Web Search-based Surveillance of Multiple Diseases in Multiple Countries

Nigo SUMAILA[†], Shoko WAKAMIYA[†], and Eiji ARAMAKI[†]

[†] Nara Institute of Science and Technology (NAIST)
E-mail: †{sumaila.nigo.sl8,wakamiya,aramaki}@is.naist.jp

Abstract SNS-based surveillance lacks when NLP resources or SNS data are scarce; Wikipedia-based is unreliable for widespread languages, and the current search-based is not applicable for low search-volume regions. Thus, how to conduct Internet-based surveillance when those conditions are not met is still an open problem. Our study serves as a first step in exploring the potential of conducting disease surveillance with relative search volume with sliced-timeframes. Our results show that our approach produces predictions with correlations against official patient numbers, ranging from 72% to 95% in countries with a high Web search volume. In countries with fewer Web search volume, our models produced correlations of about 62% for Lassa fever and 59% for Yellow fever in Nigeria. In the contexts $Cholera_{Nigeria}$ and $Cholera_{Haiti}$, our models yielded promising predictions of 47% and 42%, respectively. Furthermore, the results show that standard regression models are most suitable rather than neural-based models. Longer lookback windows on category "health" lessen the noise on signal and in overall produces the most stable results.

Key words Disease outbreak, search volume, machine learning, RNN

1 Introduction

Epidemic outbreaks cause high mortality worldwide. Malaria alone caused an estimated 450 thousand deaths in 2016, and most occurred in developing countries [35]. Disease surveillance is a crucial component of mitigation strategies to minimize outbreaks' burden on public health [14]. However, conventional surveillance systems have a significant strain of operating costs, and with lags of up to 2 or 3 Weeks to detect an outbreak [18]. In developing countries, the delays can be even longer due to fragile health infrastructures.

Given the massive data that people generate online, Internet-based surveillance has gained much traction as a complement to conventional systems. Opening opportunities for countries with a scarcity of resources to establish fast and affordable surveillance systems.

The widespread usage of social network sites (SNS) has made them be a recurrent target for several disease surveillance research [18; 3; 8; 7; 1]. Machine learning and NLP improve SNS data filtering by learning to recognize relevant content based on richer characteristic rather than keywords alone [3; 17; 1]. However, SNS-based surveillance have shortcomings. First, processing text data is a challenging task, and it gets much harder when dealing with low (NLP) resource languages. Second, it is becoming arduous to collect SNS data due to privacy policy changes.

Alternatively, several studies have used Wikipedia access log to conduct disease surveillance [19; 14; 29]. They adopt articles' language as proxies for location. However, this approach is coarse and unreliable [26], as many languages are widespread.

Another recurrent target of Internet-based surveillance is search volume data from Google Trends [20; 4; 34; 16; 28]. However, the prevailing search-based surveillance is only effective in regions with high search volume. Since the search volume data is relative to a long and fixed timeframe, i.e., the entire period of the study as timeframe.

In short, SNS-based surveillance lacks when NLP resources or SNS data are scarce; Wikipedia-based is unreliable for widespread languages, and the current search-based is not applicable for low search-volume regions. Thus, how to conduct Internet-based surveillance when those conditions are not met is still an open problem.

Our study proposes a different approach to search-based surveillance. We apply short and sliced timeframes that we call lookback windows. The advantage of sliced timeframe over fixed timeframe is that the former generates search volume even in low search volume regions, such as in developing countries. Moreover, the goal of the study is to assess the applicability of the sliced timeframe-based search volume for surveillance across several contexts, from the developed to developing countries.

Our contributions are to provide answers for the following research questions based on thousands of discrete experiments:

- What is the best regression model to define the relationship between the sliced-timeframe-based search volume and patient number?

- How far in the past should we look back (how long or short the timeframe) to produce the most reliable results?
- Google Trends categorizes search terms; While past studies have used category types, as best of our knowledge how much can type of the category impact on a model accuracy is still unknown. Thus, what is the category contribution?

To answer the questions above, experiments in a single country for a single disease are not enough. Therefore, we aim to target multiple diseases with different modes of transmission, biology, types of symptoms, length of incubation, seasonality and prevalence in several countries.

2 Related Work

Internet-based disease surveillance makes use of user-generated content on the Internet to deliver near-real-time disease surveillance [13]. The standard pipeline is to filter relevant data on SNS using keywords, then conduct inferences studies to capture trends and project prediction, and run validation against official data [18; 8]. However, Keyword filtering is limited, it does not capture well different contexts in which the keyword or phrase is being used. For example, a tweet containing the keyword "flu" might not be relevant to influenza surveillance when it's about news reporting and not the tweet's user being sick. [3**?** ; 17; 1] showed that machine learning and NLP could be used to categorize data for relevance based on richer characteristic rather than keywords alone.

However, SNS-based approaches have shortcomings. Applying SNS for multiple disease surveillance across numerous countries can be arduous work, and in some instances, even not applicable. First, processing text data is a challenging task, and it gets much harder when dealing with low (NLP) resource languages. Second, with constant-changing in privacy policies and increase of privacy awareness among social media users, it is becoming arduous to collect social media content. For instance, in response to the Cambridge Analytica scandal [30], Facebook announced heavy restrictions [31] on data collection. Moreover, Twitter provides unpaid access only to 1% of real-time tweets (via its streaming API) and a highly rate-limited search to seven days in the past (via search API) and prohibits sharing historical data among researchers.

