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).

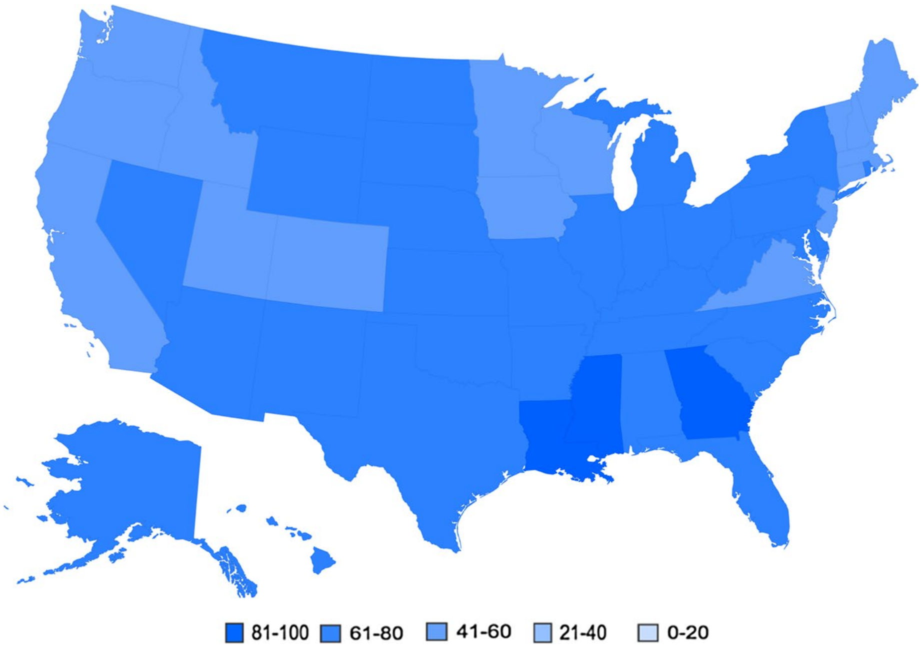

Fig. 1 Heat map of the online interest in the term 'Chlamydia' by state (2004–2016)

