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),

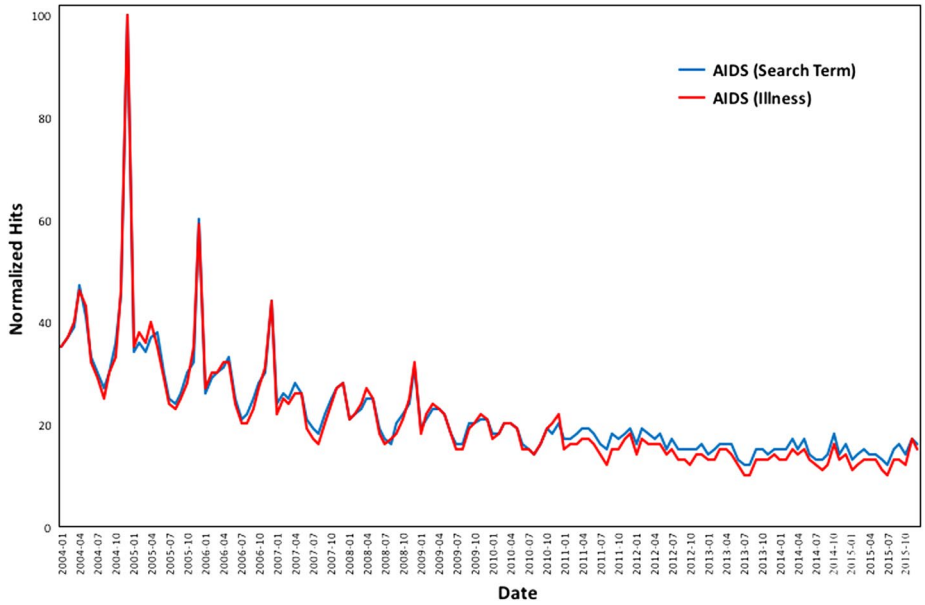


Fig. 1 Monthly Normalized Google Trends' Data for 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015

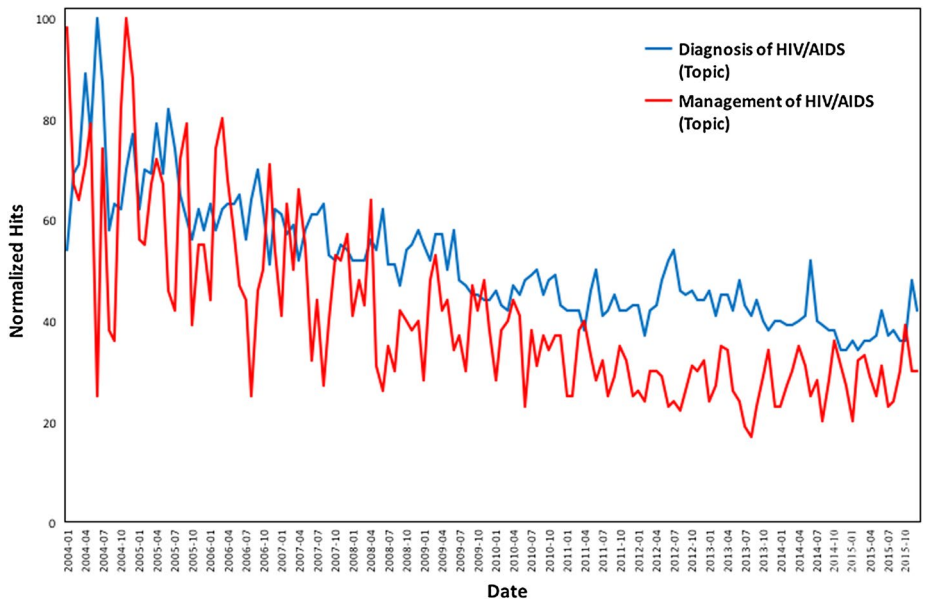


Fig. 2 Monthly normalized Google Trends' Data for 'Diagnosis of HIV/AIDS (Topic)' and 'Management of HIV/AIDS (Topic)' from January 2004 to December 2015

