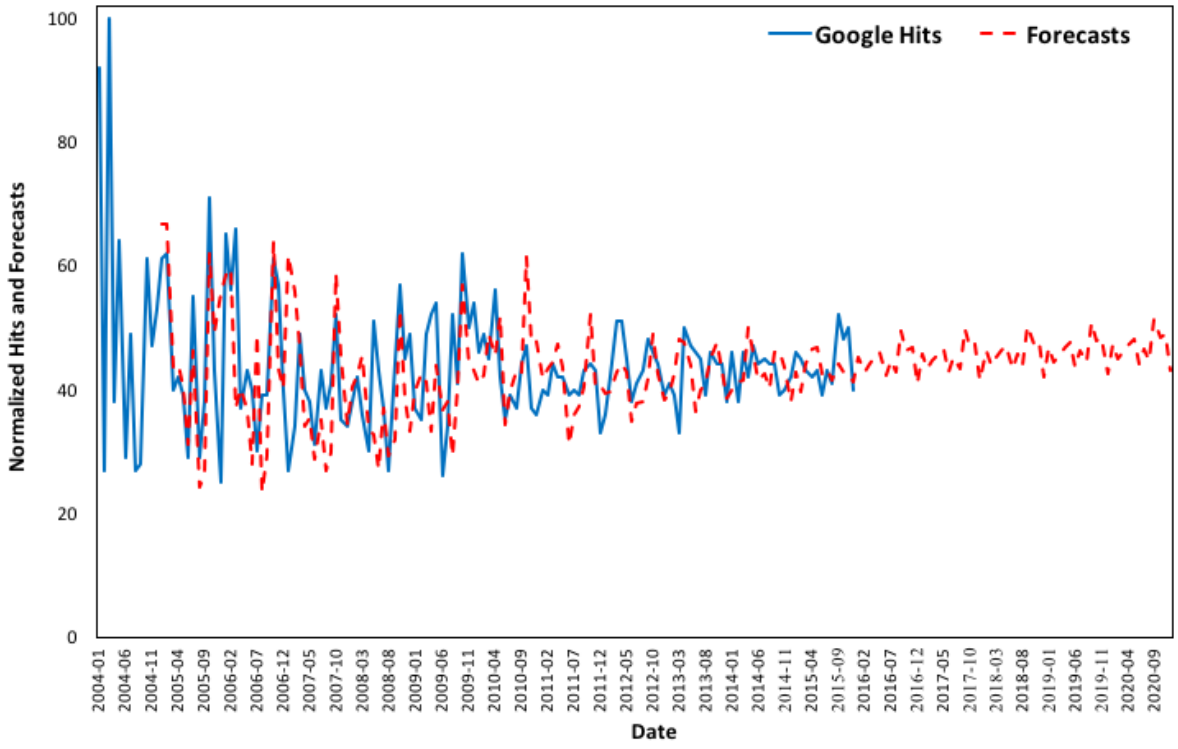


Figure B18. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Kentucky.

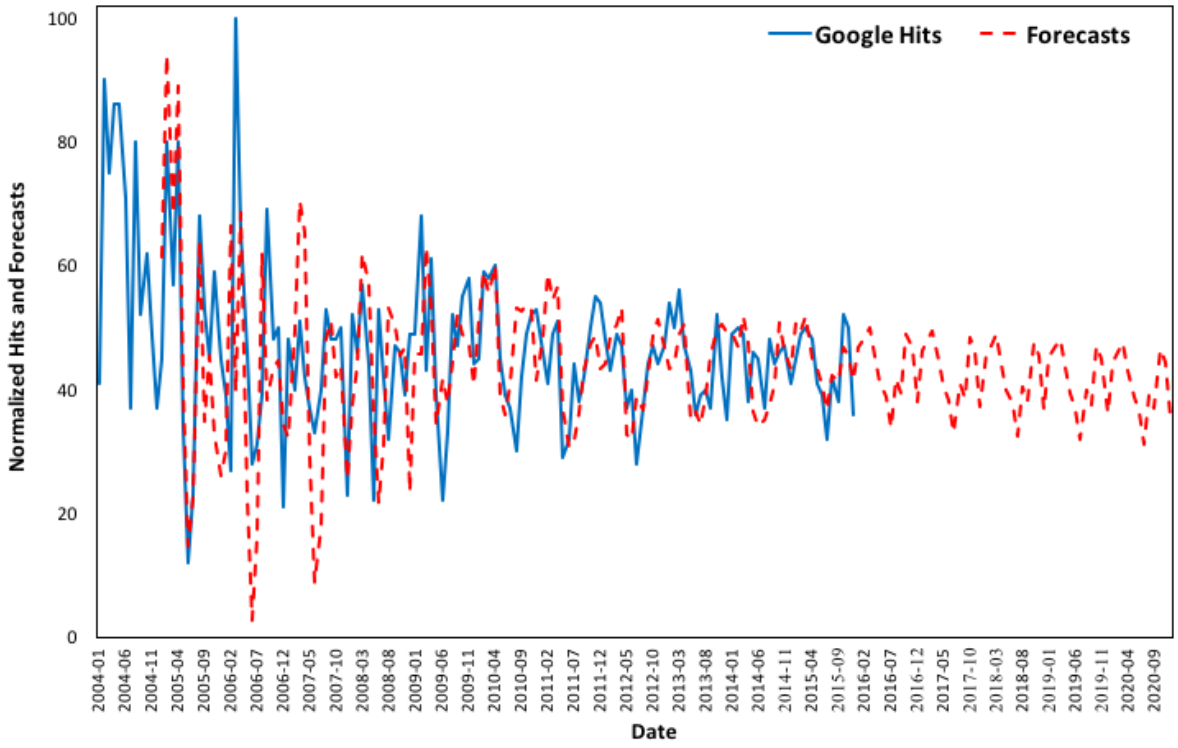


Figure B19. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Louisiana.

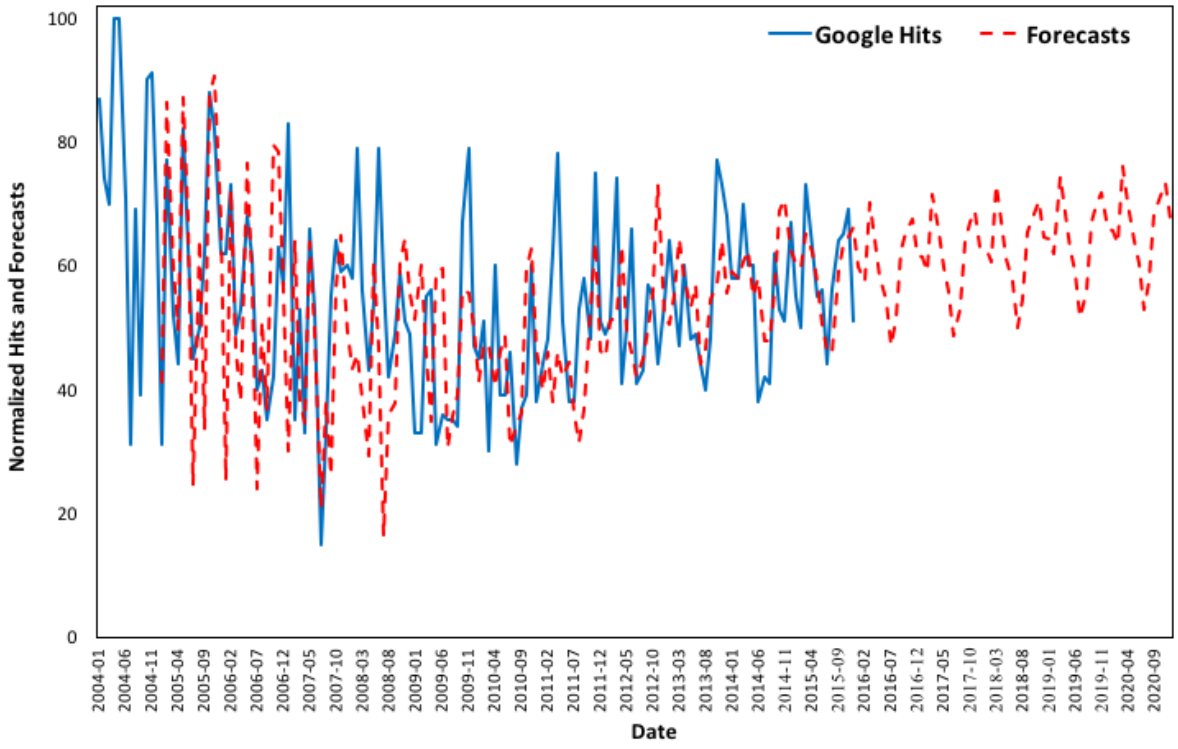


Figure B20. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Maine.

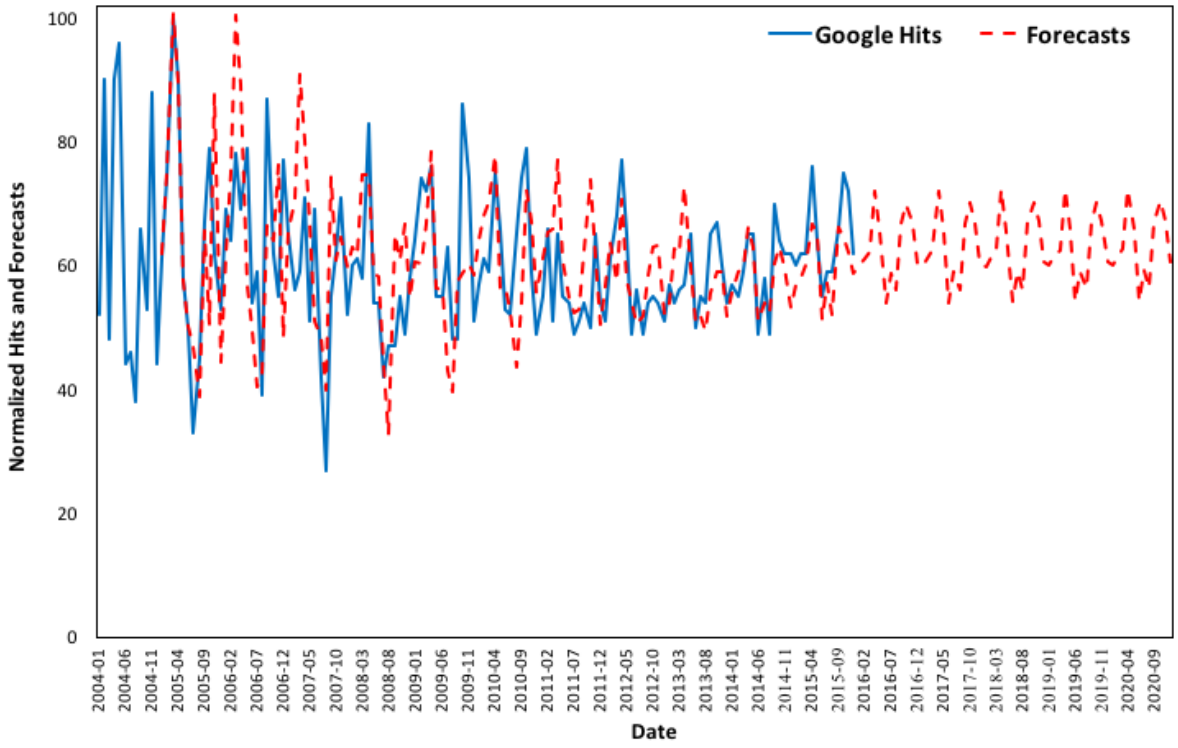


Figure B21. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Maryland.