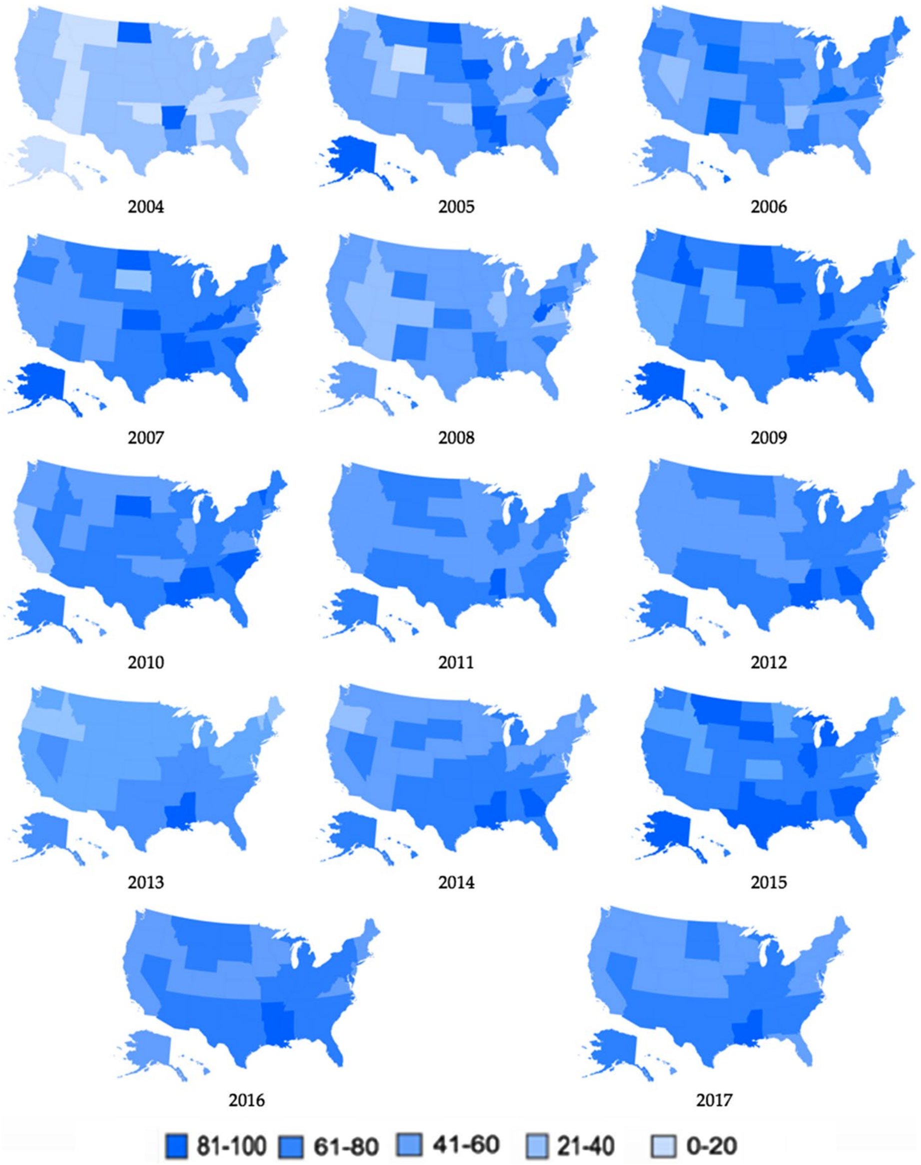

Fig. 2 Online interest heat maps for the term 'Chlamydia' by state by year (2004–2017)

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,

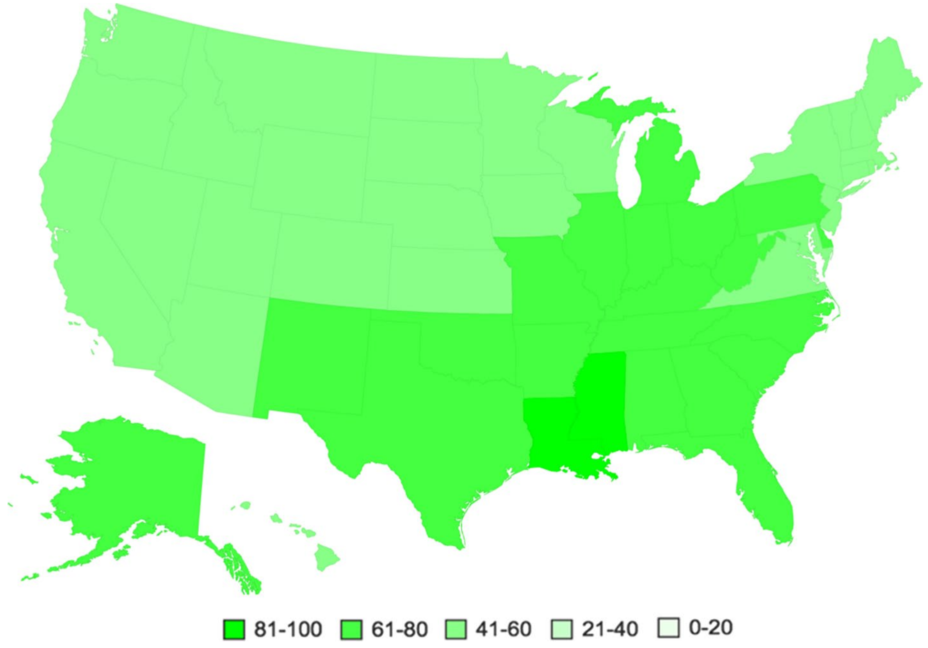

Fig. 3 Heat map of the online interest in the term 'Gonorrhea' by State (2004–2016)