'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)' ($r=0.9423$,

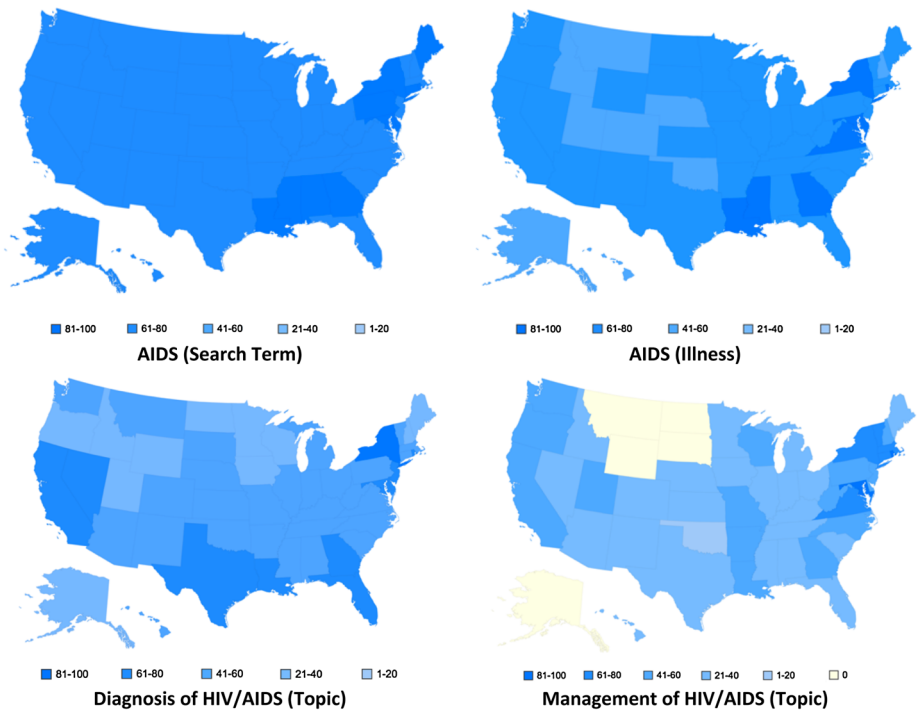


Fig. 3 Online Interest by State for 'AIDS (Search Term)', 'AIDS (Illness)', 'Diagnosis of HIV/AIDS (Topic)', and 'Management of HIV/AIDS (Topic)' from January 2004 to December 2015