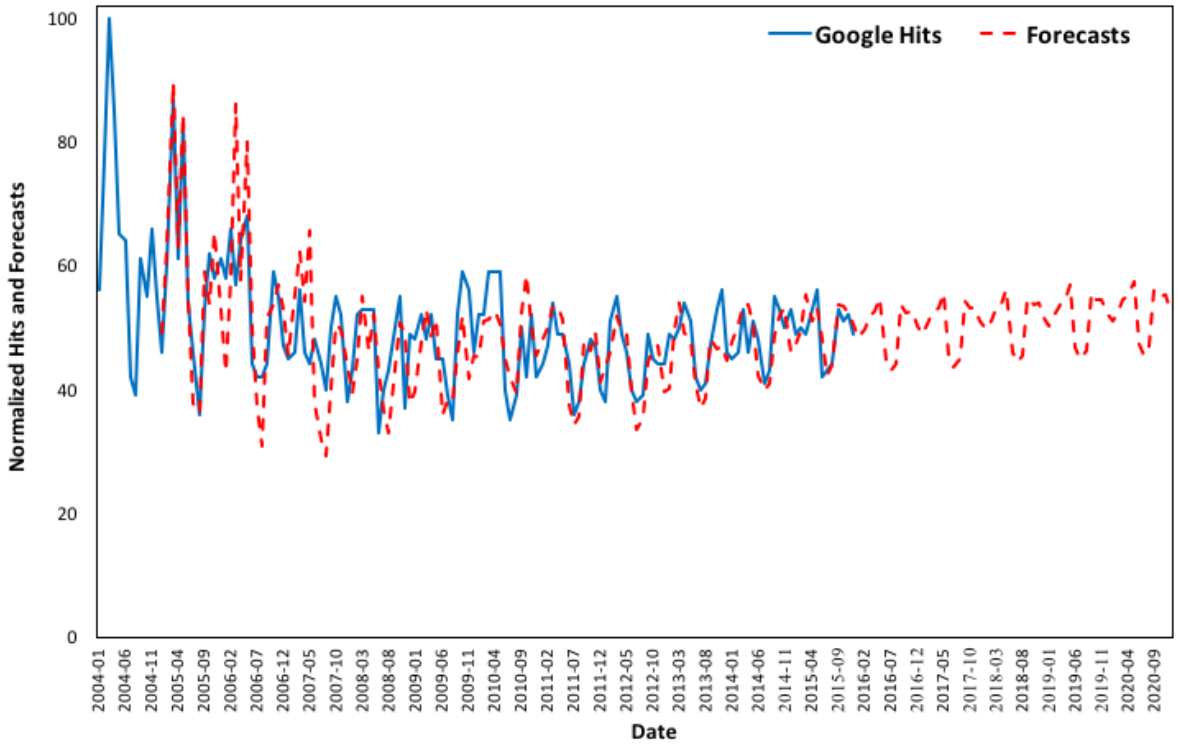


Figure B22. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Massachusetts.

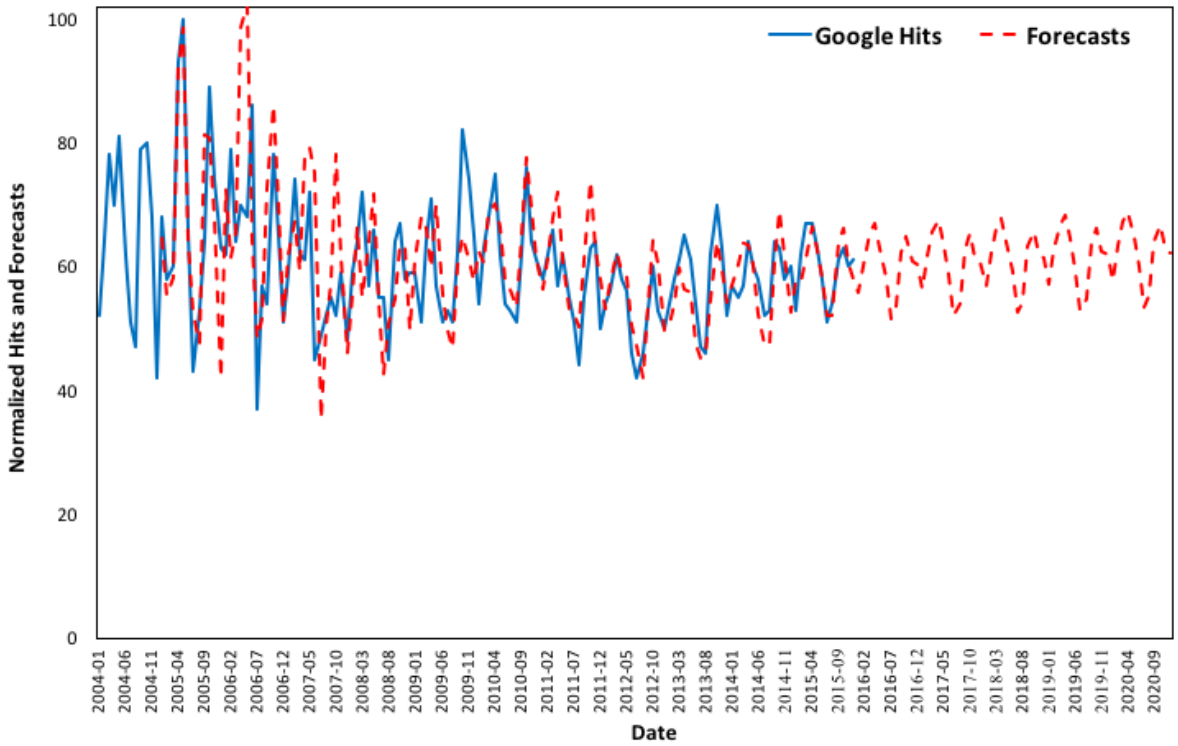


Figure B23. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Michigan.

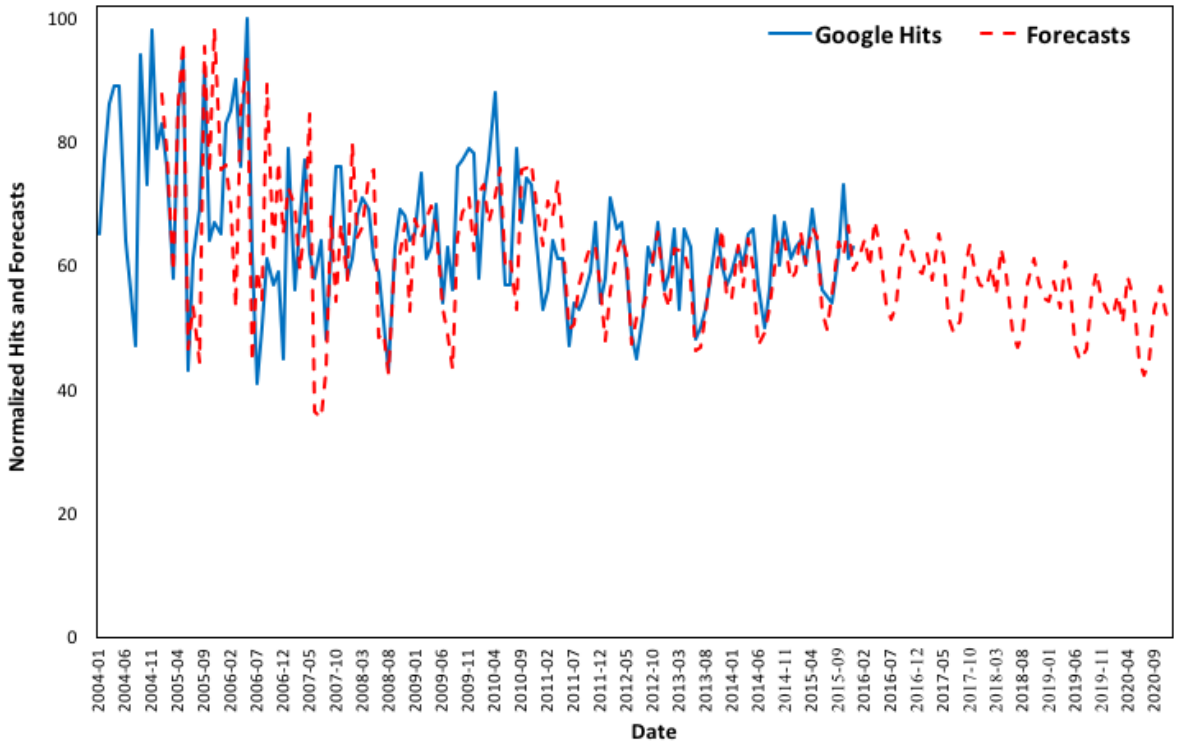


Figure B24. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Minnesota.

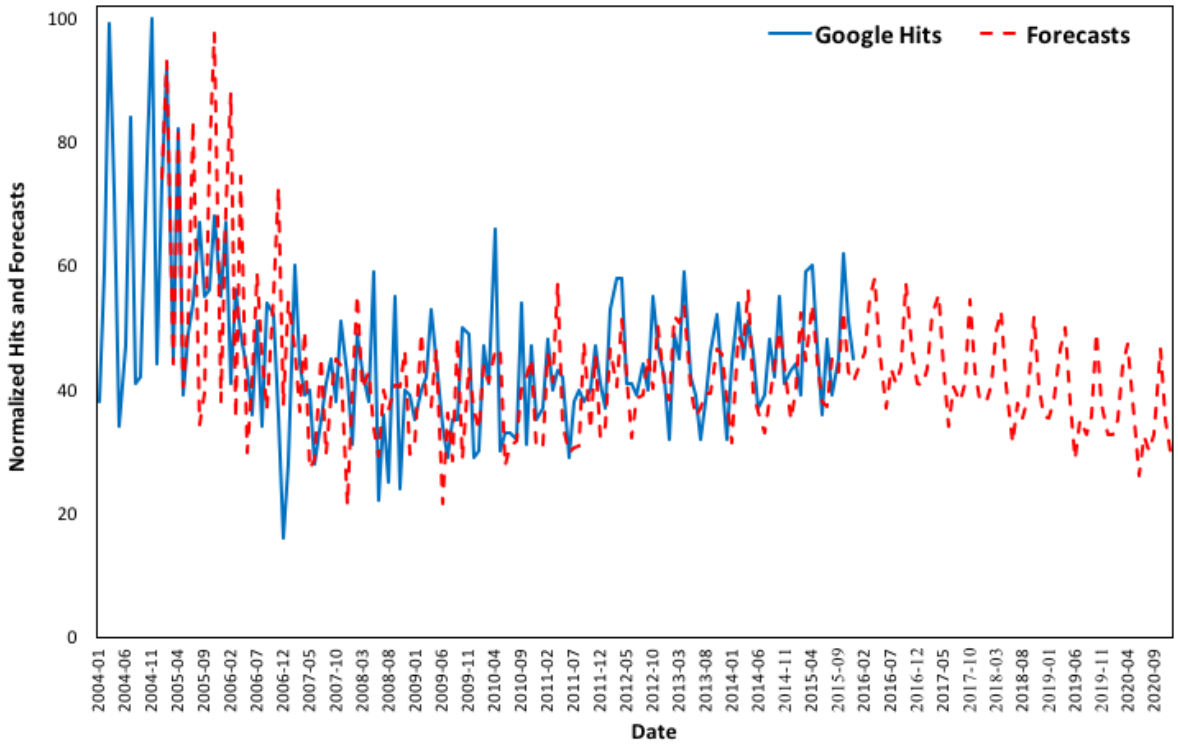


Figure B25. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Mississippi.

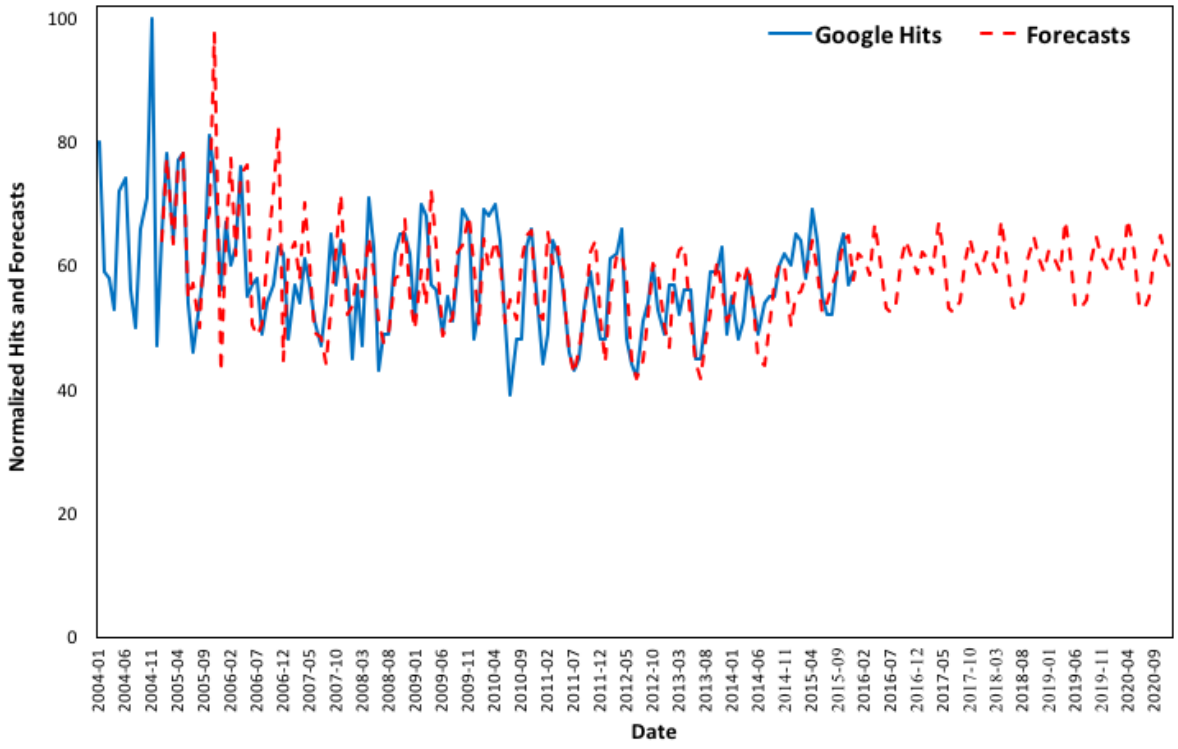


Figure B26. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Missouri.