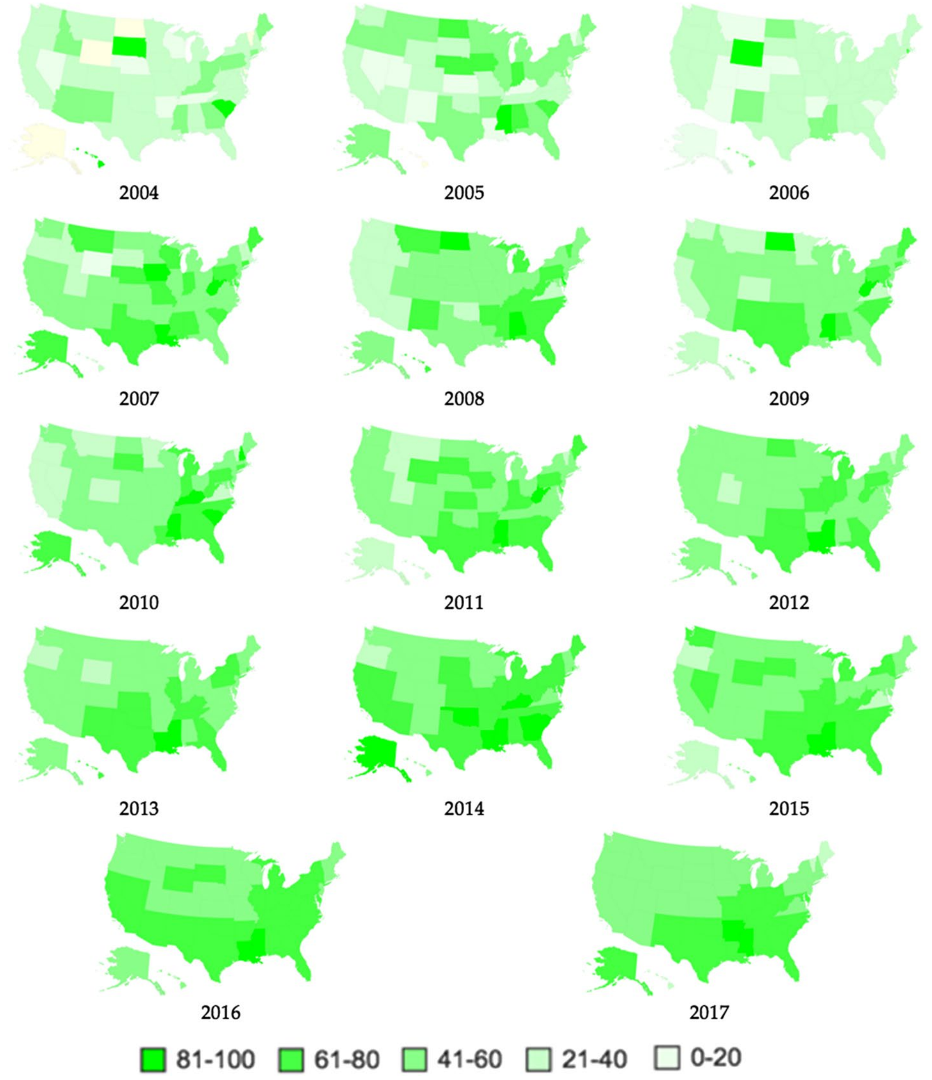

Fig. 4 Online interest heat maps for the term 'Gonorrhea' by state by year (2004–2017)

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).