Another approach is to apply SNS data as exogenous variables to autoregressive exogenous models [27], which make them contingent on the availability of past official incidence data from conventional systems. While it might not be an issue in countries with functioning conventional surveillance systems, it might render the models ineffective when those systems are inefficient or non-existing.

Studies have Wikipedia access log to conduct disease surveillance [19; 14; 29]. [14; 29] applied Wikipedia access log to global scale surveillance, and they tested their approaches in several location-disease contexts around the globe. They used article's language as proxies for location, such as solving Japanese-language article to Japan. However, this approach is coarse and unreliable [26], as many languages are widespread.

2 1 Search-based Surveillance

Google holds about 87.51% share of mobile search traffic worldwide (注1). Google Trends provides access to a largely unfiltered sample of actual search requests made to Google. It's anonymized, categorized and aggregated (grouped). For a given search term in a particular location, Google Trends provides its Relative Search Volume (RSV) as the proportion of searches related to that term against all Google Searches on that location over time period (hereafter refer to as timeframe)[23]. Thus, the RSV is scaled on a range of 0 to 100 based on the term's category, the location, and the time timeframe.

Several past studies have used Search Volume for disease surveillance. [20; 4; 34; 16; 28] obtained keywords whose Search Volume correlated with the official incidence cases. Furthermore, [6; 2] observed that Linear models fitted with a fraction of search volume for specific dengue-related queries against the official dengue case counts yielded high correlations on held-out test datasets. However, the current RSV-based surveillance have similar methodology: First, the RSV data is relative to a fixed timeframe, i.e., the entire period of the study as timeframe. For example [20; 4] used a timeframe of about 11 years. This approach applies to developed world or regions with a high search volume, in which the signal of health-related searches relative to a long and fixed timeframe is strong enough to appear in Google Trends. However, in developing countries or regions with low search users, a fixed timeframe may not even produce RSV data because the search volume may not be above the threshold that Google Trends considers a popular search for that period. Second, they had a similar purpose, which was to assess if RSV of search terms related to their disease of interest correlated with official incidence data.

In summary, SNS-based surveillance lacks when NLP resources or SNS data are scarce; Wikipedia-based is unreliable for widespread languages, and the current search-based is only effective in countries or regions with high search volume. Thus, how to conduct Internet-based surveillance when

（注1）：https://www.netmarketshare.com/search-engine-market-share.aspx those conditions are not met is still an open problem.

3 Material

3 1 Gold Standard Data: Patient Numbers

We sought to test several epidemic and pandemic-prone diseases in multiple countries, from developing to developed countries, in various climates, and with different level of Internet coverage. Similar to [14], we also sought to target diseases with diverse modes of transmission (e.g, airborne droplet, vector, sexual, ...), biology, types of symptoms, length of incubation, seasonality, and prevalence [14]. Additionally, we needed reliable gold standard data, and with high temporal granularity (preferably, weekly or daily incidence values) for at least close to two years-long of reports. However, such data are not always publicly available. Table 1 shows all diseases and their locations that make up our gold standard data. In total 8 disease-location contexts were analyzed.

3 2 Search Volume Data: Relative Search Volume (RSV)

3 2. 1 Search Term Selection

We applied three different techniques together to gather the terms we used to collect search volume data related to diseases of interest. Firstly, We used crowdsourcing (Amazon Mechanical Turk, Mturk (注2)) to ask people about what search terms they have used or would have used to search on Google about our targeted diseases. Then, from medical references, we collected the most common symptoms associated with each disease. Lastly, we added more related terms from Google Trends suggestions. Table 2 shows the list of all terms by contexts and types.

3 2. 2 Data Pulling Strategy

we defined four types of lookbacks: 7 days, 14 days, 30 days and 90 days. For each week in our Gold standard dataset for any context, we collect the past 7 days, 14 days, 30 days and 90 days of RSV data for each search-term defined on the Table 2. We collect the data two times, first using setting the category to “all” and second time setting it to “health”. We used an unofficial API for Google Trends (注3) that allows simple interface for automating downloading of reports from Google Trends.

4 Method

We sought to appraise the applicability of RSV with sliced-timeframes for multiple diseases surveillance. Furthermore, We aim to find the relationship between RSV with short timeframes and disease incidence cases. The present section describe the approach We use to find such relationship.

4 1 Task Definition

Let us define the disease surveillance problem as a regression task. Denote by $X_t \in R^{n \times T}$ the RSV for a given context at the time t, where n is the number of search terms, and T is the size of the timeframe or Lookback Window, e.g., RSV at week t of terms “flu” and “fever” for past 7 days. Thus, the task is to predict the incidence cases y of a disease associated with the search terms at a future time point $t+h$ through a regression function f, as shown on Figure 1, where h is the horizon of the prediction, and $h = 0$ is nowcasting and $h > 0$ is forecasting. For the size of T we considered 7 days, 14 days, 30 days, and 90 days.