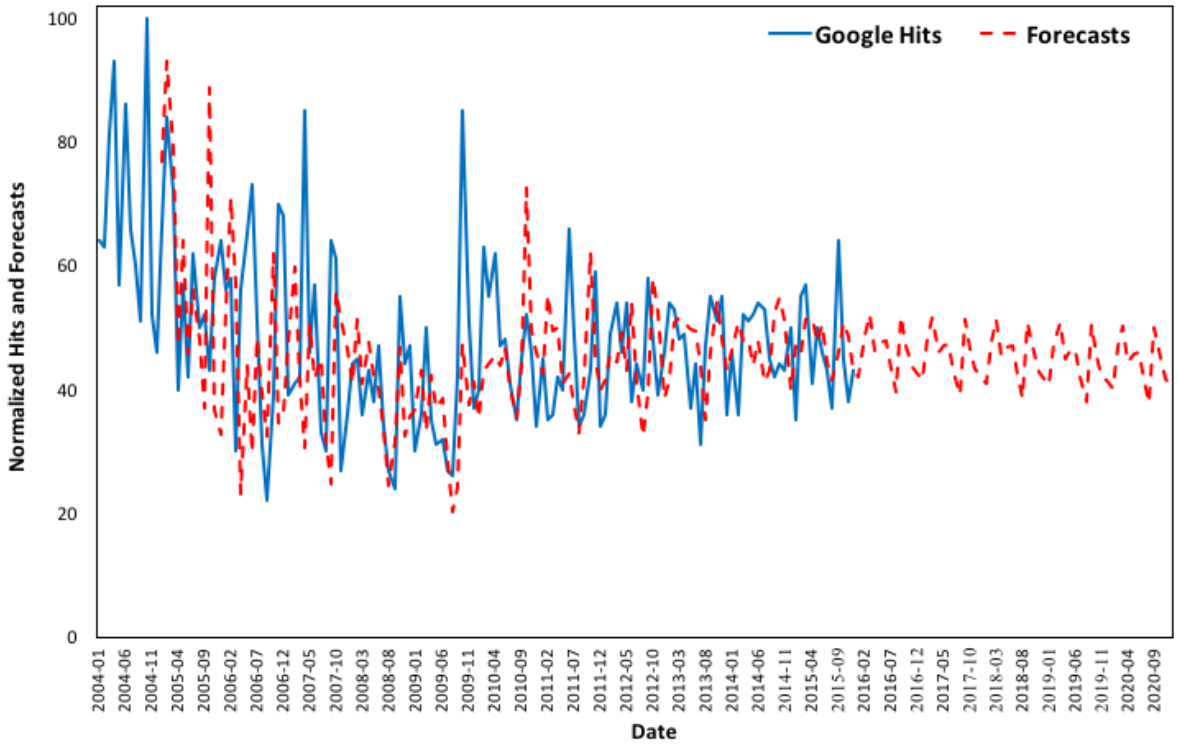


Figure B27. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Montana.

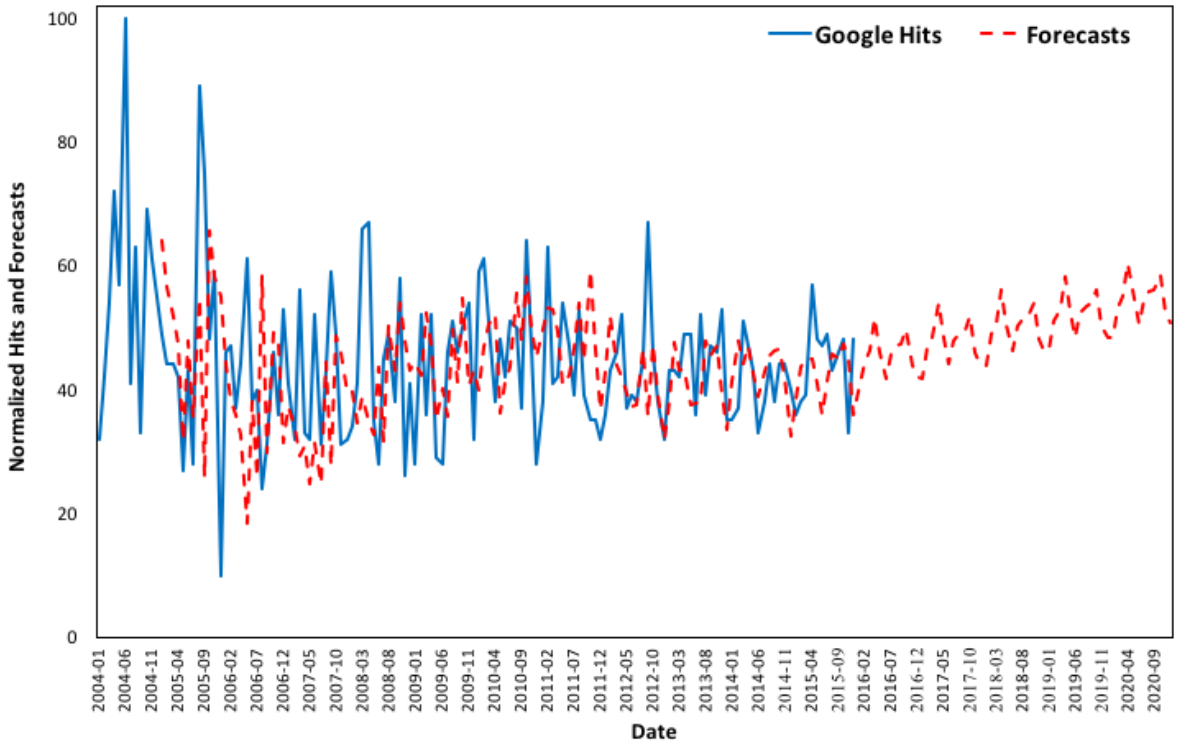


Figure B28. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Nebraska.

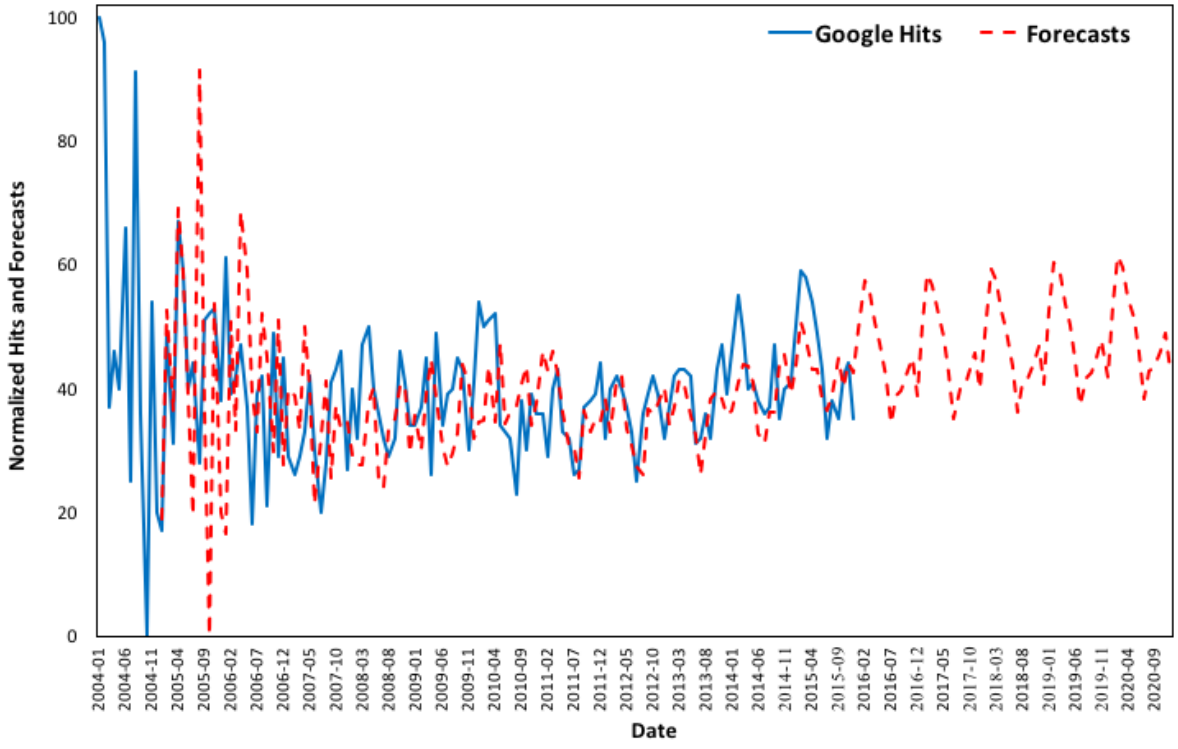


Figure B29. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Nevada.

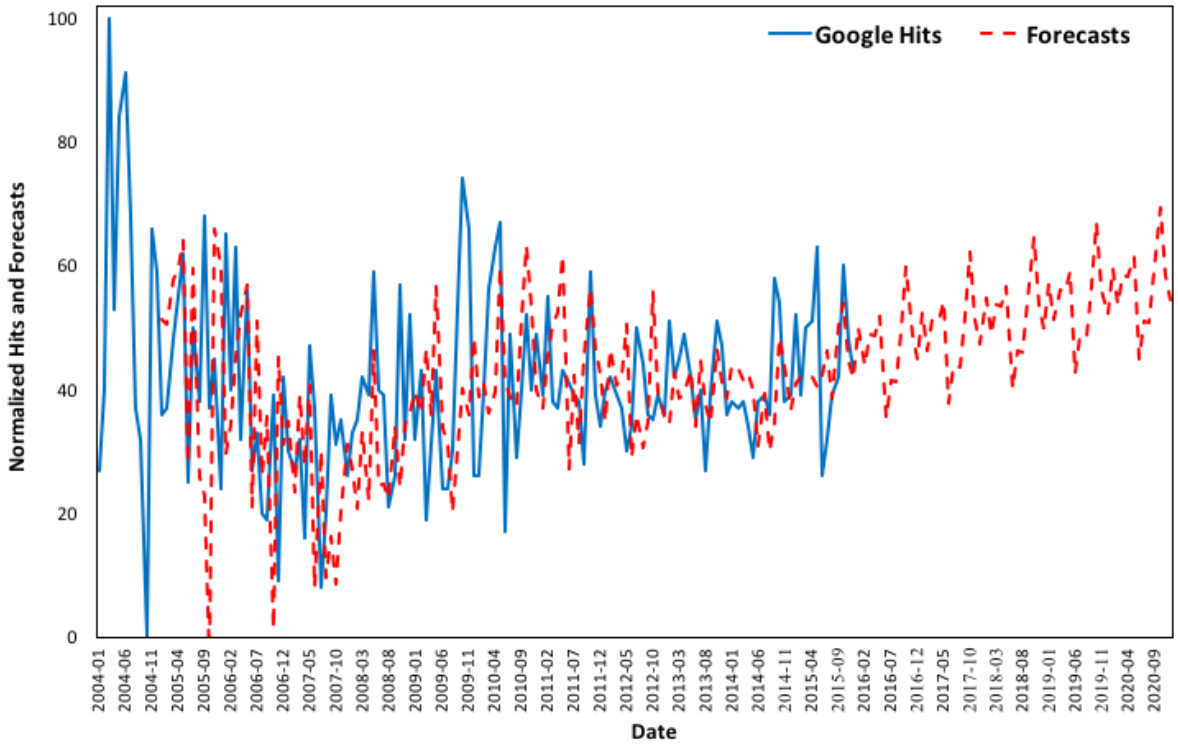


Figure B30. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in New Hampshire.

