Western Michigan University

ScholarWorks at WMU

Transactions of the International Conference on Health Information Technology Advancement

Center for Health Information Technology Advancement

11-2019

The Dynamics of Real-Time Online Information and Disease Progression: Understanding Spatial Heterogeneity in the Relationship

Blake Tindol
Western Michigan University, blake.tindol@wmich.edu

Utkarsh Shrivastava
Western Michigan University, utkarsh.shrivastava@wmich.edu

Kuanchin Chen
Western Michigan University, kc.chen@wmich.edu

The Dynamics of Real-Time Online Information and Disease Progression: Understanding Spatial Heterogeneity in the Relationship

Blake Tindol
Western Michigan University
Blake.tindol@wmich.edu

Utkarsh Shrivastava
Western Michigan University
Utkarsh.shrivastava@wmich.edu

Kuanchin Chen
Western Michigan University
Kc.chen@wmich.edu

Abstract: The re-emergence of infectious diseases such as measles and polio is creating logistics challenges for the state authorities to curb their spread and contain them. (CL, 2015) Real-time surveillance of infectious diseases is important to detect possible epidemics in advance to prevent shortages of medications (FDA, 2018). The outbreak of an infectious disease creates panic in the community and is accompanied by a sudden increase in the online interest in knowing more about the disease and its symptoms. Prior studies have found a strong relationship between web-based information and disease outbreak but the influence of dynamics of web-based information in real-time is often not considered (Zhang, 2017). The dynamics or rate of change of the online interest in a disease can inform or misinform about perspective cases of the disease in a region. Oftentimes, especially in this connected world individuals overreact to the situation which may send spurious online signals regarding the disease progression. Hence, we study the relationship between the dynamics of online information and the infectious disease outbreak. We also investigate if this relationship could be influenced by regional demographic factors. We analyze weekly online interest dynamics for five infectious diseases over a period of three years across 50 states of the United States. We control for several factors (including weather, demographics, and travelers) and utilize hierarchical functional data models to incorporate real-time dynamics and clustering at the regional level. Preliminary findings suggest that online interest dynamics have a significant relationship with disease outbreak and the effect is segregated at the regional level. These findings are important to develop a system for real-time surveillance and account for the influence of heterogonous online interest during an endemic outbreak.

BACKGROUND

Infectious diseases are a risk to public health and wellness. In the year 2016, the deaths caused by infectious diseases were ranked in the top 10 leading causes of deaths worldwide, mostly occurring in low-income countries (WHO, 2016). The prediction of the progression of these diseases based on historical estimates can help save lives by preventing drug shortages and is of increasing interest in studies (M.F.Myers, 2004). This is no dearth of research on explaining the outbreak of infectious disease, however, due to their erratic transmission patterns, it is still challenging to predict outbreaks with an acceptable level of accuracy (Presanis, 2011).

The federal care organizations at the state level are authorized to take steps for preventing the outbreaks of the infectious disease and providing adequate care for the infected. However, the duration between when the infection is contracted, treatment and reporting make it difficult for them to track or predict the future cases of the disease (Jajosky, 2004). Nowadays due to the abundance of medical resources on the internet the patients are encouraged to seek the medical advice or investigate the cause of their symptoms from the online sources.

The role of online search trends related to a disease in predicting its outbreak is an emerging area of research. Prior studies have found a strong correlation between online search trends and the active cases of an infectious disease. These studies, however, did not control for a variety of factors that can influence the impact of online activity and the actual disease count. For instance, regions where patients are not tech-savvy there they might not leave any footprints on the internet. Similarly, in the regions where health services are easy to avail or are less costly the patients there may not seek medical information on the internet. It is important to understand the regional factors which might impact the relationship between online activity and disease progression for developing a robust and effective disease surveillance system.

Over the years the internet has become a reliable and cheap source of medical information while the Medicare costs have risen consistently in the United States. It is not a surprise that about a third of the online search today is for medical information. It is possible that this high cost of availing health services in drawing some of the internet traffic for medical information. For some individuals, the cost of reaching out to physicians for medical advice might to too high to immediately approach them as the symptoms appear. The financially constrained patients may prefer to first verify their symptoms and their severity from online information sources before reaching out to the physicians.