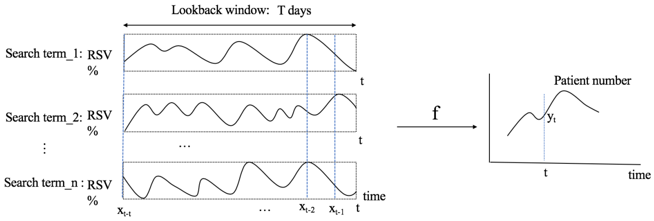


Figure 1: Proposed approach

4 2 Standard Regression Models

To define a f that describe the relationship between the RSV and disease incidence cases, we first apply standard regression models such as Lasso, Ridge (and bayesian Ridge), and support vector regression. Lasso and Ridge’s regressions are linear models. Linear models are simple, yet they provide an adequate and interpretable explanation of how the features affect output, they can sometimes outperform more complex nonlinear models, particularly in situations with small numbers of training samples, low signal-to-noise ratio or sparse data [15]. Similarly, Support Vector Regression algorithms come from SVM (Support Vector Machine), and SVM algorithms can generalize the unseen data efficiently in a high dimensional feature space [32; 22].

As we defined in the previous section, the search volume at timestamp t is represented as $X(n, T)$, and before applying into the regression models, we transform the $X(n, T)$ matrix into one dimension matrix $X(1, n * T)$, and consider each value in the matrix as an independent feature, since these models accept the input in the format of one dimension matrix.

4 3 Neural-Based Regression Models

In the previous point, we define the f from Figure 1 as a standard regression models to describe the relationship between RSV and patient numbers. We transformed the search volume matrix into one dimension matrix and assumed each

（注2）：https://www.mturk.com/help
（注3）：https://github.com/GeneralMills/pytrends

Table 1: Gold standard data with a total of 8 disease-location contexts. Include min, max and mean of patient counts.

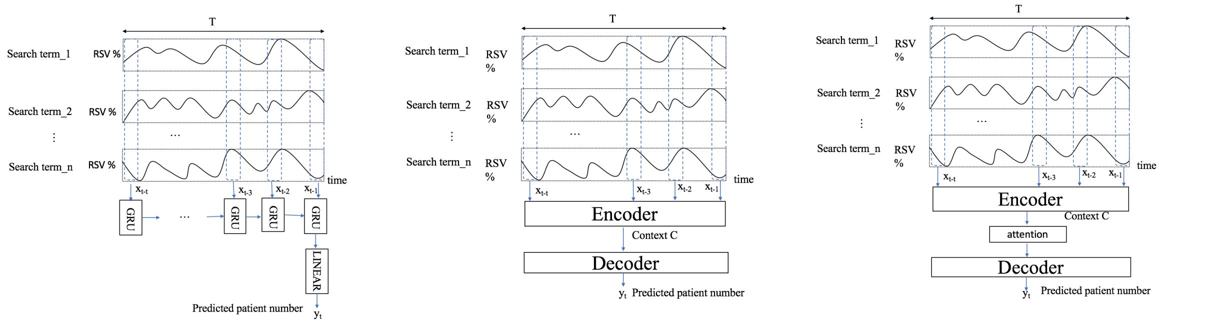
Disease	Country	Start	End	Resolution	Min	Max	Mean	Source
Cholera	Haiti	12-05-2010	12-05-2012	Daily	0	3687.28	51	[14]
	Nigeria	03-20-2017	12-30-2018	Weekly	0	2929	589.02	[21]
Dengue	Brazil	07-03-2010	03-16-2013	Weekly	419	84144	17268.74	[14]
Influenza	Japan	06-26-2010	07-05-2013	Weekly	0	1668	173.79	[14]
	Poland	10-17-2010	10-23-2013	Weekly	0.0	1668	1222.73	[14]
	US*	01-01-2011	01-10-2014	Weekly	0.0	6.06	1.74	[14]
Lassa Fever	Nigeria	01-08-2011	12-30-2018	Weekly	0	70	10.49	[21]
Yellow Fever	Nigeria	02-07-2017	12-30-2018	Weekly	20	112	51.60	[21]

* we used weighted numbers instead of the actual number of patients.

Table 2: Search terms by context and types.

Context	term types		
	name-based and related-name-based	*symptoms-based*	*generic**
$Influenza_{US}$	influenza, flu	fever,chills sweat,fatigue weakness sore throat,aching muscle headache, persistent cough	symptoms, shot treatment, vaccine signs
$Influenza_{Japan}$	インフルエンザ，インフル 風邪， かぜ	高熱，寒気	治療， アウトブレイク，症状 タミフル，リレンザ
$Dengue_{Brazil}$	dengue, chikungunya febre amarela, malaria	febre alta, enjoos, vomito, dor de cabeça manchas, coceiras, pontos vermelhos pele dor atrás olhos, mal-estar e cansaço	sintomas, sinais, tratamento vacina, propolis
$Influenza_{Poland}$	grypa	goraczka,ból głowy, chłód, ból gardła, zmeczenie	objawy, leczenie, szczepionka
$LassaFever_{Nigeria}$	lassa fever, yellow fever malaria	fever, headache, myalgia nausea, vomit, fatigue, diarrhoea sore throat, malaise	symptoms, therapy outbreak,treatment
$YellowFever_{Nigeria}$	yellow fever, lassa fever malaria, chikungunya	fever, headache, myalgia, jaundice nausea, vomit, fatigue, diarrhoea sore throat, malaise	symptoms, therapy outbreak,treatment
$Cholera_{Nigeria}$	cholera	nausea, vomit, diarrhoea ,watery diarrhoea	symptoms,therapy treatment, outbreak
$Cholera_{Haiti}$	choléra, cholera	nausée, vomi, diarrhée, diarrhée aqueuse	symptômes, thérapie traitement, épidémie

*: We use the character space to form term pair-wise between name-based terms and generic terms, e.g., “flu” and “shot” to become “flu shot”



GRU-based RNN

GRU-based enc-dec

GRU-based enc-dec with attention mechanism

Figure 3: Neural-based regression models.

value in the matrix as an independent feature. On the contrary, in this section, we imply that the elements of the RSV sequence are related to each other and their order matters. Therefore, we attempt to use a class of neural networks that can preserve that temporal dependency of our inputs (RNN, encoder-decoder and att. encoder-decoder).