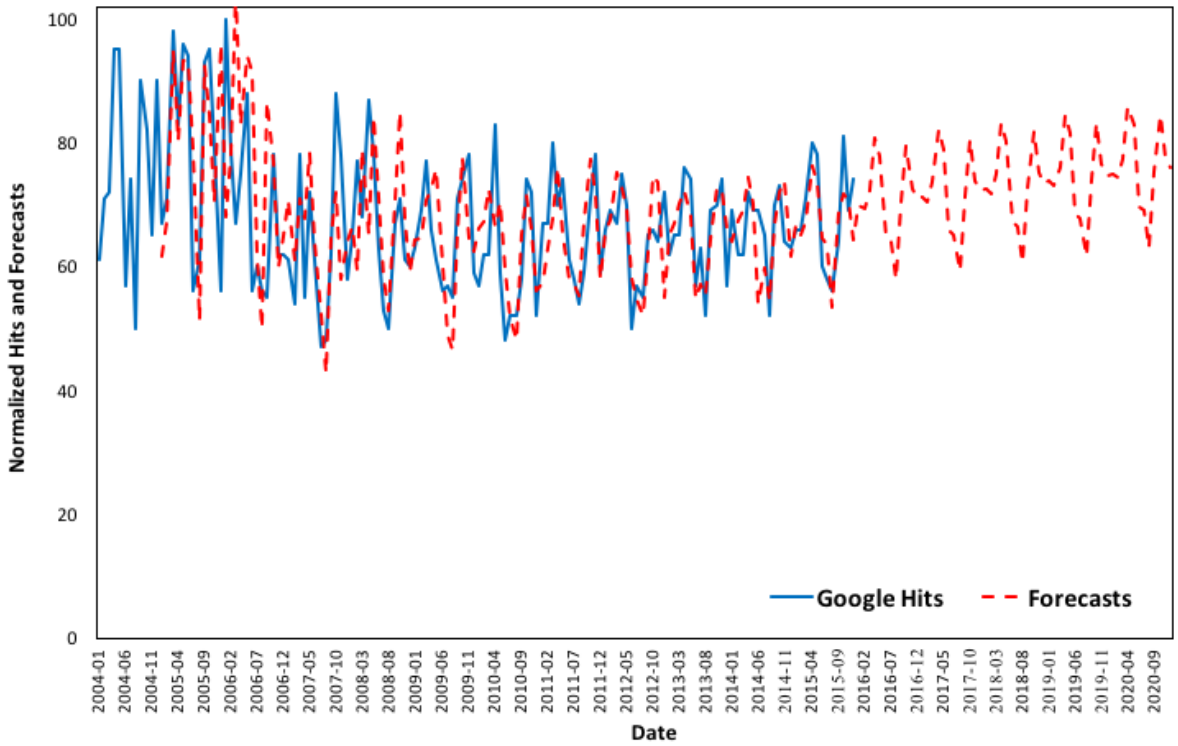


Figure B31. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in New Jersey.

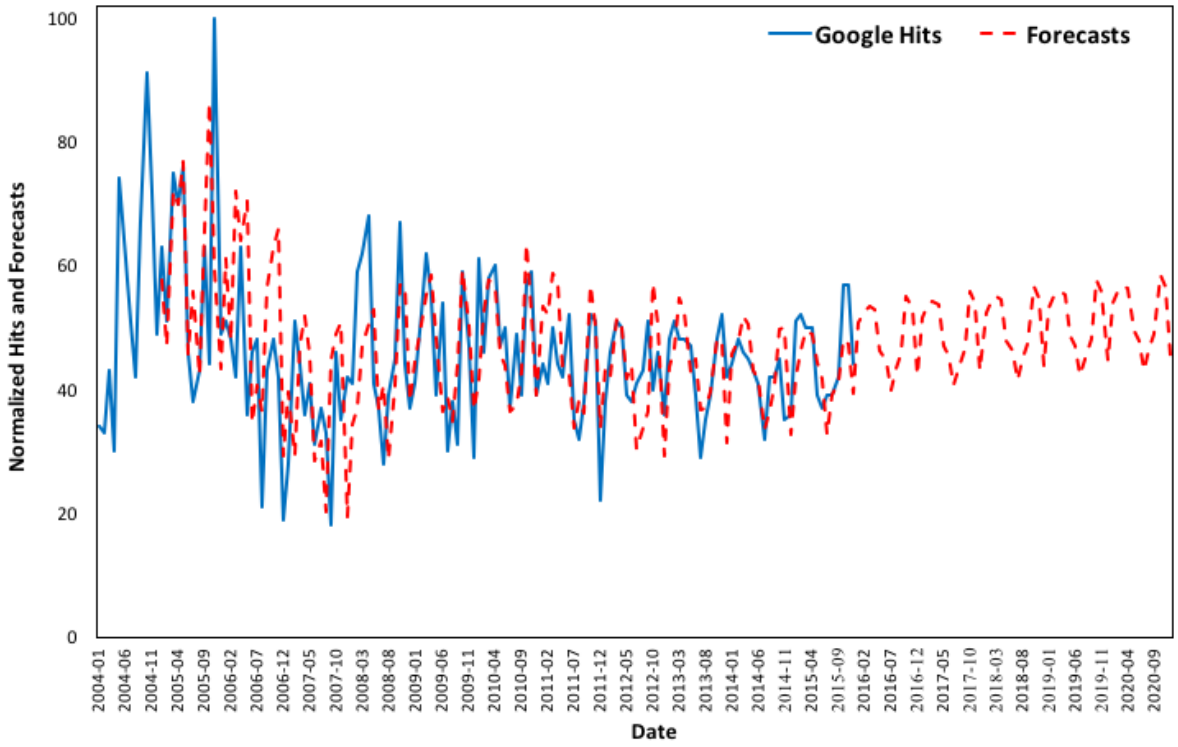


Figure B32. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in New Mexico.

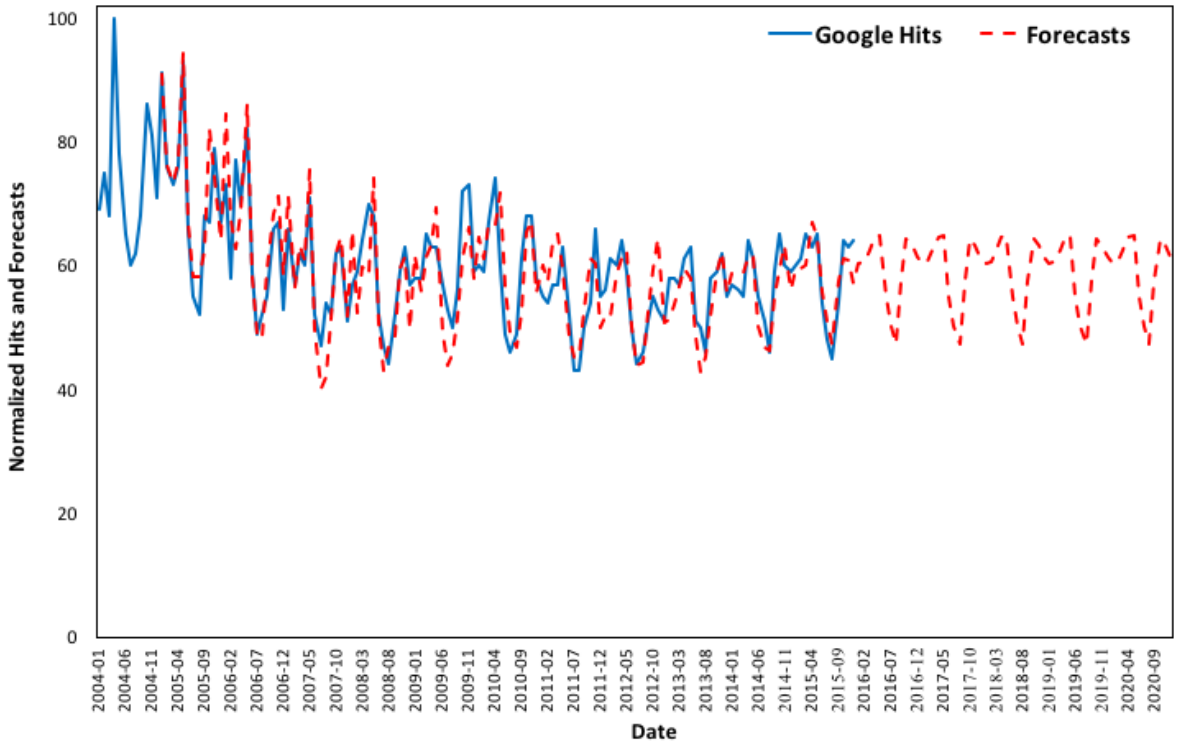


Figure B33. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in New York.

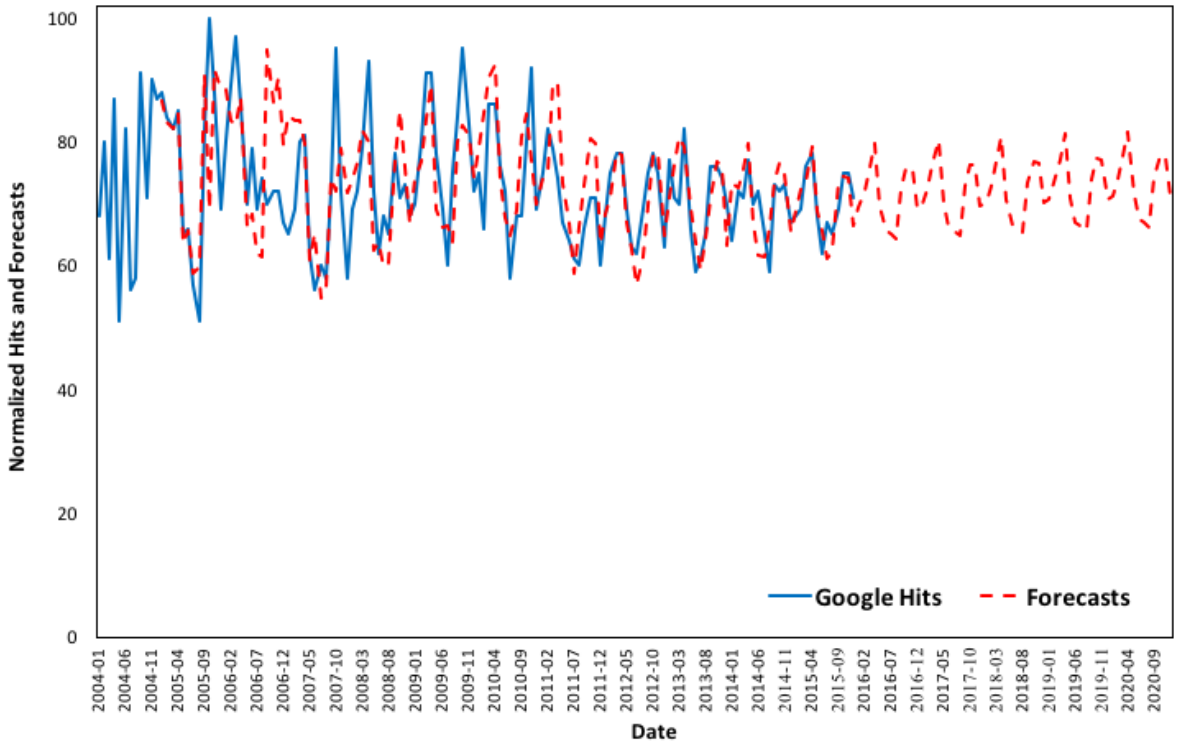


Figure B34. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in North Carolina.