Hence, in this study, we first investigate if there are any regional variations in the relationship between online search activity of a disease and its progression. Next, we study the role of Medicare costs in a state in determining the direction of this relationship. We investigate if the higher Medicare costs strengthen the positive association between the online search for a disease and its actual cases or not. The findings from this study would help the stakeholders to better understand the dynamics of online search activity and disease progression. This could be the first step for developing a more robust disease surveillance system. While form the policy perspective the finding can have implications for the healthcare expenditures for the state.

We investigate the progression of 5 infectious diseases over a period of five years across all 50 states of the United States to answer the above questions. We control for several variables that may contribute the cases of infectious disease including weather, travelers, population density, income and age of the residents. We multilevel modeling approach to account for the unobserved heterogeneity across the states. We also allow the relationship between the search activity and disease cases to vary across each state by incorporating random coefficient hierarchical models. The next sections discuss prior work, data, methods and key findings from this study.

PRIOR WORK

Internet search queries have shown to help improve the disease prediction by several prior studies (Santillana, 2015) (Chae, 2018) (Zhang, 2017) (Fuente, 2018) (Milinovich, 2014). A survey study found that people having trouble getting access to health care might be more likely to query the internet and can help the accuracy and reliability of surveillance systems (Lee, 2015). However, no team was able to predict influenza season milestones using the online data during a challenge sponsored by the Center for Disease Control and Prevention (CDCP). The problem with forecasting models proposed by most teams during the challenge was the lack of interactions between model developers and public health decision-makers as well as a limited amount of data.

The geospatial studies have shown that there is likely an effect of long-range airline transportation on the spread of diseases (Duygu Balcan, 2009). Along with long-range geospatial effect, the short-distance spatial infectious disease spread is likely to affect school season resulting in faster close proximity spread and slower long-distance spread (Gog, 2014). As one might expect the close proximity and higher density of population have shown likely to affect disease rate spread (Hu, 2013). Different diseases have different responses in spread rate in varying weather conditions as well as other descriptive and predictive studies have shown (Chae, 2018) (Song, 2015)

There is also some evidence that individuals living in poor conditions are more inclined to obtain and transmit infectious diseases that those living in an affluent environment (Xia, 2013). Some studies have also found that the demographics of the population essentially the age group of the individuals living in an area also impact the progression of an infectious disease (Valle, 2014). These different sited environmental factors have all be used in these studies either for prediction or assessing if they are relevant factors in improving infectious disease surveillance systems. In this study, we control for most of the factors cited in the prior studies that could have an impact on the progression of the disease. The table 1 below discusses the prior work in detail.

Table 1: Literature Review

PAPER	OBJECTIVE	METHOD	Online Search	FINDINGS
Ray, E. L. (2018).	Predicting influenza progression.	Featured weighted density ensemble models	No	Component models showed more variability and ensemble methods showed slightly better average performance
Zhang, Y. (2017)	To analyze usefulness of internet search query's in pertussis surveillance.	Time Series Method (SARIMA)	Yes	Google trends information improves forecast
Milinovich, G. J. (2014).	To investigate the potential of using internet search data for early warning of a wide range of disease.	Correlation Analysis	Yes	17 Diseases were found to be significantly correlated with a search result
Biggerstaff, M. (2016).	Predicting the timing of start, peak and intensity of influenza	Multiple prediction models	Yes	No team was entirely accurate in forecasting influenza season milestones.
Chretien, J.-P. (2014).	To review influenza forecasting models	Literature Review	N/A	Comparing the accuracy of the forecasting applications in prior studies is difficult as forecasting methods, outcomes, and validation metrics varied widely.
Shaman, J. (2012).	To predict influenza disease outbreaks in New York from 2003-2008	SIRS-EAKF ensemble method	Yes	Real time skillful predictions of peak timing possible up to 7 weeks in advance
Santillana, M. (2015).	Getting improved predictions using- Google searches for influenza surveillance	Machine Learning Methods (Adaboost, SVM etc.)	Yes	Ensembles ML are more accurate than any of the algorithms alone.
Duygu Balcan. (2009)	To analyze the geospatial effect of transportation methods on global epidemics.	Generalized Linear Models	No	The spatiotemporal patterns of disease spreading are mainly determined by long-range airline transportation.
Chowdhury , F. R. (2018).	To study the effect of weather on disease spread	ANOVA test	No	Heterogeneity in impact of temperature and humidity on disease progression.
Chae, S. (2018).	To use big data to deep learning to help reduce disease reporting delays	Deep Neural Network, ARIMA	Yes	The deep learning methods DNN performed much better that time series approaches like ARIMA.
Song, Y. (2015).	To study the usefulness of weather variables in the prediction of hand foot and mouth disease.	Time series method (SARIMA)	No	Strong relationship between weather and the progression of diseases considered
Fuente, M. O. (2018).	To estimate the rate of influenza epidemics using information from public sources	GLS regression models	No	The GLS estimators performed much better than OLS estimators.