We use GRU-based RNN (Gated Recurrent Unit) [9; 10; 11] to capture the temporal dependencies within our RSV sequence, Figure 2a. For encoder-decoder, we use a simple architecture similar to that proposed by the original authors, [9] and [33]. Having a simple architecture is more suitable to our problem because we deal with data-deficient situations for most of our contexts (country-disease). Furthermore, in our architecture, as we show in Figure 2b, we use GRU-based RNN as both the encoder and the decoder.

The attention mechanism improves the performance of enc-dec in longer sequences [**?**] by focusing much more at specific parts of the encoded-sequence when decoding it. In our problem, we attempt to use the attention mechanism to force the decoder to give adequate focus at specific parts of the encoded RSV sequence.

5 Experiment

In this section, we describe the procedures we took to implement our models, the variety of experiment we conducted, and how to evaluate the results.

5 1 Implementation Setup

We attempt to establish the relationship between the RSV and disease incidence cases first applying standard and robust regression models such as Lasso, Ridge (and bayesian Ridge), and support vector regression. Moreover, to implement those models, we utilised Scikit-learn tool [（注4）]. The results we report are from the default configuration for each method.

（注4）：https://scikit-learn.org/stable/

Alongside using standard regression models, we also implied that the elements of the RSV sequence were related to each other and their order mattered. Using PyTorch [25], we implemented a GRU-based RNN, an encoder-decoder, and an encoder-decoder with an attention mechanism. To train these models, we used mean squared error (MSE) with a mean reduction, applying Stochastic Gradient Descent (SGD) optimizer with a learning rate set to 0.001. We used 30 hidden states for the GRU cells, dropout set to 0.2.

5 2 Evaluation

For evaluation, we used a moving window approach with one-week held-out, starting at 50% of the datasets, i.e., we initially train with the first 50% and predict one week ahead. Add that week to the train set and move one step forward and repeat the process. We use Pearson's Correlation Coefficient (CORR) to compare our predictions against the gold std data.

We heuristically created several groups of the search terms (see Table 2), i.e., we take a search term set (e.g., name-based search terms) alone and consider it one group, combine it with other search terms set and consider it a different group, and so on. Per each search term group in a given context, we evaluate in both categories ("all" and "health"), for all standard regression models. In each category, we try all the four lookback windows (7 days, 14 days, 30 days and 90 days). For the case of neural-based regression models, we did not run an exhaustive evaluation. Instead, in each context, we chose settings (search term group, category and lookback) based on the performances from standard regression models, highest for each case.

6 Results and Discussion

Table 3 shows a summary of best performances from each lookback, category, and for all contexts from an exhaustive evaluation using standard regression models. Our results show that in general, models built with RSV with short timeframes can predict the patient number in high volume Internet users countries and moderate successful predictions in countries with low volume Internet users. Figure 8b shows the results from a Bayesian ridge regression trained with name-based search terms and a lookback of 7 days predicted with success about two Japanese influenza's seasons even though that by the first predicted season the model had only seen one peak season, and that shows the model generalized quite well. In the contexts $influenza_{US}$ and $dengue_{Brazil}$, Figure. 8a 8c, our models produced predictions overall close to the official patient numbers.

$Influenza_{Poland}$ and $LassaFever_{Nigeria}$ (Figure 8d and 8e) present fascinating cases, in both, our models failed to catch the height of the outbreaks but predicted peaks similar to those previously known. These cases could probably be some isolated incidences caused by noise on the signal or a limitation of our approach. By definition, RSV of a given term for a specific period is the query share of the term normalized by the highest query share of that term over the specified period. In our assumption, two outbreaks peaks with different intensity could produce similar RSV leading to fault prediction such as those we observed in Figure 8d and 8e. However, to assert this assumption, more contexts with similar data characteristics are required.

In developing countries or rural areas, fewer people have access to Internet services. Moreover, as accounted by [6; 5], Web-query based surveillance depends on sufficient Web search volume in order to be statistically representative of the country or area health situation. Taking that into perspective, we consider our results from $LassaFever_{Nigeria}$ and $YellowFever_{Nigeria}$ to be relatively successful results despite that our models produced correlations bellow 70 %. Our models failed to produce correlation above 50 % in the contexts $Cholera_{Nigeria}$ and $Cholera_{Haiti}$. The data from Haiti are from right after the 7.0 Magnitude earthquake [12] in 2010 which left the country with about half of the infrastructure damaged including communication systems, hence decreasing the already-small number of people with an Internet connection.

Another limitation, not particular to our approach but to the Web-based surveillance, in general, is that even if we capture the real signal from sick people, we can not ensure what the next action of the individual is going to be. For example, [24] found that in a rural area in Cambodia, about 67% of cases of hemorrhagic fever were treated at home instead of at a "health" clinic. Self-treatment do not make it to official patient number recorded by the government, and this can lead to problems in Web-based surveillance systems.