$p < 0.01$). 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

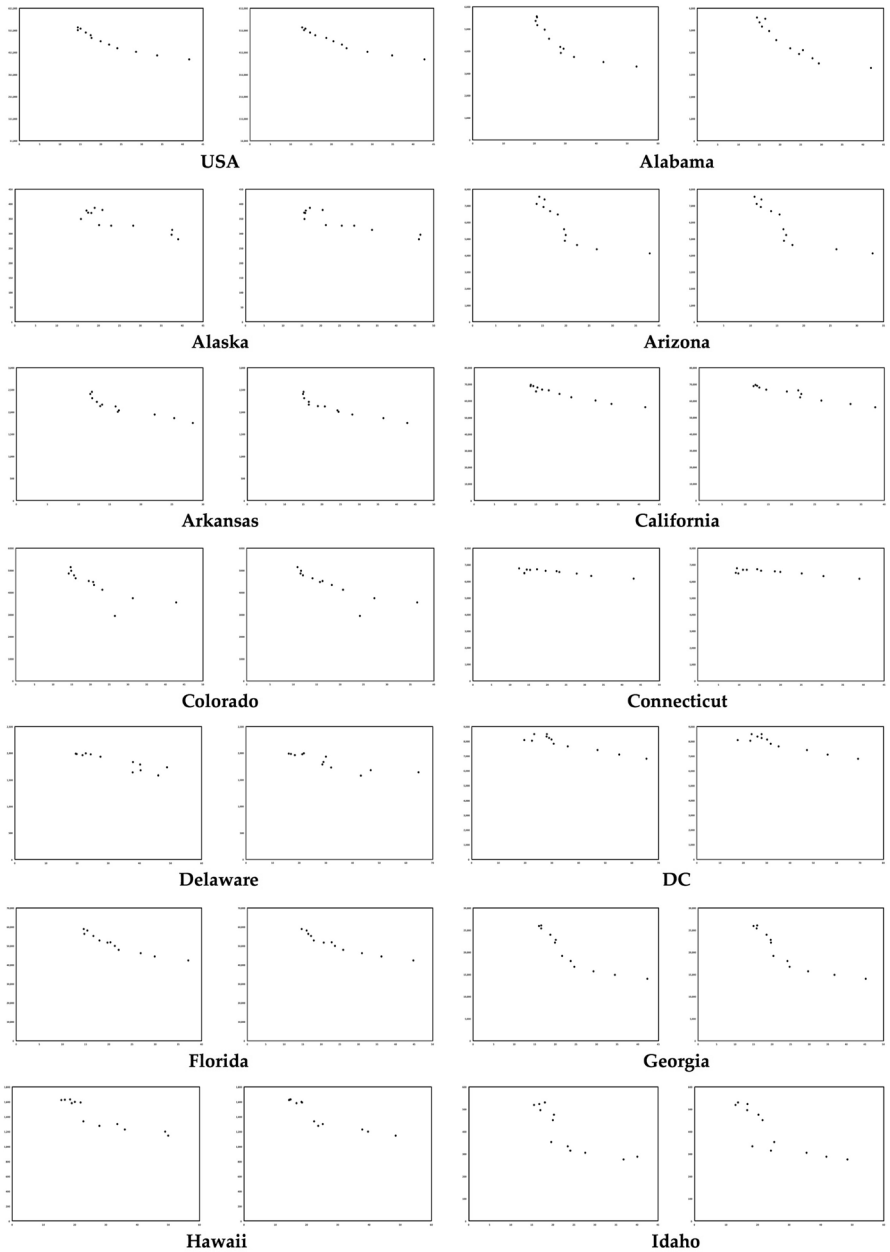


Fig. 4 'AIDS Prevalence' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015 (USA; Alabama–Idaho)