Most of the prior studies focused on a specific disease such as influenza or utilized black-box approaches (e.g. machine learning methods) to predict disease progression. To our knowledge, none of the prior studies investigated the heterogeneity in the relationship between online activity and disease progression. In this paper, we aim to observe the combination of most of the previously studied environmental factors in a more comprehensive study to find out how Medicare costs are associated with disease transmission.

DATA

To understand the relationship between the online search trend of a disease and its reported cases, we combined data from multiple databases. First, we obtained the states level disease data from the National Notifiable Disease Surveillance System (NNDSS,2019). The NNDSS system coordinates the data gathering and organization from several state-level reporting systems managed by the Center for Disease Control (CDC). We collected weekly from 50 states for five different diseases including Chlamydia, Gonorrhea, Campylobacteriosis, Salmonellosis, and Syphilis for the years 2015-2017.

To account for online search activity associated with each of the aforementioned diseases, we collected weekly google trends search results associated with each disease across all the states under consideration. The Google search trend measure associated with a disease is scaled between 0 -100 and is adjusted at the geography and topic level to make comparison easier between the terms. It should be noted however that different regions that show the same search interest don't always have the same total search volumes. The Google search data is an unbiased sample of the actual searches and only a percentage of online search sample is used to compile trends. We picked the most relevant search terms for each of the diseases, for example, the disease 'Campylobacteriosis' did not have as much information as the more commonly searched term is 'Campylobacter' which refers to the same disease. This was the only one of the diseases where the scientific name of the disease was not used in the google trends search.

The Medicare cost per capita data for each state were obtained from the Centers for Medicare and Medicaid Services (CMS,2019). The CMS works with the state authorities to monitor the distribution of the Medicaid and health insurance portability standards. The key state-level Medicare cost variables collected included, the actual per capita Medicare costs, hospital inpatient (IP) per capita actual cost and hospital outpatient (OP) per capita actual costs.

To control for the tourists arriving in the state which may lead to an increase in the count of infectious disease we collated the air travel data. The travel data was obtained from the bureau of transportation statistics website (Transstat, 2019). The dataset contains information on the number of arrivals and departures for each airport for each state in the U.S.

We also control for the weather conditions in a state during a certain week by obtaining data from the Automated Surface Observing System (ASOS) database. The information of this database covers all weather stations in each state containing weather observations on multiple variables every 20 minutes for each station. To reduce processing time for gathering the weather data we decided to pick 10 stations for each state in the U.S. that were picked on the spatial proximity to other weather stations locations in that state. We extracted two different weather variables including the air temperature and relative humidity and standardized them at the state and week level.

We also control for various state-level characteristics including the population density, median age, and median income for each state from the U.S. census bureau from the US census website. These variables are estimated at the annual level, the table xxx below describes the data used in the study.