6 1 Regression Model Type

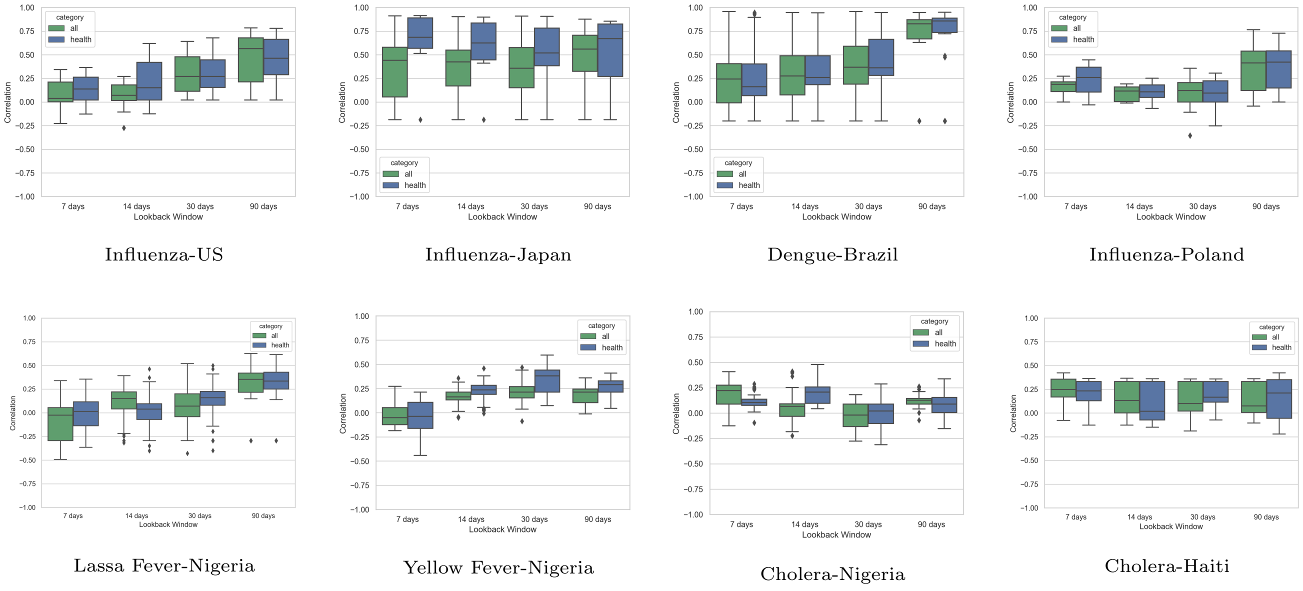
We attempted to use two classes of regression models to interpret the relationship between RSV and patient numbers. Table 4 presents a summary of the performance of neural-based regression models. It distinctly shows that they are not adequate models for this problem; the data size is too small for neural networks. Moreover, we have left with standard regression models.

On the previous point, we analysed Figure 7 from the category perspective, and now we are going to look at it from the regression model standpoint. One impression that we can have from it immediately is that all regression models have close maximum values (top of the interquartile), which means that any regression model can reach the top performance at a specific configuration (search terms, lookback and category). However, for the performances in all config-

Table 3: Top results by context, category and lookback windows

context		Category: all				Category: health			
		Lookback				Lookback			
		7 days	14 days	30 days	90 days	7 days	14 days	30 days	90 days
$Influenza_{US}$	corr	0.343	0.270	0.641	**0.843**	0.364	0.618	0.677	0.828
	model	RR	B-RR	SVR_{linear}	LR	B-RR	SVR_{linear}	B-RR	SVR_{linear}
	$search_{terms}$	name	all	generic	all	name + sympt	generic	generic	name
$Influenza_{Japan}$	corr	0.910	0.904	0.907	0.875	**0.914**	0.897	0.905	0.855
	model	B-RR	B-RR	B-RR	B-RR	B-RR	B-RR	B-RR	RR
	$search_{terms}$	name*	name*	name*	all	name	name	name	name
$Dengue_{Brazil}$	corr	**0.957**	0.947	0.955	0.946	0.947	0.943	0.945	0.950
	model	B-RR	B-RR	B-RR	RR	B-RR	B-RR	B-RR	RR
	$searc_{terms}$	generic	generic	generic	all	generic	generic	generic	all
$Influenza_{Poland}$	corr	0.271	0.191	0.355	**0.764**	0.446	0.251	0.300	0.727
	model	LR	LR	LR	B-RR	LR	B-RR	LR	B-RR
	$search_{terms}$	sympt	sympt	name + sympt	all	generic	sympt	all	sympt
$LassaFever_{Nigeria}$	corr	0.338	0.391	0.518	**0.624**	0.355	0.462	0.498	0.613
	model	LR	SVR_{linear}	LR	SVR_{poly}	B-RR	LR	B-RR	SVR_{linear}
	$search_{terms}$	name*	genric	name*	name*	generic	genric	name*	name*
$YellowFever_{Nigeria}$	corr	0.272	0.357	0.469	0.358	0.211	0.457	**0.595**	0.405
	model	SVR_{poly}	RR	SVR_{poly}	SVR_{poly}	SVR_{sig}	SVR_{poly}	B-RR	B-RR
	$search_{terms}$	generic	generic	all	name*	generic	name*	all	generic
$Cholera_{Nigeria}$	corr	0.409	0.409	0.180	0.261	0.288	**0.477**	0.287	0.338
	model	SVR_{poly}	RR	SVR_{Poly}	LR	SVR_{linear}	SVR_{linear}	SVR_{poly}	SVR_{poly}
	$search_{terms}$	generic	generic	generic	generic	generic	generic	generic	generic
$Cholera_{Haiti}$	corr	**0.422**	0.366	0.357	0.361	0.363	0.360	0.359	0.421
	model	SVR_{poly}	SVR_{sig}	SVR_{poly}	SVR_{sig}	SVR_{linear}	SVR_{sig}	SVR_{sig}	B-RR
	$search_{terms}$	sympt	name	name	all	name	all	name	name