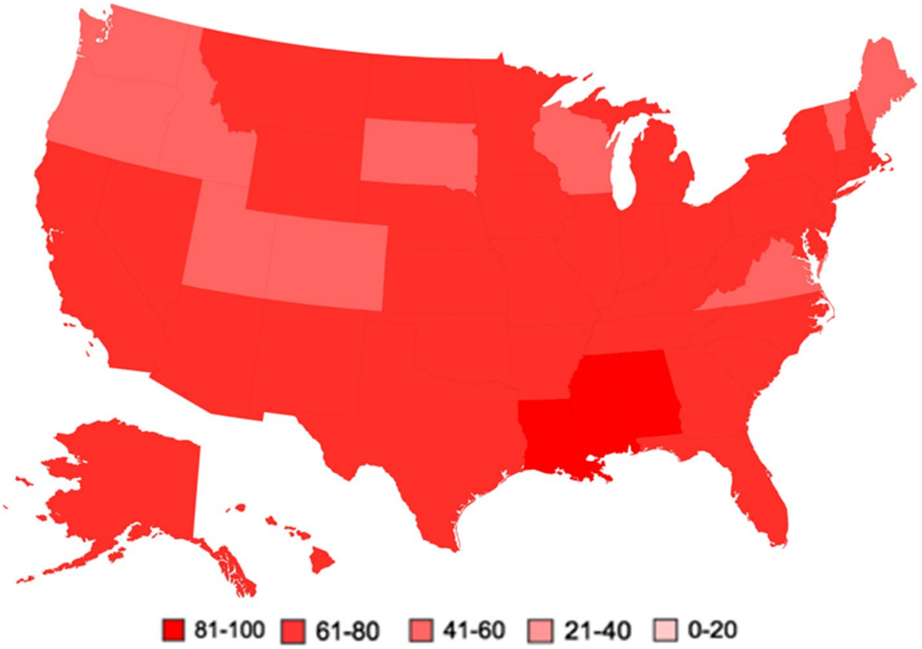

Fig. 5 Heat map of the online interest in the term 'Syphilis' by state (2004–2016)

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.

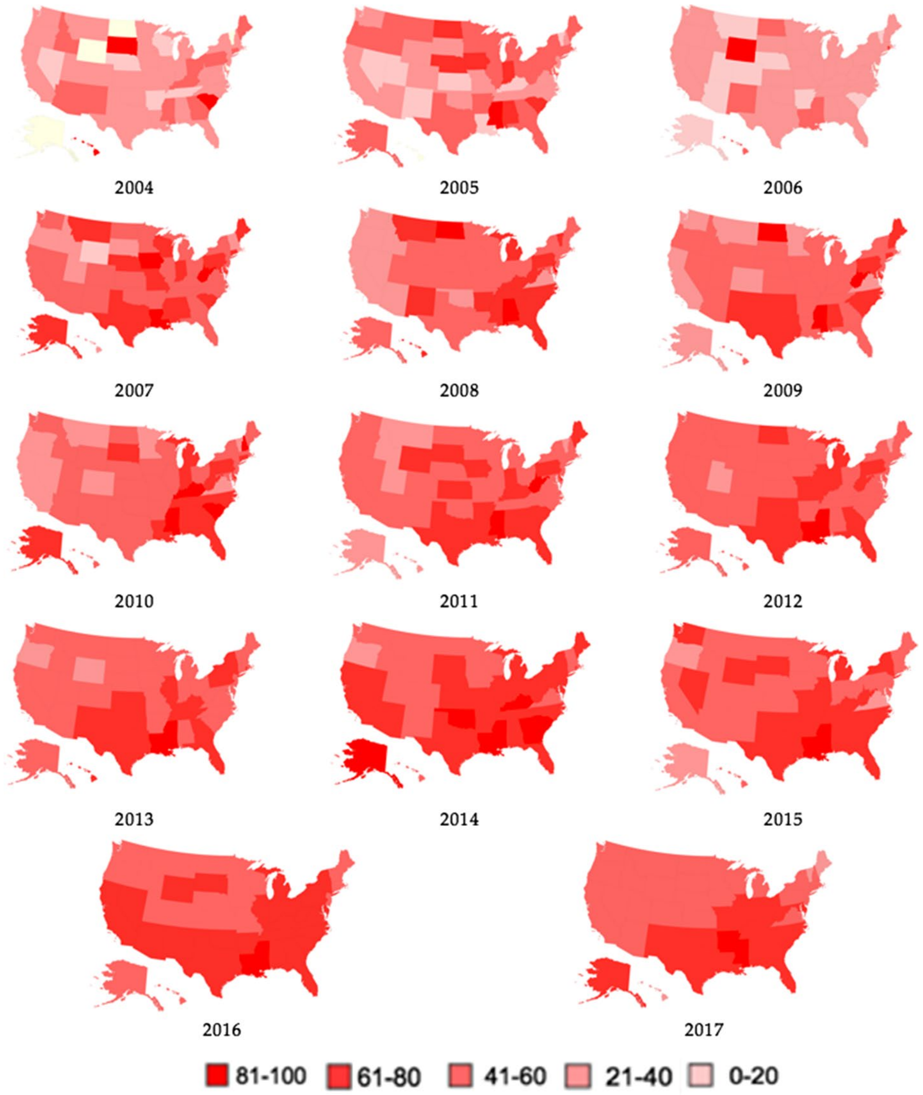

Fig. 6 Online interest heat maps for the term 'Syphilis' by state by year (2004–2017)