Table 2: Variables Description

Variable Name	Descriptive Statistics	Description
Disease_Count (Count)	Min = 0, Mean =72.25, Max = 4331.00	This is a daily count of disease from the NNDSS aggregated and summed by week from 01-02015- 12-2017
Online_Trend	Min = 0, Mean = 18.02, Max = 100	Search trend for the disease in past one week.
Week	Time series	This is the week of the year
Temperature	Min = -12.4, Mean = 54.14, Max = 92.11	The average air temperature in Fahrenheit in the state over the week.
Humidity	Min = 14.17, Mean = 69.51, Max = 95.21	The average relative humidity in % over the week.
Population Density	Min = 1.292, Mean = 200.7, Max=1208.66	This is population estimates for each year divided by the area per square mile in each state
Median_Income	Min = 40037, Mean = 58966, Max = 81084	This is median income per year estimates from U.S. Census Bureau for each state.
Arrival_Rate	Min = 0, Mean = .04, Max = .25	The average number tourist visiting each week as a % of total population
Median_Age	Min = 30.10, Mean = 38.06, Max = 44.30	This is the Median Age for each state for each year from the U.S. Census Bureau
Medicare Cost	Min = 29539, Mean = 574115, Max = 2725203	Actual per capita Medicare costs
Inpatient Cost	Min = 10017,Mean =183655,Max = 766103	Hospital inpatient (IP) actual costs as a percent of total actual Medicare costs
Outpatient Cost	Min = 4111, Mean = 103517,Max = 376058	Hospital outpatient (OP) actual per capita Medicare costs.

EMPIRICAL ANALYSIS AND RESULTS

The data structure considered in this study consists of repeated weekly observations for each state on the disease count, weather, online search, state demographics, and Medicare cost variables. The observations are not independent and are clustered at the state level. One of the questions of interest in our study is the relationship between the online search trends for a disease and its actual cases in the state and how this relationship varies for a different state. To account for the multilevel structure of the data we use a multilevel model with random coefficient to investigate heterogeneity in the relationship between online interest and the actual count of infectious disease. The model equation is described below:

$Count_{ij} = \beta_{0i} + \beta_{1i} Online\_Trend \ + \beta_3 Temperature + \beta_3 Humidity + \beta_4 Population\_Density + \beta_5 Arrival\_Rate + \beta_6 Median\_Income + \beta_7 Median\_Age + \beta_7 Medicare\_Cost + \beta_8 Week + \sum_j \beta_j Disease \ + \sum_k \beta_k Year + \ \varepsilon_{ij}$ ---- (1)

$\beta_{1i} \ = \gamma_{00} + U_{0i}$ ---- (2)

The dependent variable of interest in our model is the number of cases for a disease ‘j’ in the state ‘i’ ($Count_{ij}$). We account for state-level fixed effects through the coefficient β_{0i} while the disease-specific effects are accounted by the coefficients β_j. To account for heterogeneity in the disease count cases due to unobserved year-specific event we control for the ‘Year' of observation in our model. We also control for the weather, demographics, and Medicare-related variables in our model. The level 2 model the random part U_{0i} which is normally distributed with mean zero accounts for heterogeneity in the relationship between the search trend and the weekly cases for a disease. Both level 1 and level 2 models are estimated using the restricted maximum likelihood method which has been shown to be better than the typically used MLE estimation method. We compare multiple models in table 3 and 4 below to test the robustness of the hierarchical modeling structure.

Table 3: Model Results

	Model 1 Linear Regression	**Model 2 Random Intercept**	**Model 3 Random Coefficient Fixed Intercept**
(Intercept)	810.53262	52.11749	753.46739$^{\cdot}$
	(493.60455)	(189.30897)	(398.83138)
Online interest	0.41248***	0.40728***	4.40549**
	(0.08291)	(0.08287)	(1.58178)
Week	0.11197$^{\cdot}$	0.11144$^{\cdot}$	0.08870$^{\cdot}$
	(0.06100)	(0.06086)	(0.04926)
Temperature	0.16731**	0.16907**	0.18893***
	(0.06297)	(0.06278)	(0.05083)
Humidity	-0.13684	-0.14731	-0.15480$^{\cdot}$
	(0.11394)	(0.11351)	(0.09199)
Population density	-0.60472	0.08601$^{\cdot}$	0.27498
	(0.81784)	(0.04604)	(0.66031)
Median income	-0.00005	-0.00012	-0.00023
	(0.00057)	(0.00051)	(0.00046)
Arrival rate	113.09268	119.88411	40.69653
	(124.86290)	(115.52776)	(100.80944)
Median age	-21.79922$^{\cdot}$	-3.73173	-20.96897*
	(12.38203)	(4.82108)	(10.00208)
Total per-capita cost	0.00006	0.00013***	0.00009
	(0.00009)	(0.00003)	(0.00007)
Disease Fixed Effects	Yes	Yes	Yes
Adj. R^2	0.38684	0.3925	0.807
Num. obs.	39000	39000	39000
AIC	514784.5	514997.36179	498138.46452
Num. groups: state		50	50
Var: state (Intercept)		6386.28272	
Var: Residual		31571.60789	20561.24840
Var: state online_interest			124.67191