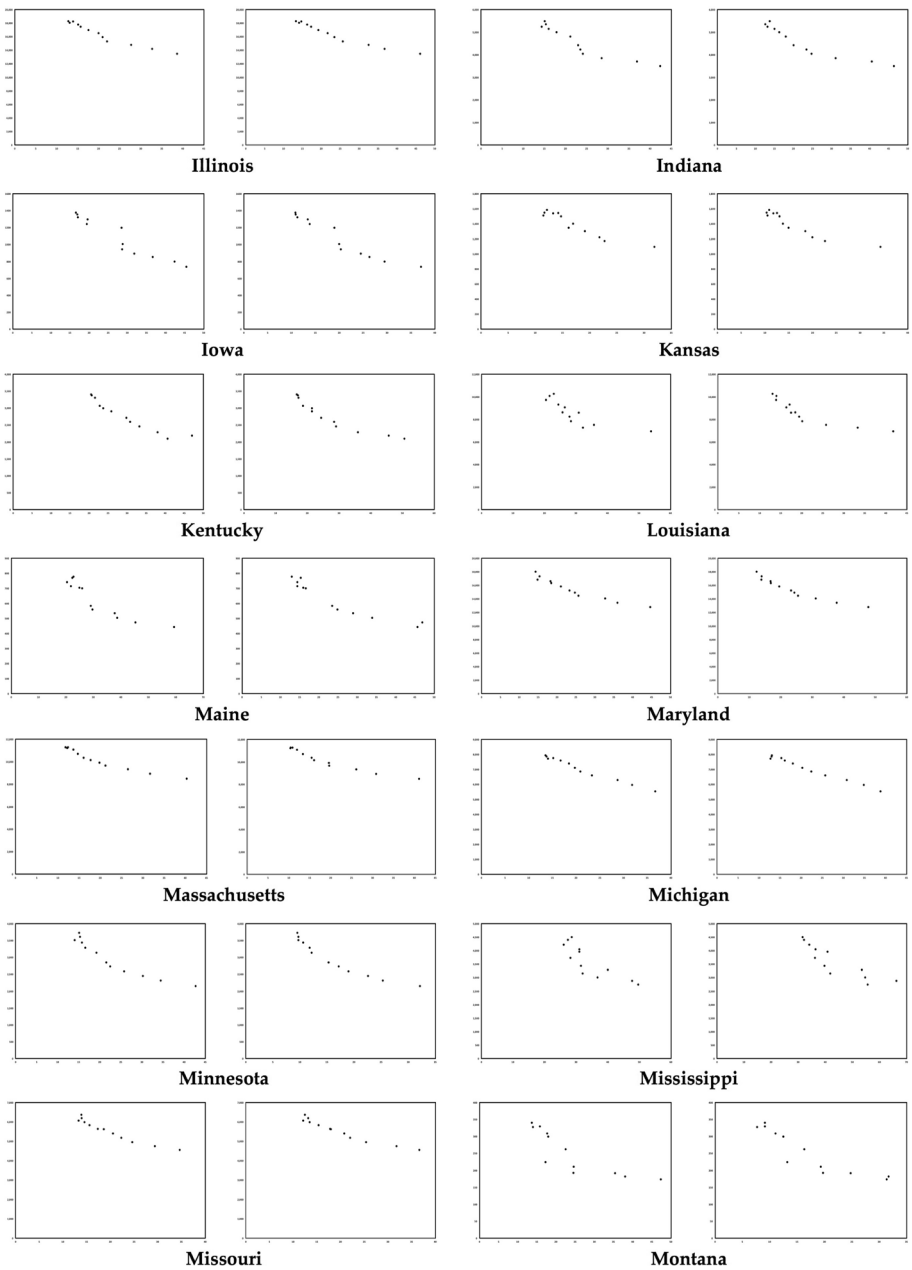


Fig. 5 'AIDS Prevalence' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015 (Illinois–Montana)

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.