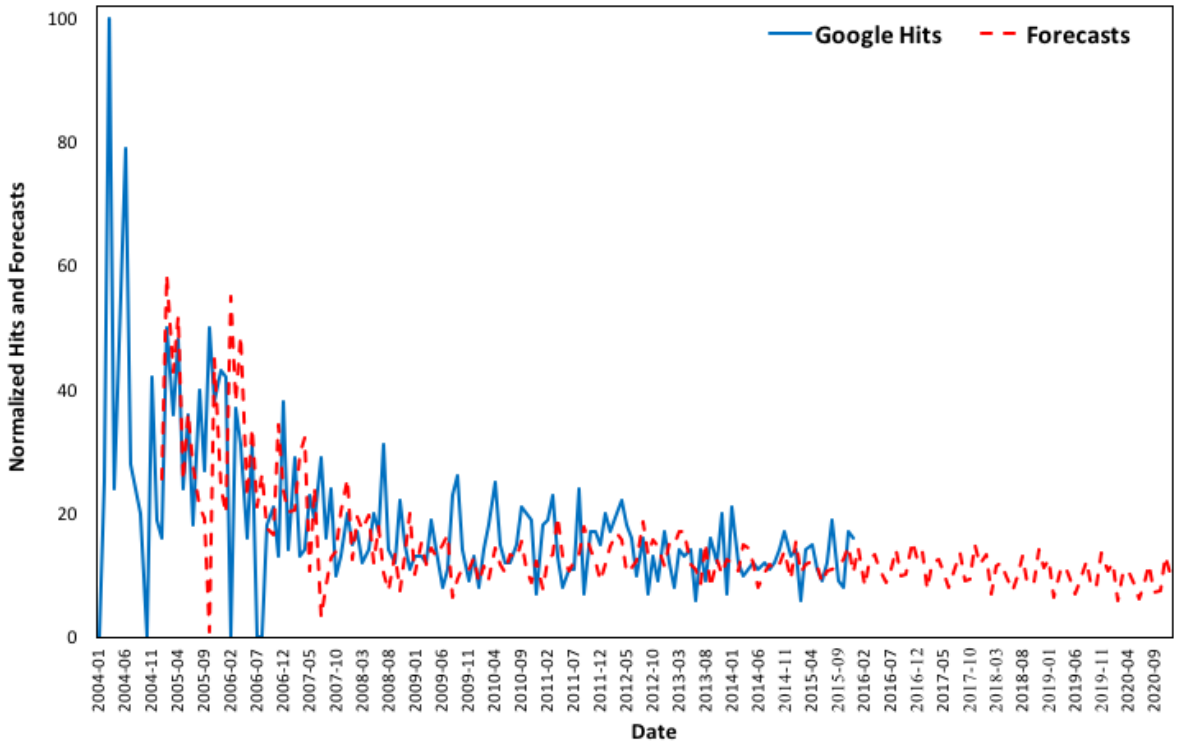


Figure B35. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in North Dakota.

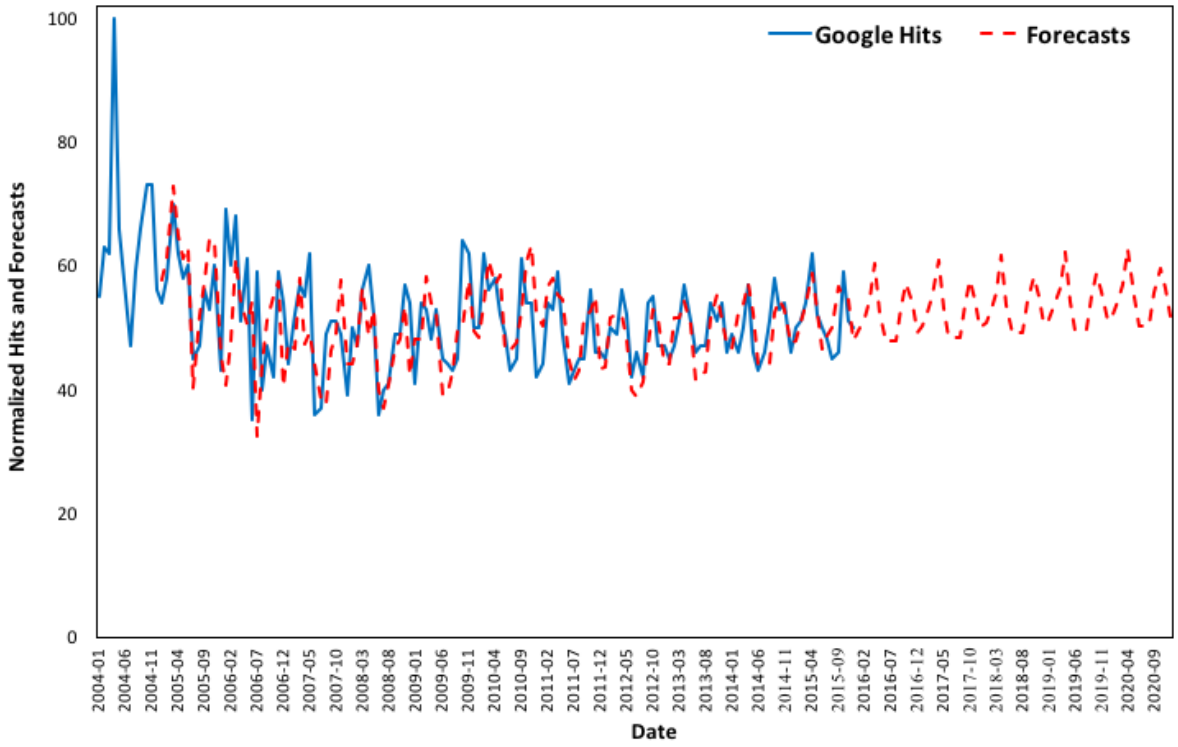


Figure B36. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Ohio.

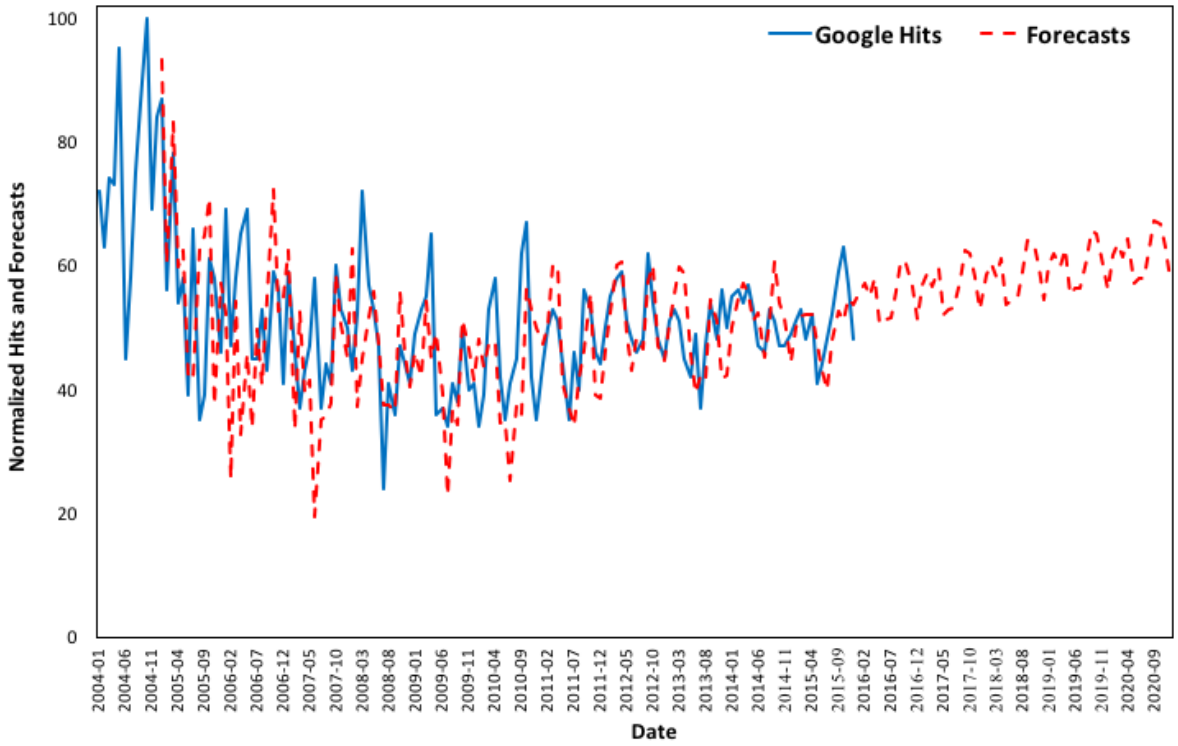


Figure B37. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Oklahoma.

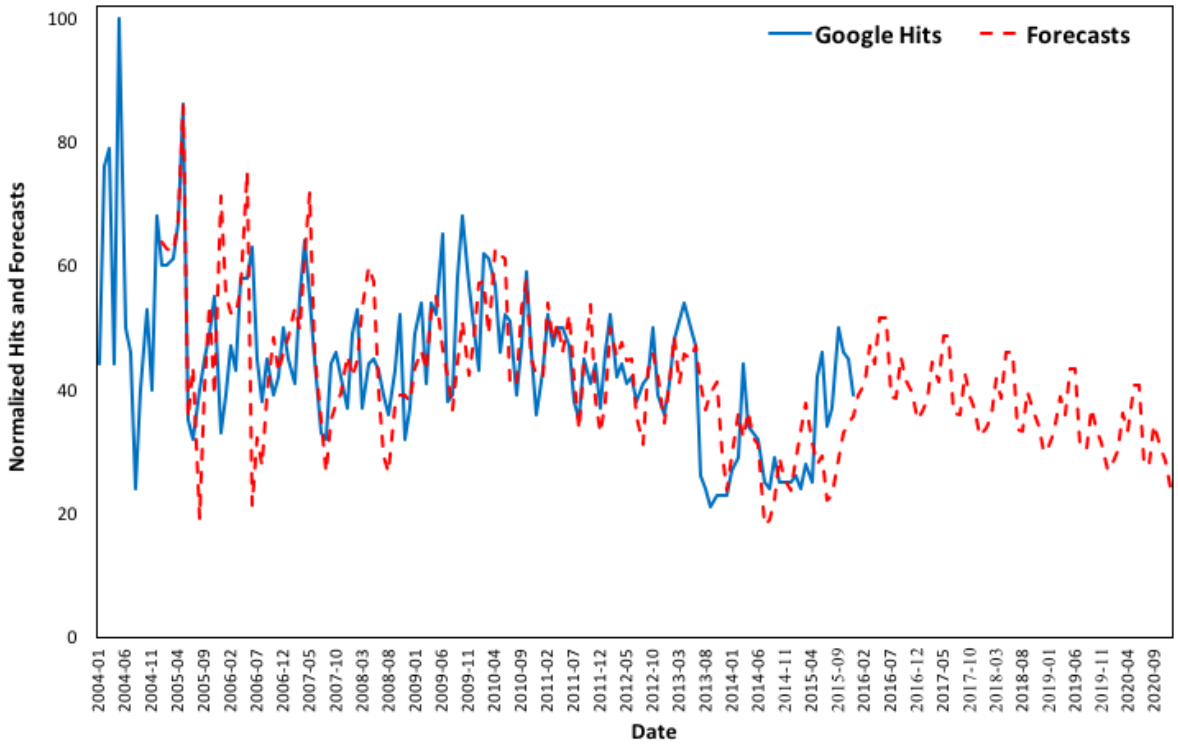


Figure B38. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Oregon.