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

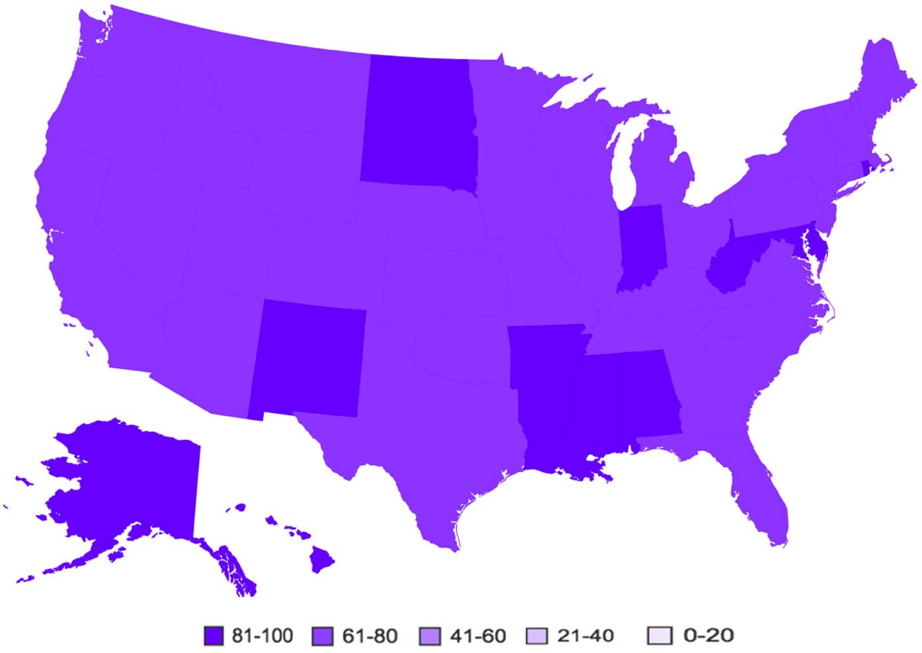

Fig. 7 Heat map of the online interest in the term 'Tuberculosis' by state (2004–2016)

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.