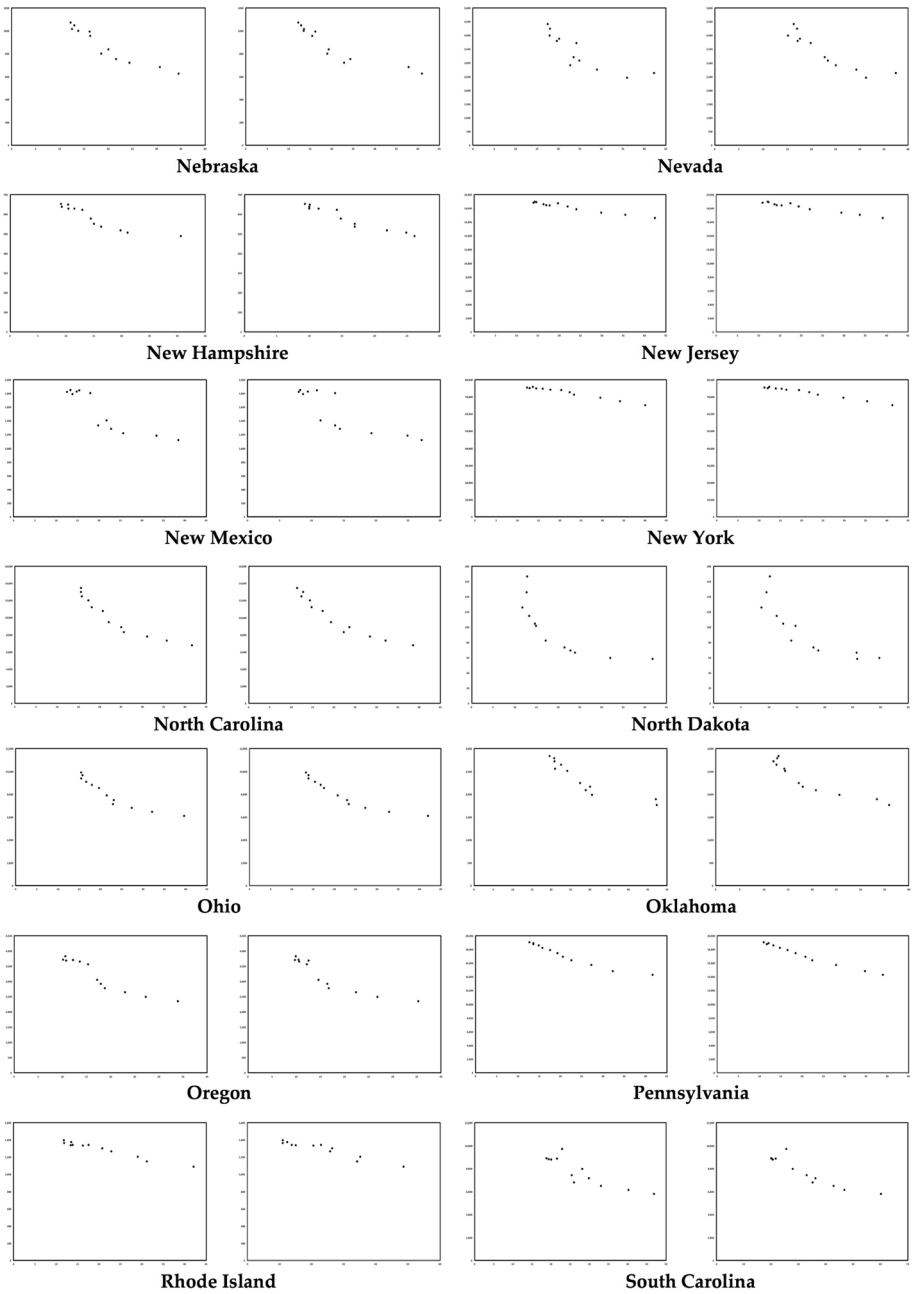


Fig. 6 'AIDS Prevalence' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015 (Nebraska–South Carolina)