B-RR - Bayesian Ridge Regression, LA - Lasso Regression, RR - Ridge Regression;
SVR_{poly}, SVR_{linear}, SVR_{sig} - Support Vector Regression with Polynomial,Linear and sigmoid kernel, respectively.
name: name-based search terms; name* : name-based plus related-name search terms; generic: generic-based search terms;
sympt: symptoms-based search terms; all: using all available search terms.



Influenza-US

Influenza-Japan

Dengue-Brazil

Influenza-Poland

Lassa Fever-Nigeria

Yellow Fever-Nigeria

Cholera-Nigeria

Cholera-Haiti

Figure 5: Performances' distribution by lookback, category and context

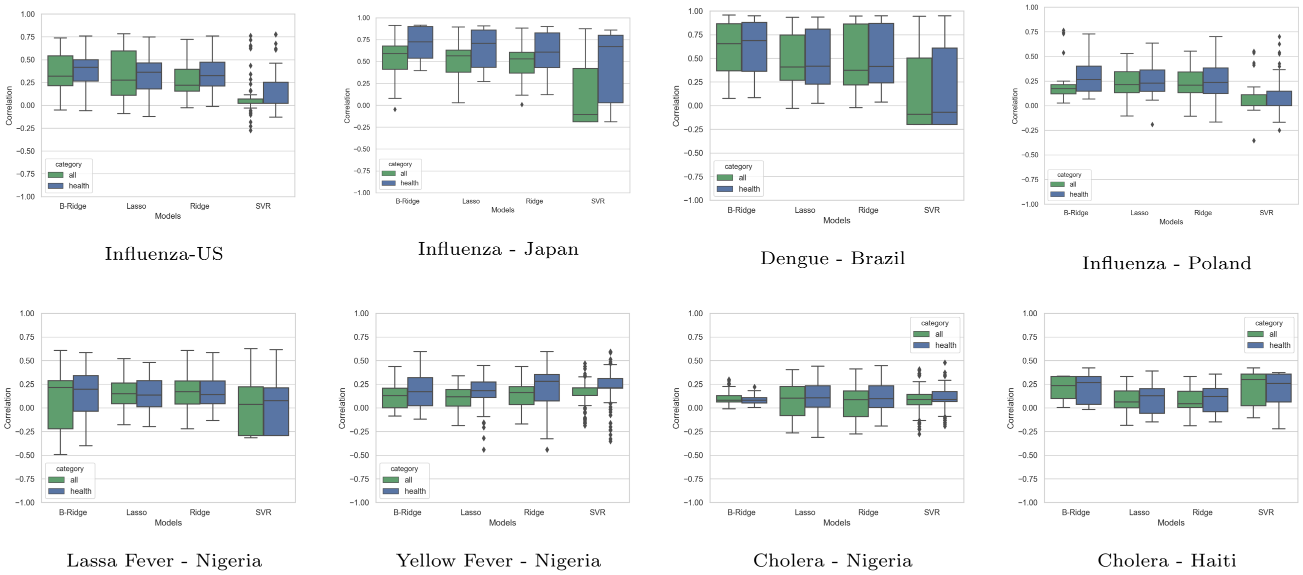
Table 4: Average performance of RNN-based regression models

Context	Neural-based Models		
	GRU-based RNN	*Enc-Dec*	*Att. Enc-Dec*
$Influenza_{US}$	0.201	0.160	0.141
$Influenza_{Japan}$	0.289	0.260	0.210
$Dengue_{Brazil}$	0.301	0.281	0.207
$Influenza_{Poland}$	0.224	0.215	0.198
$LassaFever_{Nigeria}$	0.091	0.042	0.041
$YellowFever_{Nigeria}$	0.090	0.030	0.033
$Cholera_{Nigeria}$	0.072	0.068	0.071
$Cholera_{Haiti}$	0.099	0.089	0.561

urations, the Support Regression Models have lower results compared to other models in most of the contexts. In overall Bayesian Ridge Regression have higher results.

6 2 Lookback Window

One of the main concern raised about using Web data for disease surveillance is that this approach suffers from false warnings caused by noise signals. Not everyone posting content on social media mentioning a disease, or individual searching about an illness on the Internet is ill. Sensationalistic media coverage of an outbreak can lead to an increase in search activity or on volume created on social media, consequently producing a false warning on surveillance systems. Systems built with social media content address this issue through the use of machine learning and Natural Language Processing to filter out irrelevant content. However, current systems built with RSV via Google Trends are still susceptible to this issue because the original searches are not available. Our results on Figure 5 show that in general, longer lookbacks performed better. Fake spikes in search activity caused by panic-induced searches fade away over time, and this is expressed by the improvement in the performance of our models when the lookback increased. Influenza is an epidemic in Japan and with a consistent seasonality, and probably made it less prone to panic-induced searches dur-