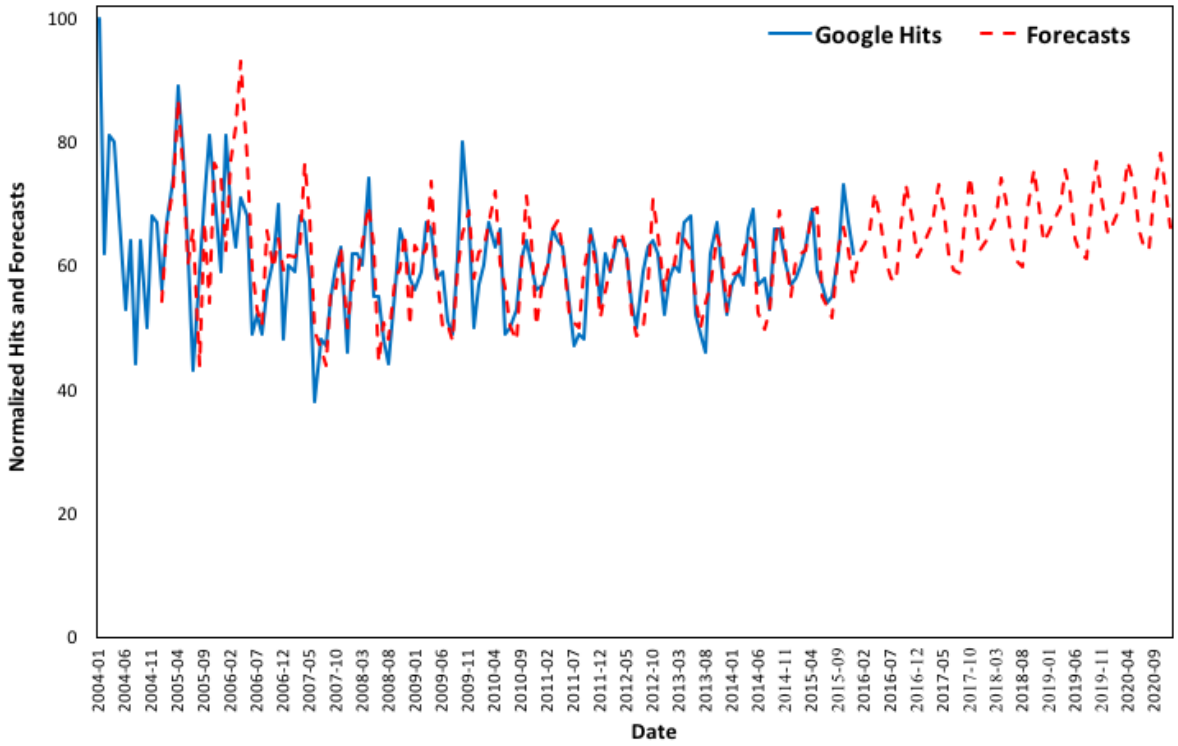


Figure B39. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Pennsylvania.

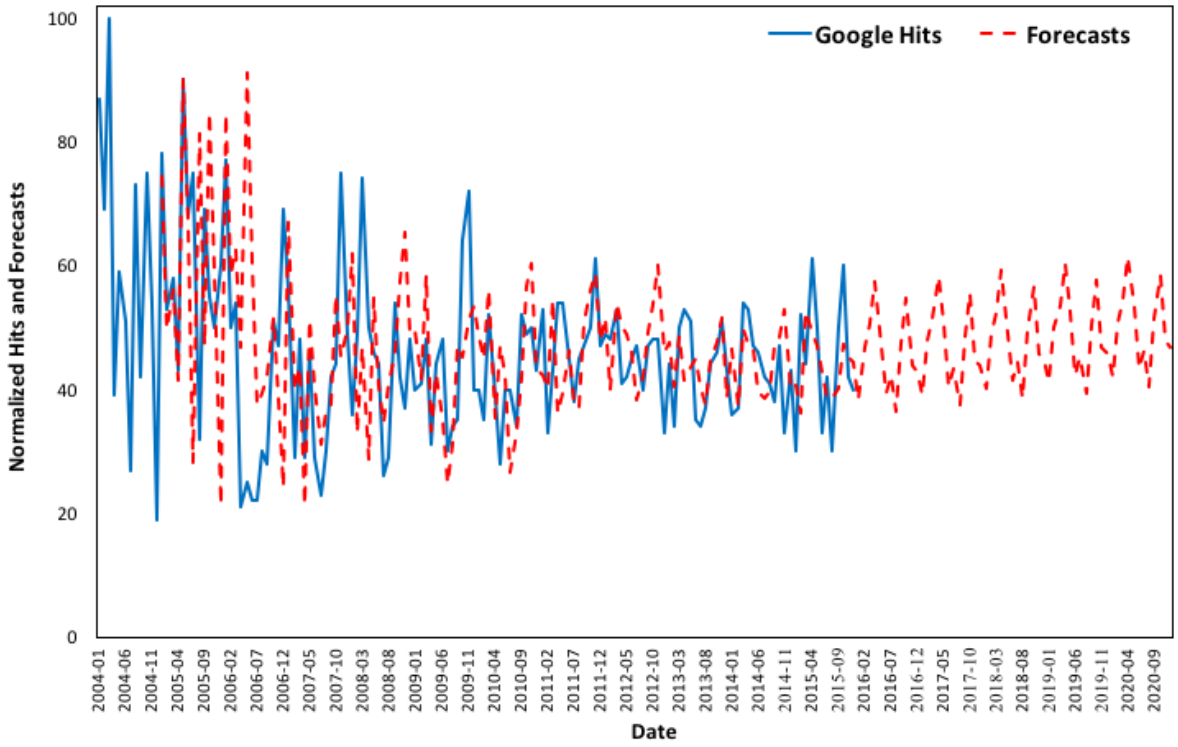


Figure B40. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Rhode Island.

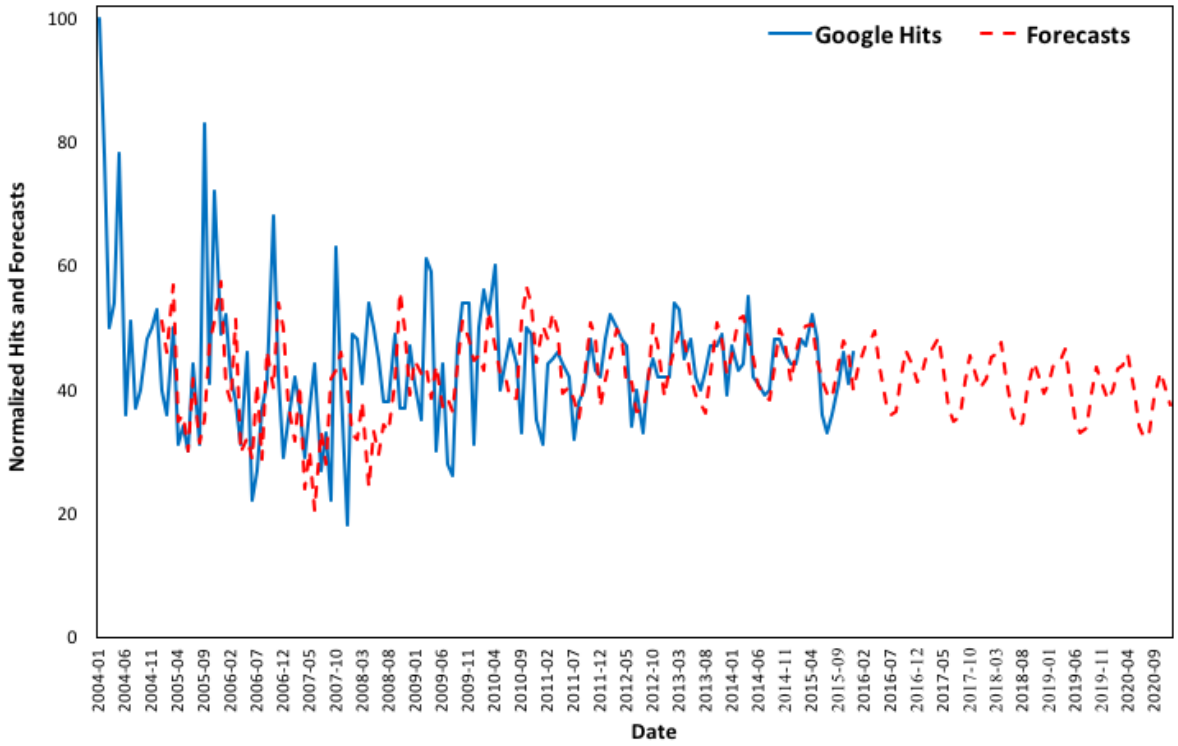


Figure B41. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in South Carolina.

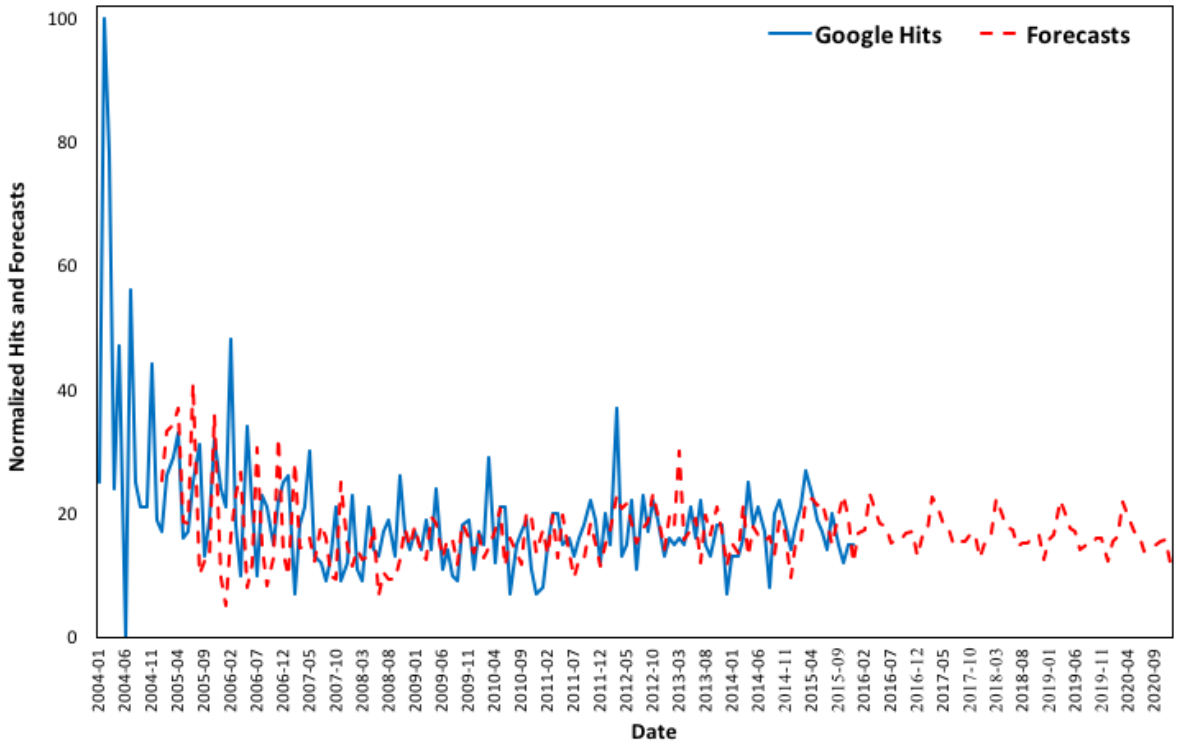


Figure B42. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in South Dakota.

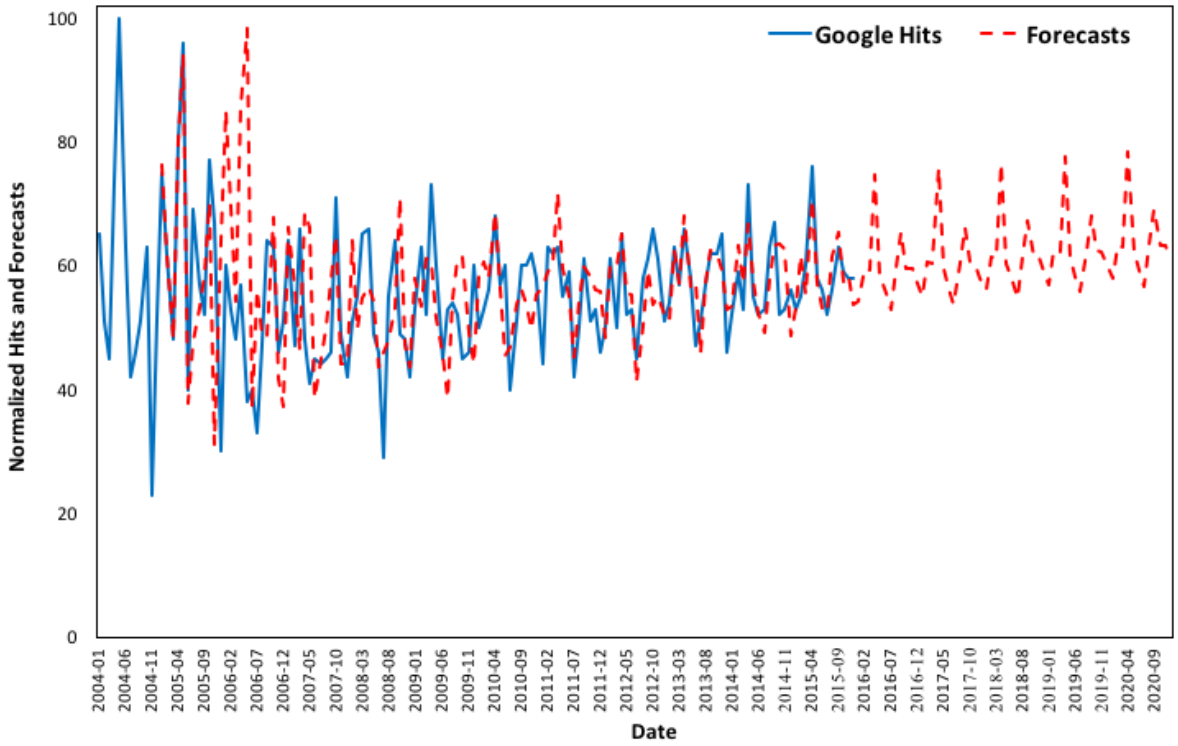


Figure B43. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Tennessee.