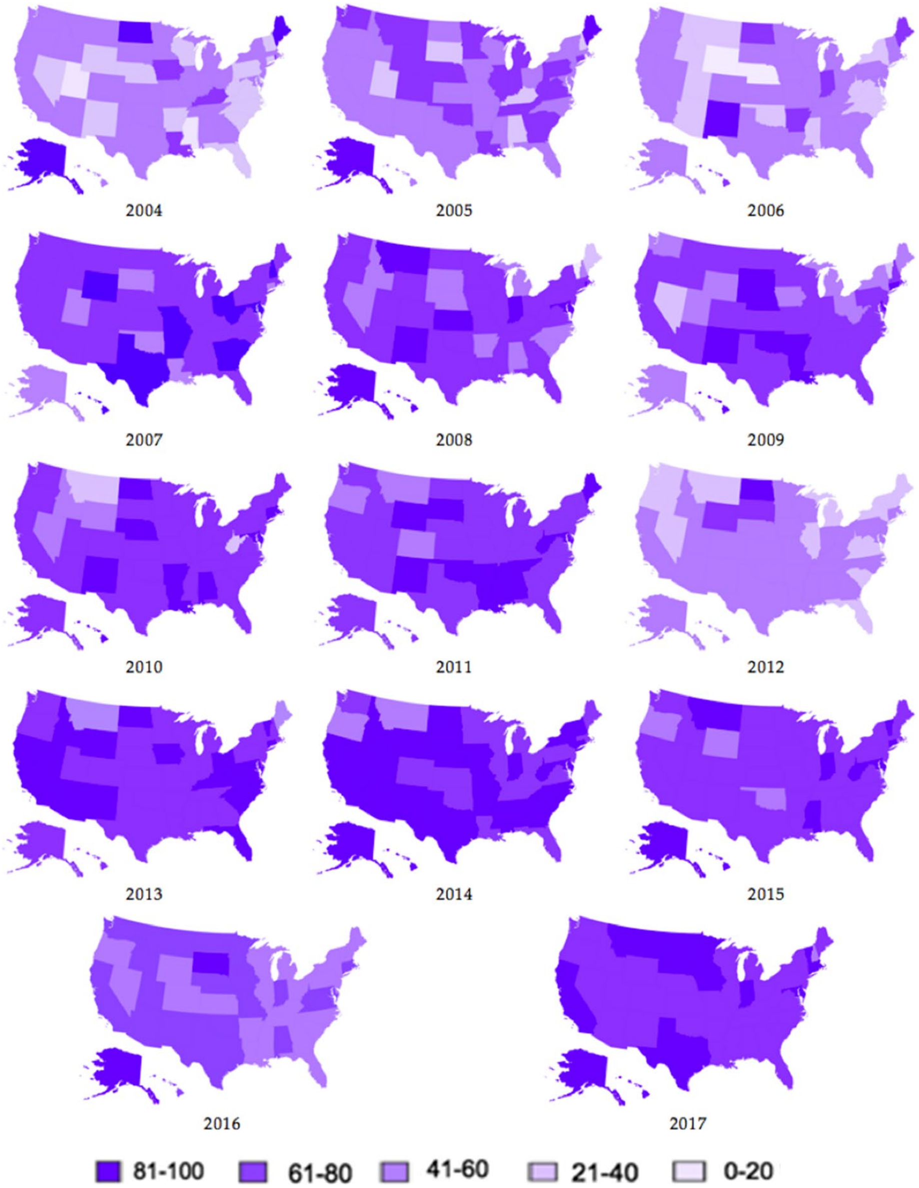

Fig. 8 Online interest heat maps for the term 'Tuberculosis' by state by year (2004–2017)

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

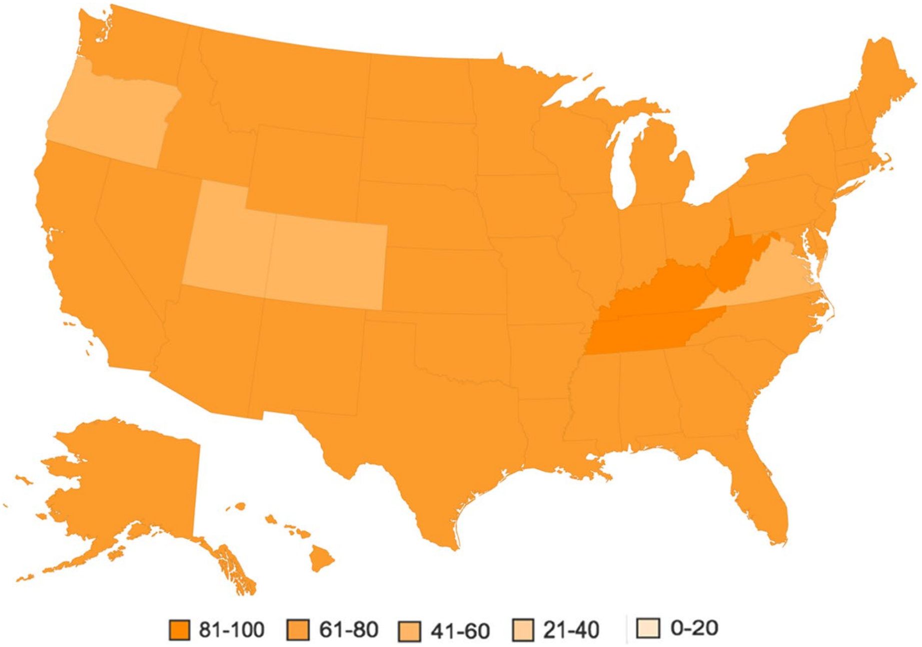

Fig. 9 Heat map of the online interest in the term 'Hepatitis' by state (2004–2015)

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 ($r=0.9045$, $\mathrm{p}<0.01$); 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 $y = 261.44\mathrm{x} - 8197.4$ 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.