Influenza-US

Influenza - Japan

Dengue - Brazil

Influenza - Poland

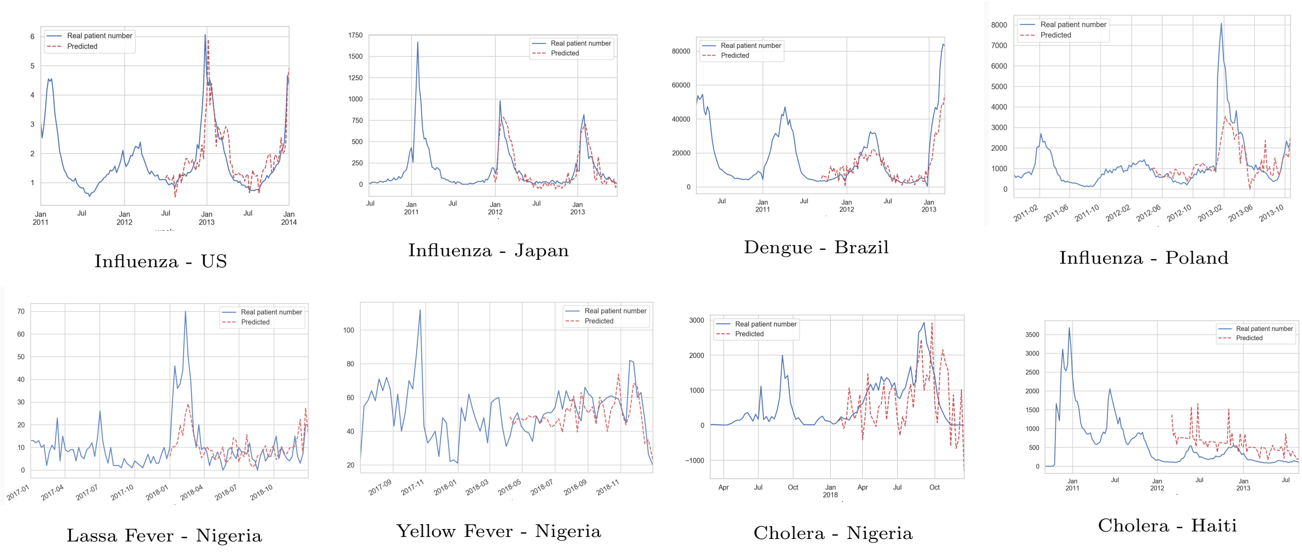
Lassa Fever - Nigeria

Yellow Fever - Nigeria

Cholera - Nigeria

Cholera - Haiti

Figure 7: Performances' distribution by regression model, category and context



Influenza - US

Influenza - Japan

Dengue - Brazil

Influenza - Poland

Lassa Fever - Nigeria

Yellow Fever - Nigeria

Cholera - Nigeria

Cholera - Haiti

Figure 9: Real patient number against predictions by the context's best model (see Table 3)

ing the period of the study which may explain the close performances among all lookbacks on this context but still the longest being the best.

6 3 Category

Figures 5 and 7 show a side-by-side comparison between our models' performances in both categories across lookbacks and regression models, respectively. In Figure 5, comparisons across lookbacks, 32 times out of 40 comparisons (about 80%) the category "health" showed better results. We used two criteria to compare. First, we compared the absolute value of the median of the interquartile range, higher meaning better. Second, when the medians had equal or close absolute values, we considered better the category with a more compact interquartile. Fig. 7 shows much more clear and comparable results, within the same regression model the category "health" performed better than the category "all".

Our results show that category's type have a significant impact on model accuracy and in our task.

7 Conclusion

We proposed sliced timeframe-based search volume approach for disease surveillance. Our results showed that our models produced predictions with correlations against official patient numbers, ranging from 70% to 95% in countries with a high Web search volume, such as US, Japan, Brazil and Poland. In countries with fewer Web search volume, our models produced correlations of about 62% for Lassa fever and 59% for Yellow fever in Nigeria, which we considered successful mid results. In the contexts Cholera - Nigeria and

Cholera - Haiti, our models yielded promising predictions of 47% and 42%, respectively. We showed that our approach performed adequately with longer lookback windows and on category health using standard regression models.

Acknowledgments

This study was supported in part by JSPS KAKENHI Grant Number JP19K20279.

References

[1] Ali Alessa and Miad Faezipour. Tweet Classification Using Sentiment Analysis Features and TF-IDF Weighting for Improved Flu Trend Detection. In *ML and Data Mining in Pattern Recognition*, Lecture Notes in Computer Science, pp. 174–186, Cham, 2018. Springer International Publishing.

[2] Benjamin M. Althouse, Yih Yng Ng, and Derek A. T. Cummings. Prediction of Dengue Incidence Using Search Query Surveillance. *PLOS Neglected Tropical Diseases*, Vol. 5, No. 8, p. e1258, August 2011.

[3] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 1568–1576, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[4] Benjamin N. Breyer, Saunak Sen, David S. Aaronson, Marshall L. Stoller, Bradley A. Erickson, and Michael L. Eisenberg. Use of google insights for search to track seasonal and geographic kidney stone incidence in the united states. Vol. 78, No. 2, pp. 267–271, 08 2011.

[5] Herman Anthony Carneiro and Eleftherios Mylonakis. Google trends: A web-based tool for real-time surveillance of disease outbreaks. Vol. 49, No. 10, pp. 1557–1564, 2009.