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

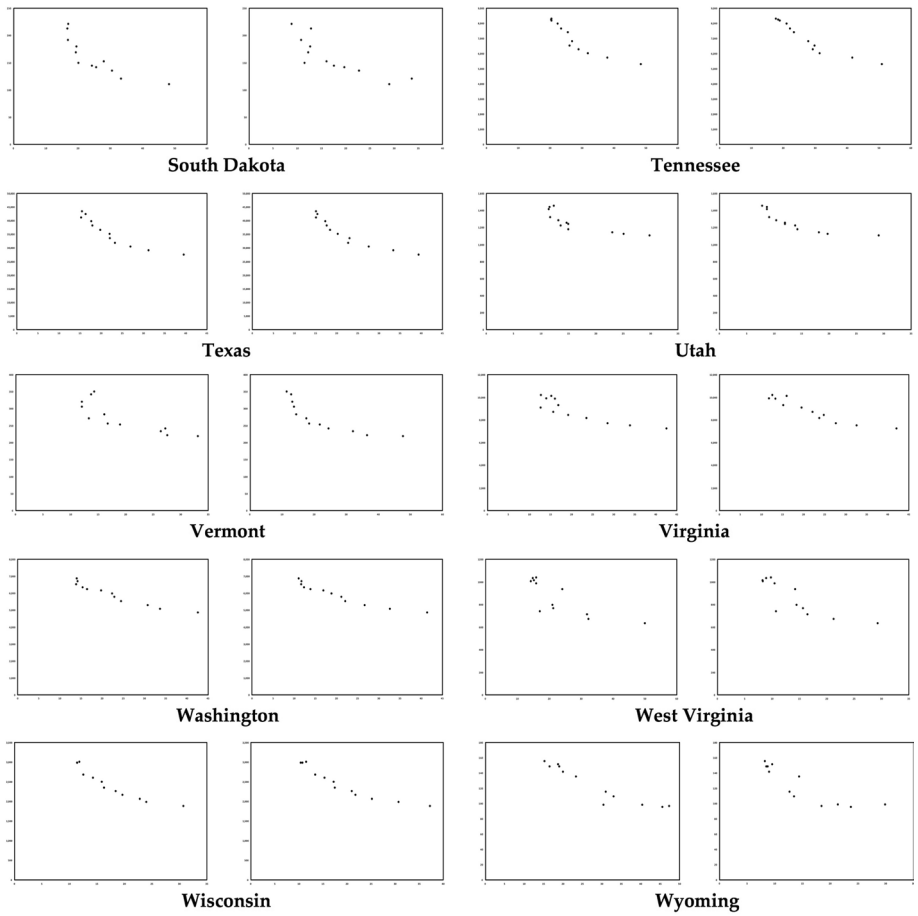


Fig. 7 'AIDS Prevalence' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015 (South Dakota–Wyoming)

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.

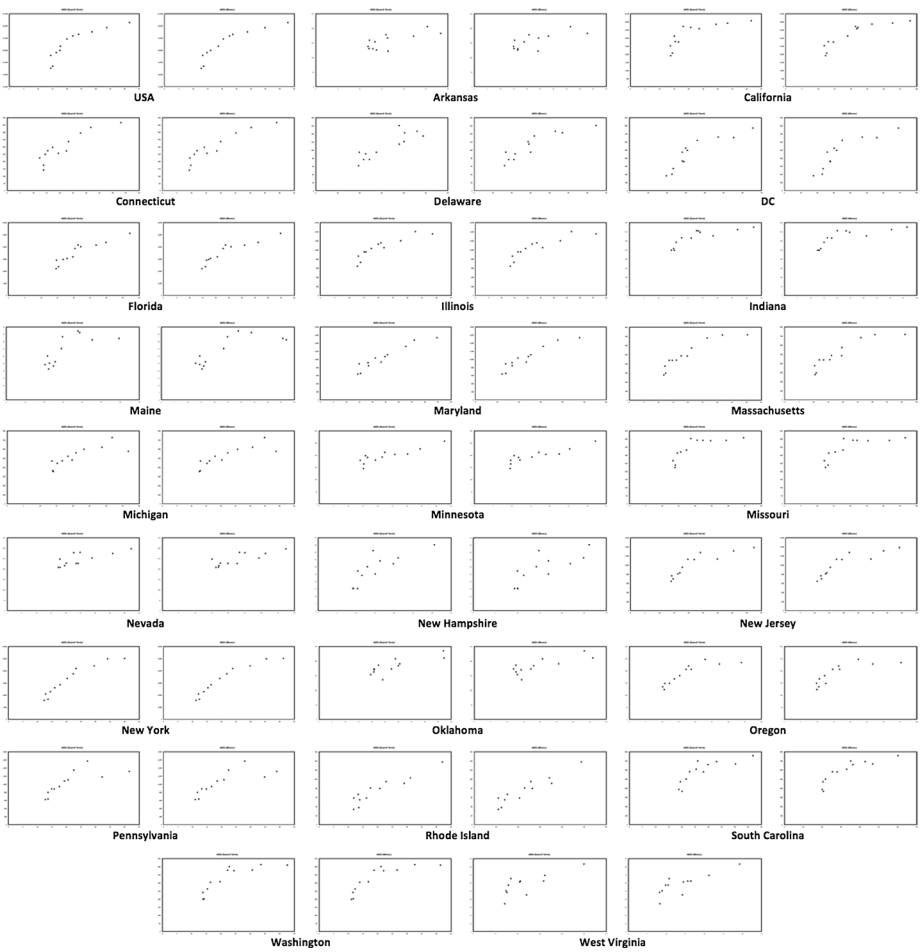


Fig. 8 'AIDS Diagnoses' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015

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.