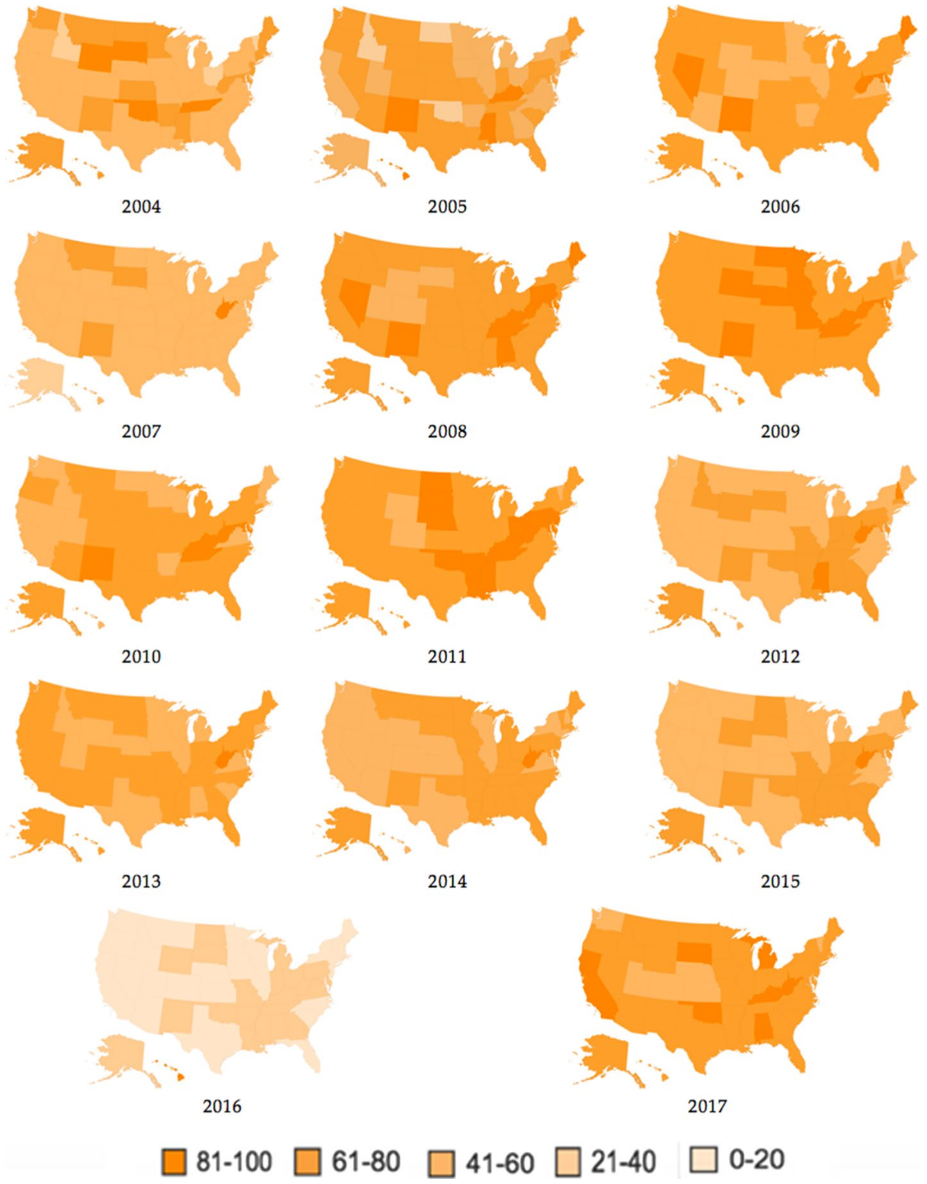

Fig. 10 Online interest heat maps for the term 'Hepatitis' by state by year (2004–2017)

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

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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.