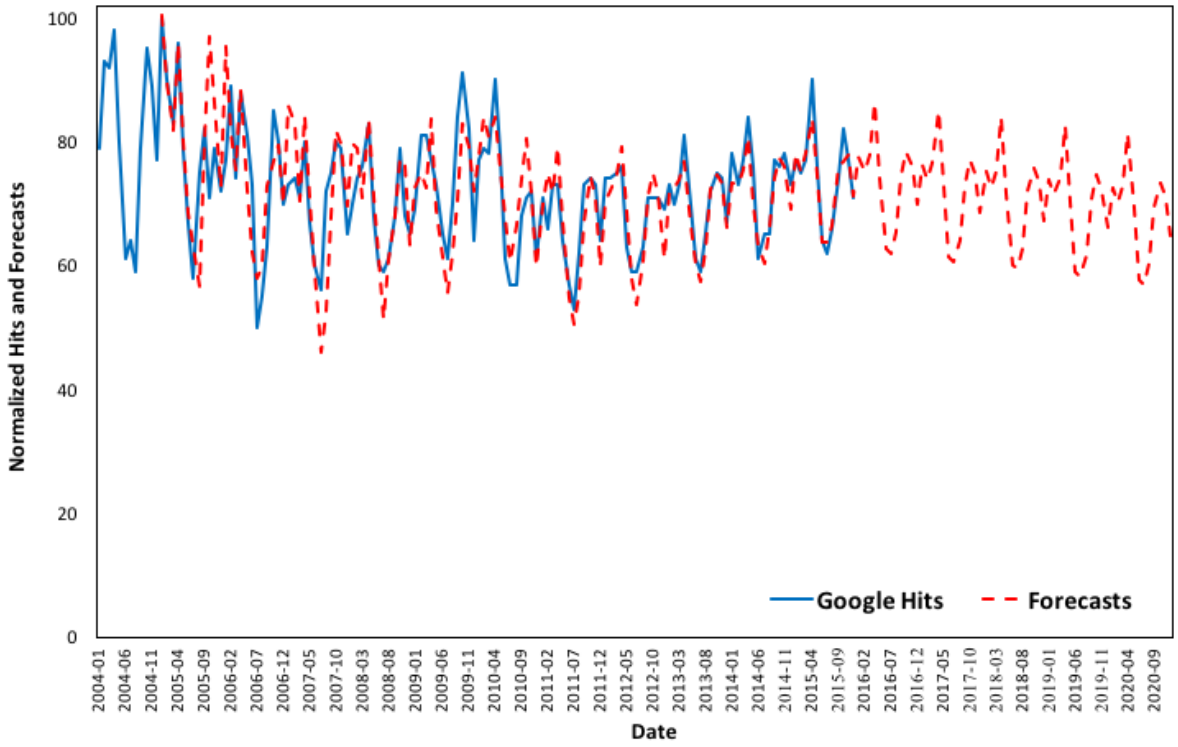


Figure B44. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Texas.

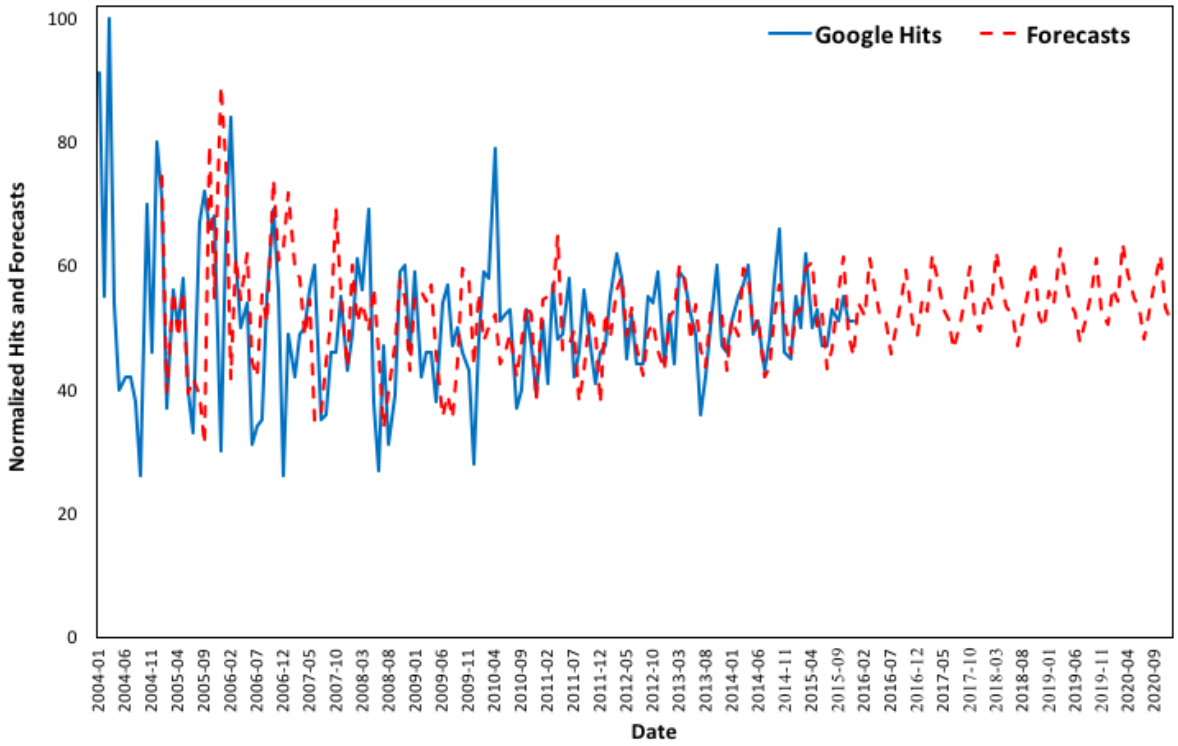


Figure B45. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Utah.

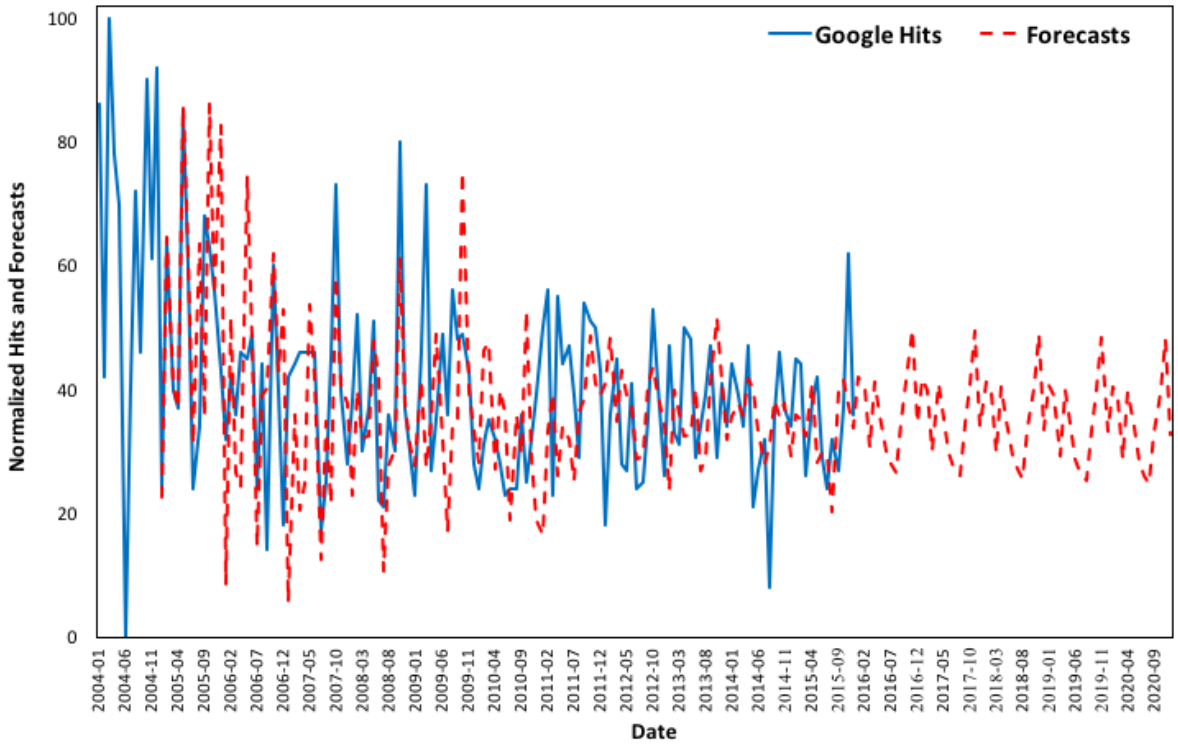


Figure B46. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Vermont.

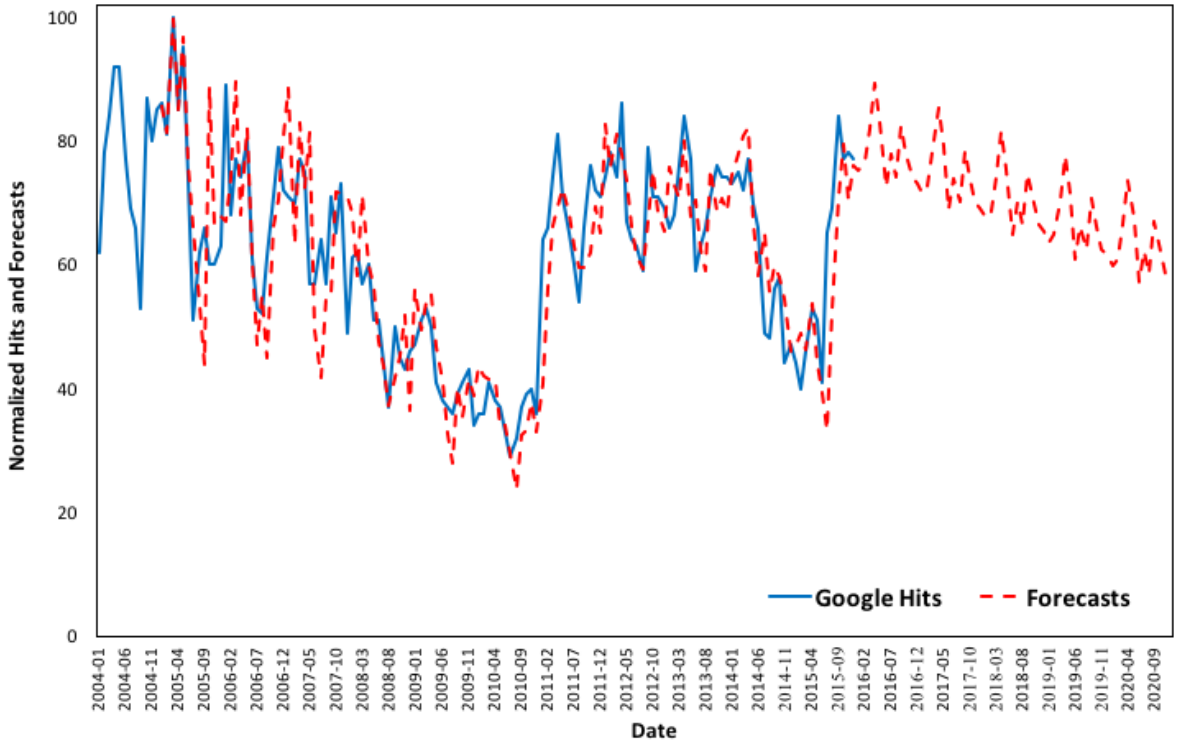


Figure B47. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Virginia.