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4: Model Results

Random Coefficient Fixed Intercept	**Model 4 (Total per-capita cost)**	**Model 5 (Outpatient cost)**	**Model 6 (Inpatient costs)**
(Intercept)	823.56289*	311.37766	898.20245*
	(399.02551)	(416.03195)	(385.67617)
Online interest	-2.37955	0.89658	-1.48232
	(1.83843)	(1.67893)	(1.90155)
Week	0.08946$^{\cdot}$	0.08938$^{\cdot}$	0.09007$^{\cdot}$
	(0.04925)	(0.04924)	(0.04926)
Temperature	0.18815***	0.18653***	0.18897***
	(0.05083)	(0.05082)	(0.05083)
Humidity	-0.15648$^{\cdot}$	-0.14849	-0.15844$^{\cdot}$
	(0.09198)	(0.09202)	(0.09196)
Population density	0.38973	0.58321	0.32557
	(0.66063)	(0.66268)	(0.65945)
Median income	-0.00021	-0.00021	-0.00019
	(0.00046)	(0.00045)	(0.00047)
Arrival rate	37.98723	45.55041	34.93340
	(100.80423)	(100.76199)	(100.79987)
Median age	-19.62189*	-8.86586	-21.44970*
	(10.00468)	(10.60499)	(9.83232)
Total per-capita cost	-0.00013		
	(0.00009)		
Online interest: Total per-capita cost	0.00001***		
	(0.00000)		
Outpatient cost		0.00005	
		(0.00026)	
Online interest: Outpatient cost		0.00003***	
		(0.00001)	
Inpatient cost			-0.00044$^{\cdot}$
			(0.00024)
Online interest: Inpatient cost			0.00003***
			(0.00001)
AIC	498140.51932	498131.06150	498144.52389
Adj. R Squared.	0.774	0.7935	0.7737
Num. obs.	39000	39000	39000
Num. groups: state	50	50	50
Var: state*online interest	86.15747	110.46377	90.89845
Var: Residual	20558.57534	20549.56109	20561.62823

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

The proposed modeling structure (Model 3 in Table 3) explains about 80% of the variations in the observed data which is significantly higher than the linear model (Model 1) and the models with state-level random effects (Model 2). The model fit statistics AIC and likelihood values also suggests that the model that accounts for the randomness in the relationship between online search and the disease count explains most variation in the observed data. The model statistics described in Table 4 confirm the interaction effect of Medicare costs on online search activity. For instance, the interaction coefficient (Online interest: Total per-capita cost) is significant and positive in Model 4. We also investigate the interaction effect of inpatient (Model 5) and outpatient cost (Model 6) per capita and we observe a similar relationship. This finding further confirms the robustness of the proposed relationship. This indicates that the states where Medicare costs are higher the effect of online activity on disease count would also be higher. Figure 1 below described the variation in the effect size of online search for a disease on disease count for low (first quartile) and high (second quartile) levels of Medicare costs across the states under consideration.