[6] Emily H. Chan, Vikram Sahai, Corrie Conrad, and John S. Brownstein. Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. Vol. 5, No. 5, p. e1206, 2011.

[7] Liangzhe Chen, K. S. M. Tozammel Hossain, Patrick Butler, Naren Ramakrishnan, and B. Aditya Prakash. Syndromic surveillance of Flu on Twitter using weakly supervised temporal topic models. *Data Mining and Knowledge Discovery*, Vol. 30, No. 3, pp. 681–710, May 2016.

[8] Cynthia Chew and Gunther Eysenbach. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1n1 Outbreak. *PLOS ONE*, Vol. 5, No. 11, p. e14118, November 2010.

[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *arXiv preprint arXiv:1412.3555*, 2014.

[11] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Gated feedback recurrent neural networks. In *International Conference on Machine Learning*, pp. 2067–2075, 2015.

[12] Reginald DesRoches, Mary Comerio, Marc Eberhard, Walter Mooney, and Glenn J. Rix. Overview of the 2010 Haiti Earthquake. *Earthquake Spectra*, Vol. 27, No. S1, pp. S1–S21, October 2011.

[13] Gunther Eysenbach. Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance. *AMIA Annual Symposium Proceedings*, Vol. 2006, pp. 244–248, 2006.

[14] Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, and Reid Priedhorsky. Global Disease Monitoring and Forecasting with Wikipedia. *PLOS Computational Biology*, Vol. 10, No. 11, p. e1003892, November 2014.

[15] Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer series in statistics. Springer, 2nd ed edition, 2017.

[16] Min Kang, Haojie Zhong, Jianfeng He, Shannon Rutherford, and Fen Yang. Using Google Trends for Influenza Surveillance in South China. *PLOS ONE*, Vol. 8, No. 1, p. e55205, January 2013.

[17] Alex Lamb, Michael J. Paul, and Mark Dredze. Separating fact from fear: Tracking flu infections on twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 789–795, 2013.

[18] V. Lampos and N. Cristianini. Tracking the flu pandemic by monitoring the social web. In *2010 2nd International Workshop on Cognitive Information Processing*, pp. 411–416, June 2010.

[19] David J. McIver and John S. Brownstein. Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. *PLOS Computational Biology*, Vol. 10, No. 4, p. e1003581, April 2014.

[20] Gabriel J. Milinovich, Simon M. R. Avril, Archie C. A. Clements, John S. Brownstein, Shilu Tong, and Wenbiao Hu. Using internet search queries for infectious disease surveillance: screening diseases for suitability. Vol. 14, No. 1, p. 690, December 2014.

[21] NCDC. Nigeria Centre for Disease Control, 2019.

[22] Robert Nisbet, Gary Miner, Ken Yale, John F. Elder, and Andrew F. Peterson. *Handbook of statistical analysis and data mining applications.* Academic Press, second edition edition, 2018. OCLC: on1012917853.

[23] Sudhakar V. Nuti, Brian Wayda, Isuru Ranasinghe, Sisi Wang, Rachel P. Dreyer, Serene I. Chen, and Karthik Murugiah. The use of google trends in health care research: A systematic review. Vol. 9, No. 10, p. e109583, 2014.

[24] Sophal Oum, Daniel Chandramohan, and Sandy Cairncross. Community-based surveillance: a pilot study from rural cambodia. *Tropical medicine & international health*, Vol. 10, No. 7, pp. 689–697, 2005.

[25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[26] Michael J. Paul and Mark Dredze. *Social Monitoring for Public Health.* No. 1947-945X. Morgan & CLAYPOOL, North Carolina, USA, 2017.

[27] Michael J. Paul, Mark Dredze, and David Broniatowski. Twitter Improves Influenza Forecasting. *PLoS Currents*, Vol. 6, , October 2014.

[28] Camille Pelat, Clément Turbelin, Avner Bar-Hen, Antoine Flahault, and Alain-Jacques Valleron. More diseases tracked by using google trends. Vol. 15, No. 8, pp. 1327–1328, 2009.

[29] Reid Priedhorsky, Dave Osthus, Ashlynn R. Daughton, Kelly R. Moran, Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, and Sara Y. Del Valle. Measuring Global Disease with Wikipedia: Success, Failure, and a Research Agenda. In *Proc. of the '17 ACM Conf. on CSCW and Social Computing*, CSCW '17, pp. 1812–1834, New York, NY, USA, 2017. ACM.

[30] Christophe Olivier Schneble, Bernice Simone Elger, and David Shaw. The Cambridge Analytica affair and Internet-mediated research. *EMBO reports*, Vol. 19, No. 8, p. e46579, August 2018.

[31] Mike Schroepfer. An Update on Our Plans to Restrict Data Access on Facebook, April 2018.

[32] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. Vol. 14, No. 3, pp. 199–222, 2004.

[33] I Sutskever, O Vinyals, and QV Le. Sequence to sequence learning with neural networks. *Advances in NIPS*, 2014.

[34] Brian P. Walcott, Brian V. Nahed, Kristopher T. Kahle, Navid Redjal, and Jean-Valery Coumans. Determination of geographic variance in stroke prevalence using internet search engine analytics. Vol. 30, No. 6, p. E19, 2011.

[35] World Health Organization. *World Health Statistics 2018: monitoring health for the SDGs : sustainable development goals.* 2018. OCLC: 1043859485.