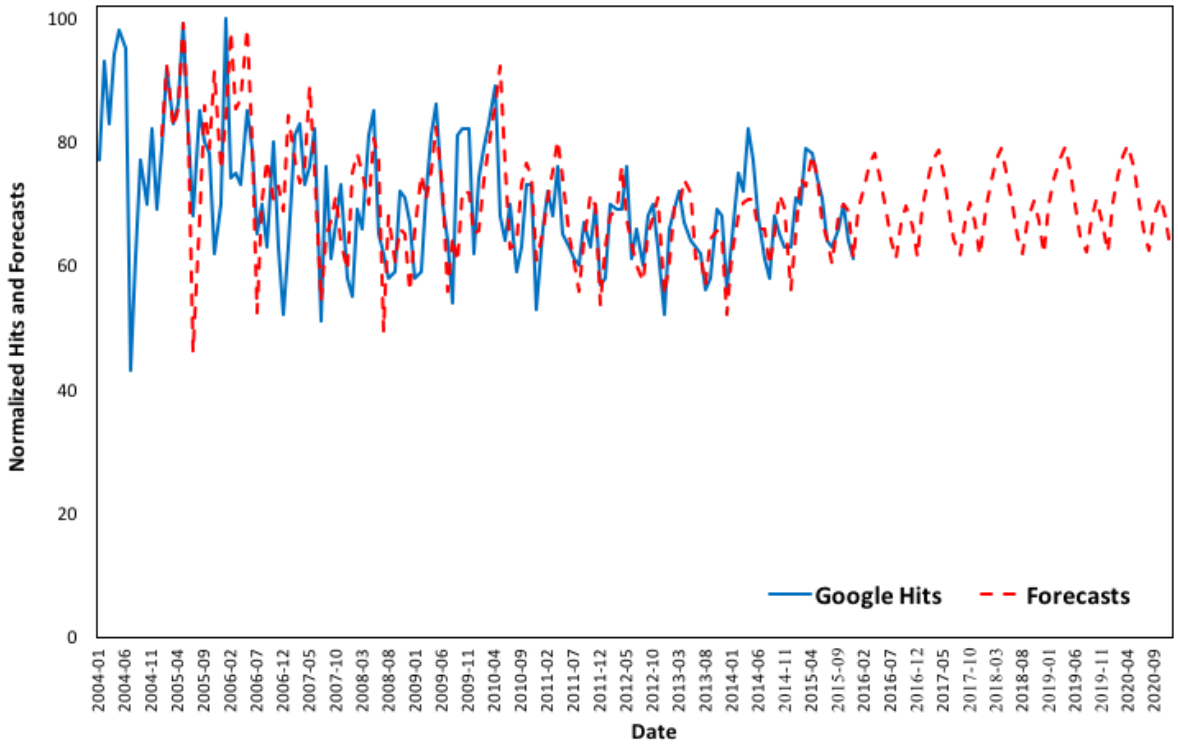


Figure B48. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Washington.

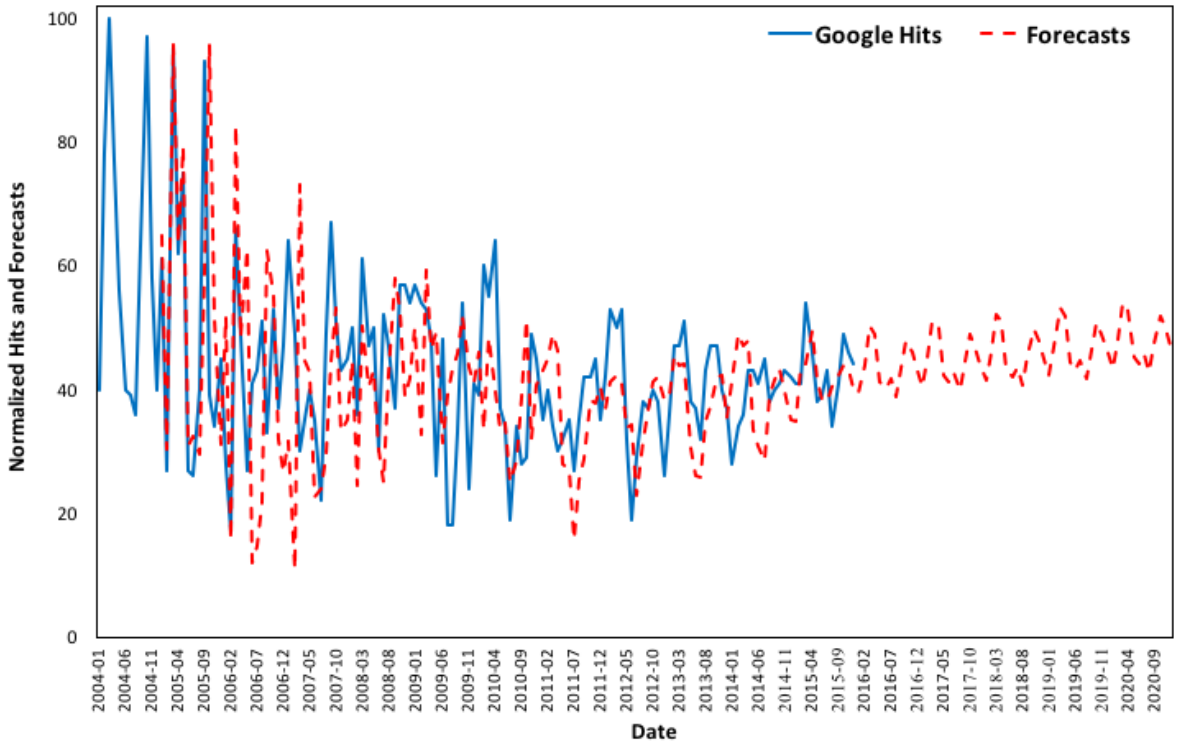


Figure B49. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in West Virginia.

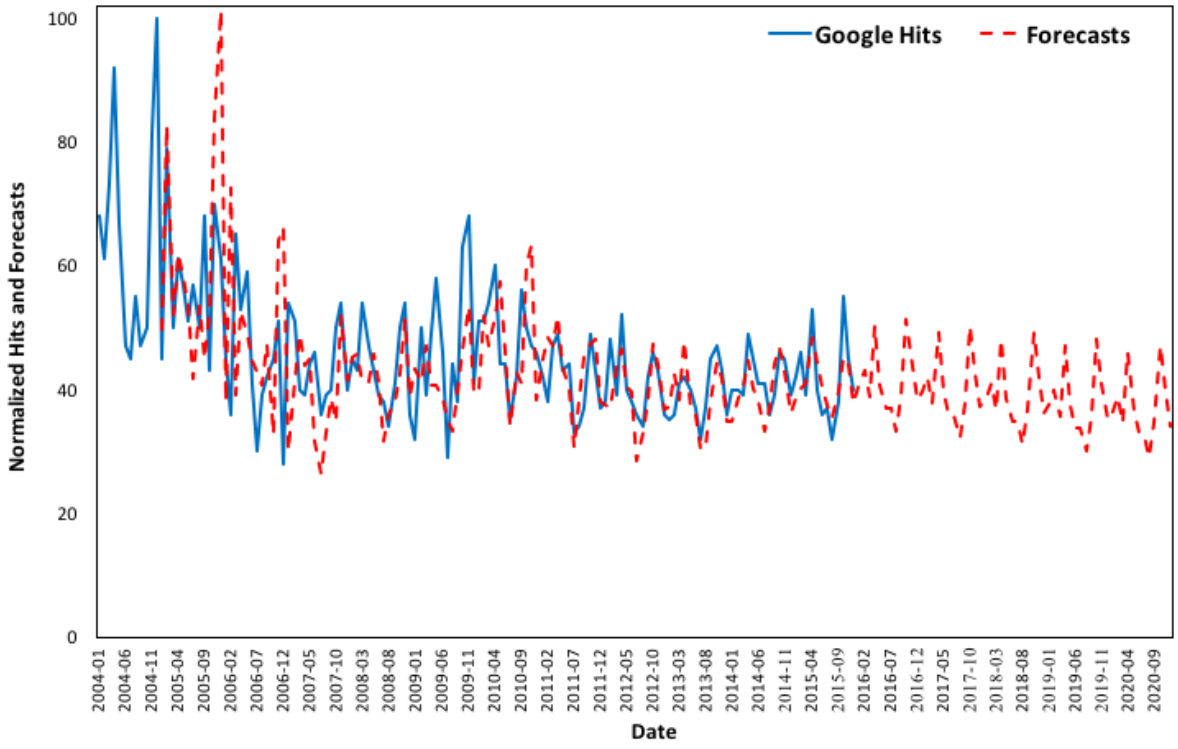


Figure B50. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Wisconsin.

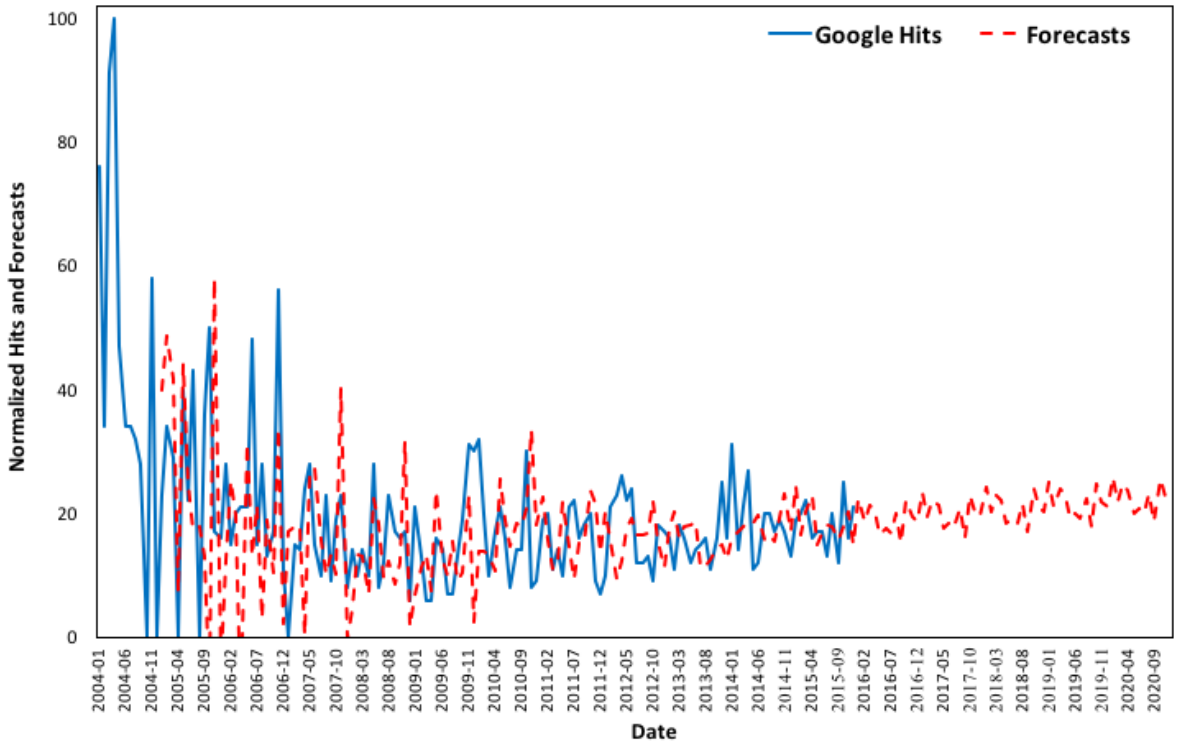


Figure B51. 'Asthma' Google Trends (2004-2015) vs. forecasts (2005-2020) in Wyoming