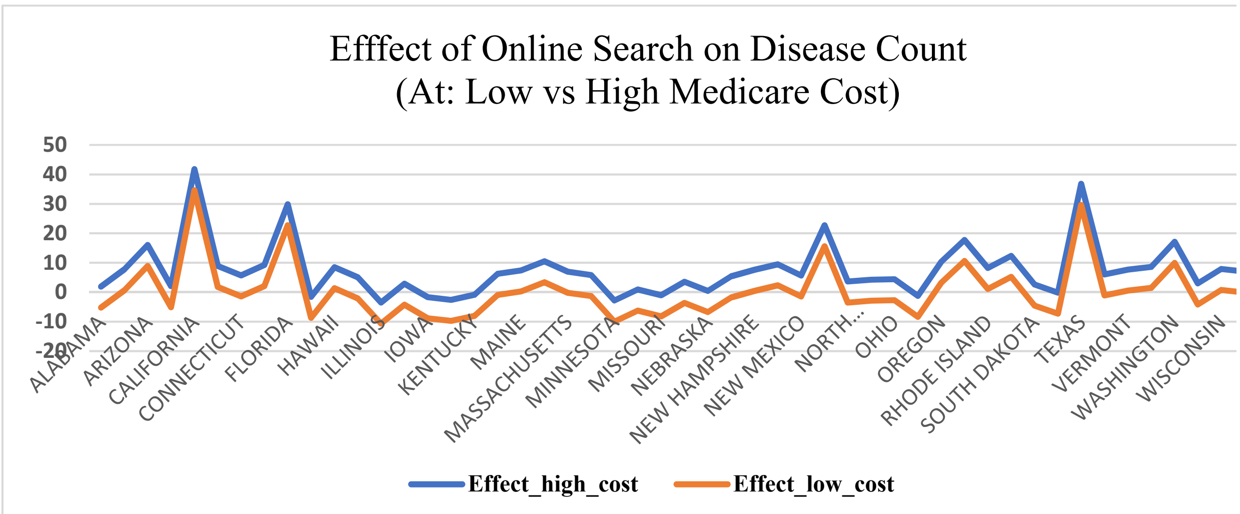

Figure 1: Effect of Online Search on Disease Count

RESULTS AND DISCUSSION

The findings of our study suggest that there is a significant variation in the relationship between the online search activity related to a disease and its actual cases in the following week across 50 states of USA. On further investigation, we found that medicare costs play is a key role in explaining this heterogeneity in the relationship across states. The states where medicare cost is higher the online search activity has a stronger influence on the actual reported cases of the disease in the coming weeks. One of the reasons for this finding could be the hesitations of the residents to avail medical services in the states with higher medicare costs. Instead to spending money on tests for medical diagnosis the residents in these states prefer to go online and look for information relevant to their symptoms. Hence, in states with higher costs of Medicare services a sudden rise in online search activity related to a disease should be taken more seriously as it might bring a sudden rise in number of disease cases.

We also find that for the five infectious diseases that we have considered the effect on online activity is not moderated by the inpatient cost. This finding reflects upon the severity of the diseases under consideration. As the diseases considered are not fatal in nature inpatients visits are not required to treat them. Hence, higher inpatient costs does not have any significant impact on the relationship between online activity and disease cases. The future disease surveillance systems should account for this heterogeneity in the effect of online search activity on disease cases for an improved forecast. The proposed linear model that accounts for state and disease level fixed effects as well as the heterogeneity in the relationship through random coefficient explains the maximum (~80%) amount of variation in the cases of infectious diseases in near future. The variation explained is about twice of that that explained by the models that only account for state level random effects.

LIMITATIONS & FUTURE WORK

We used weekly information from the CDC which is the standard for national disease surveillance systems research. This weekly CDC information oftentimes is unable to capture all the disease occurrences as only the individuals who choose to seek medical treatment at the hospitals are reported. In addition, variables such as arrival rate that we have controlled for in our models does not capture the information of number of individuals arriving in the state via road, trains or any personal transport.

The majority of diseases in this study were related to sexual based diseases and do not account for variables that could be instrumental to the spread of such diseases, for example, number of conversation interactions between users of an online dating application or number of social meeting locations in a state or city. We leave this investigation to future research. Our findings are based on search data shared by Google which is used in about 75% of global internet searches. We collected data on search trends linked to the exact names of the diseases under consideration. We do not collect trends data on other synonyms or local names of the disease that users might search. In addition, it is difficult to determine if the search trends were generated by true disease enquiries or due to associated events such as research breakthroughs or new treatment drugs. Prior studies have also argued that search data may be spurious for the diseases with high media exposure as non-patient searches would drive most of the online activity. On the other hand the extensive weekly collection of search data for five disease for over three years help control for some of the biases that may arises due to unconventional events.