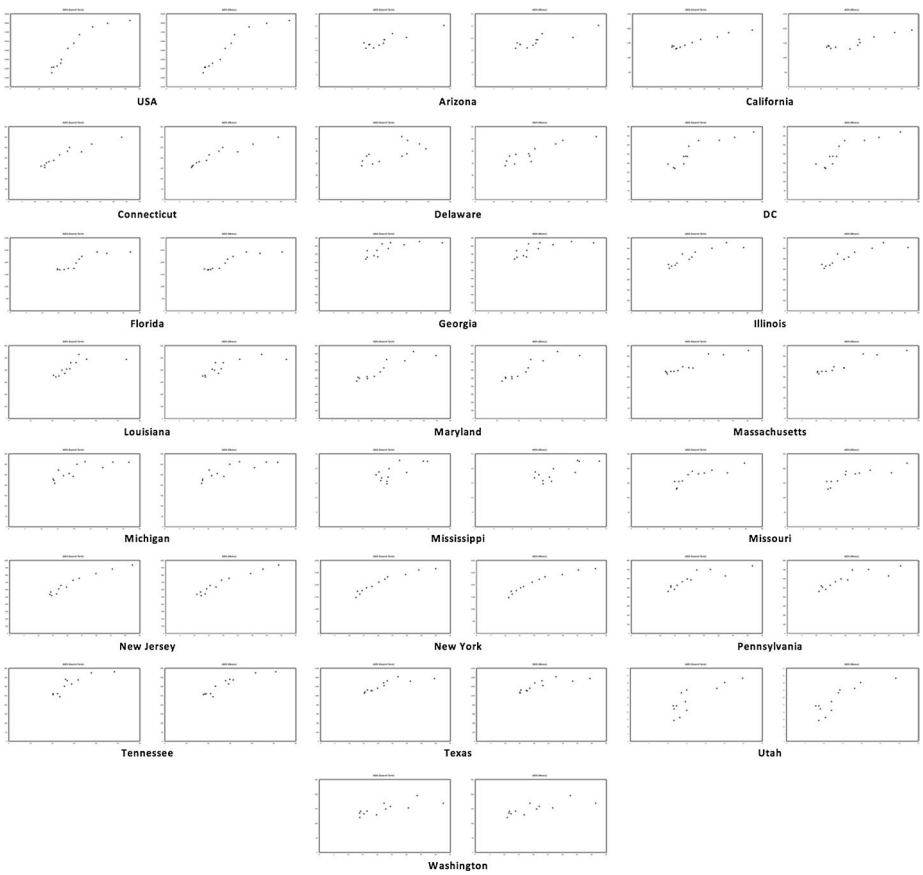


Fig. 9 'AIDS Deaths' vs. Google Trends 'AIDS (Search Term)' and 'AIDS (Illness)' from January 2004 to December 2015

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

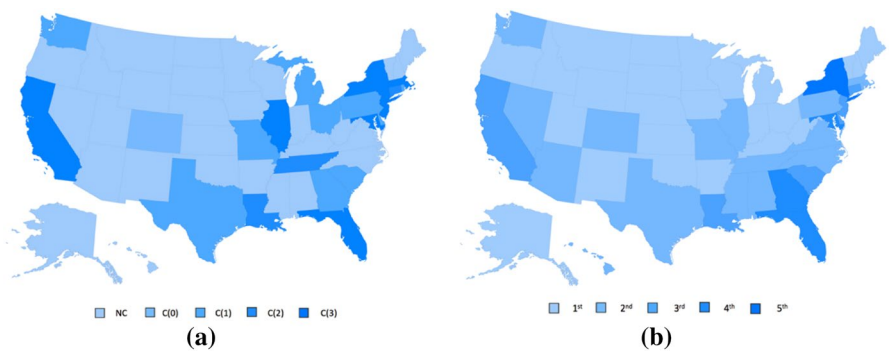
Forecasting model significance vs. AIDS rates in the US

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.