Future work in this direction can understand the spillover effect of the online activity in neighboring states on the disease cases in the home state. The effectiveness of the real-time dynamics or rate of change of online activity on disease progression could also be investigated. The research also opens a potential debate on the implications of online search behavior for health care policy. How can online search activity inform heath care policy makers? Is excessive online health information seeking an indicator of inaccessible medical services in a geographical region?

REFERENCES

Biggerstaff, M. (2016). *Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge*. Retrieved from BMC: https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-016-1669-x

Chae, S. (2018). *Predicting Infectious Disease Using Deep Learning and Big Data*. Retrieved from NCBI PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6121625/

Chowdhury, F. R. (2018). *The association between temperature, rainfall and humidity with common climate-sensitive infectious diseases in Bangladesh*. Retrieved from PLOS : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6013221/

Chretien, J.-P. (2014). *Influenza Forecasting in Human Populations: A Scoping Review*. Retrieved from PLOS: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0094130

Çiğdem Ak. (2018). *Spatiotemporal prediction of infectious diseases using structured Gaussian processes with application to Crimean–Congo hemorrhagic fever*. Retrieved from NCBI PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6114917/

CL, A. (2015). The Reemergence of Measles. *NCBI*, 1.

Duygu Balcan. (2009). *Multiscale mobility networks and the spatial spreading of infectious diseases*. Retrieved from PNAS: https://www.pnas.org/content/106/51/21484.short

Eubank, S. (2004). *Modelling disease outbreaks in realistic urban social networks*. Retrieved from Nature: https://www.nature.com/articles/nature02541

FDA. (2018). *Drug Shortages*. Retrieved from FDA: https://www.fda.gov/drugs/drug-shortages/text-version-drug-shortages-infographic

Fuente, M. O. (2018). *Predicting seasonal influenza transmission using functional regression models with temporal dependence*. Retrieved from PLOS: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194250

Gog, J. R. (2014). *Spatial Transmission of 2009 Pandemic Influenza in the US*. Retrieved from PLOS: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003635

Google. (n.d.). *How trends is adjusted*. Retrieved from Google: https://support.google.com/trends/answer/4365533?hl=en

Hu, H. (2013). *The scaling of contact rates with population density for the infectious disease models*. Retrieved from ScienceDirect: https://www.sciencedirect.com/science/article/pii/S0025556413001235

Jajosky, R. A. (2004). *Evaluation of reporting timeliness of public health surveillance systems for infectious diseases*. Retrieved from NCBI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC509250/

Lee, Y. J. (2015). Access to Care and Use of the Internet to Search for Health Information: Results From the US National Health Interview Survey. *NCBI*, 1.

M.F.Myers. (2004). *Forecasting disease risk for increased epidemic preparedness in public health*. Retrieved from Science Direct: https://www.sciencedirect.com/science/article/abs/pii/S0065308X00470132

Milinovich, G. J. (2014). *Using internet search queries for infectious disease surveillance: screening diseases for suitability*. Retrieved from NCBI : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4300155/

Presanis, A. M. (2011). Changes in severity of 2009 pandemic A/H1N1 influenza in England: a Bayesian evidence synthesis. *The BMJ*.

Ray, E. L. (2018). *Prediction of infectious disease epidemics via weighted density ensembles*. Retrieved from PLOS: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005910#sec018

Santillana, M. (2015). *Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance*. Retrieved from journals.plos: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004513

Shaman, J. (2012). *Forecasting seasonal outbreaks of influenza*. Retrieved from PNAS: https://www.pnas.org/content/109/50/20425

Song, Y. (2015). *Time Series Analyses of Hand, Foot and Mouth Disease Integrating Weather Variables*. Retrieved from PLOS: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0117296

Valle, S. Y. (2014). Mathematical models of contact patterns between age groups for predicting the spread of infectious diseases. *Ncbi*, 1.

WHO. (2016). *Top 10 causes of death*. Retrieved from World Health Organization: https://www.who.int/gho/mortality_burden_disease/causes_death/top_10/en/

Xia, S. (2013). *Combating infectious diseases of poverty: a year on*. Retrieved from NCBI PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3892074/

Zhang, Y. (2017). *Monitoring Pertussis Infections Using Internet Search Queries*. Retrieved from nature: https://www.nature.com/articles/s41598-017-11195-z

Transstant, (2019). https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=2