

Predicting Vaccination Uptake using Web Search Queries

Niels Dalum Hansen
University of Copenhagen
nhansen@di.ku.dk

Christina Lioma
University of Copenhagen
c.lioma@di.ku.dk

Kåre Mølbak
Statens Serum Institut
KRM@ssi.dk

ABSTRACT

We present a method that uses ensemble learning to combine clinical and web-mined time-series data to make predictions about future vaccination uptake. As clinical data we use official vaccination registries, and as web data we use query frequencies collected from Google Trends. Experiments with all official children vaccines in Denmark show that combining clinical and web data leads to more accurate predictions than using either data types alone. Applying ensemble learning (with default settings) to this task results in predictions of overall low error rate (as low as 4.7 Root Mean Squared Error), which shows the potential for significant improvements with optimising settings. We further show that predictions based only on web data are close to predictions combining clinical and web data. This indicates that web data can be a complementary data source of vaccination uptake in countries that do not have a national vaccination registry.

1. INTRODUCTION AND RELATED WORK

Predicting public health events, such as how many people may get vaccinated in the near future, is an area of research interest. Computationally, predicting an event given large amounts of heterogeneous time-series data is a challenge in itself. Medically, being able to make such predictions can reduce the reaction time of public health professionals, resulting in more efficient services and improved public health. Traditionally, public health event prediction relied on *clinical data* (e.g. microbiological results, medication, or patient registries) that was collected from designated bodies. In the last decade however, non-clinical *web data* (e.g. search engine queries or microblog messages), has been shown useful to the task of predicting public health events. Clinical and web data are complementary sources of evidence: Whereas clinical data contributes expert and curated information to the prediction, web data contributes near real-time information on a large scale about e.g. symptoms or health concerns that may go undetected or unreported by the official clinical

channels.

We present a method for predicting vaccination uptake by combining clinical and web data using ensemble learning. So far, research on vaccination uptake has focused on the effect of physician recommendations on vaccination uptake [4]; how combined sources of information (e.g. physician, television, friends) influence people's decisions about vaccination [5]; the effects of media coverage on vaccination uptake with respect to influenza vaccination [11], HPV vaccination [8], and MMR vaccination [12, 18]. To our knowledge, our study is the first to predict vaccination uptake using web data (with and without clinical data). Predicting vaccination uptake based only on web data is interesting in itself, as some countries do not have vaccination registries covering the whole population, and web data could in those places constitute a timely counterpart to the survey-based reports of vaccination uptake.

Web and/or clinical data have been used before for other types of health event predictions, e.g. influenza activity [6, 13, 14, 15, 17, 20], dengue fever [1] and cholera [3]. How the different types of data used in the prediction should be handled has evolved from using a unified model for both web and clinical data [9, 19], to using ensemble methods that model separately clinical and web data and then combine the outputs [16]. When the type of web data used for prediction is query frequencies [16, 17, 19], a single linear model is used to combine the query frequencies into a prediction. Methods using query frequencies select queries either by (i) timely correlation between query search frequency and the health event [6, 16, 17, 19], or by (ii) expert selection of queries [1, 15, 20]. Both approaches have disadvantages. Approach (i) relies on calculating the correlation between the health event time-series and all queries, which is computationally expensive. It also assumes that historic correlation equals predictive power in the future, which may not always be the case. Approach (ii) relies on human experts, which is costly and does not scale well. In this work we propose a third approach: We select queries based on web descriptions of the health event, in our case of the vaccine in question, using bagging and weighted majority to improve prediction quality compared to previous methods (explained in Section 2).

2. ENSEMBLE LEARNING PREDICTION

Vaccination uptake prediction with time-series data can be formulated as: $\hat{E}(t) \approx E(t-1)$, where $\hat{E}(t)$ is the predicted vaccination uptake at time t , and $E(t-1)$ is the observed vaccination uptake at time $t-1$. We compute

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '16 July 17 - 21 2016, Pisa, Italy

© 2016 ACM. ISBN 978-1-4503-2138-9.

DOI: 10.1145/1235

NOT FOR PUBLIC RELEASE

$\hat{E}(t)$ using ensemble learning by combining separate predictions on vaccination uptake based on clinical and web data into one prediction. Ensemble learning combines predictions from an ensemble of level-0 models into one prediction using a level-1 meta model. We use an ensemble method called *stacking*. First, all level-0 models are trained. Then, a level-1 model is trained to make a final prediction using all the predictions of the level-0 models as input. We experiment with three different types of level-1 models: a linear model, support vector regression (SVR) with a linear kernel, and SVR with a Gaussian kernel. Both our clinical and web data are time-series, i.e. each data point has a temporal reference.

2.1 Level-1 models

Stacking with linear model. We define a linear model with two explanatory variables: $\hat{E}(t) = \mu + \beta_1 \hat{E}_c(t) + \beta_2 \hat{E}_w(t)$, where $\hat{E}_c(t)$ is the prediction based on clinical data at time t , $\hat{E}_w(t)$ is the prediction based on web data at time t , and μ , β_1 and β_2 denote the coefficients that need to be optimized. We use ordinary least squares to find the coefficients that minimize: $\min_{\mu, \beta_1, \beta_2} \sum_t (E(t) - \mu - \beta_1 \hat{E}_c(t) - \beta_2 \hat{E}_w(t))^2$.

Stacking with SVR. SVR solves the same problem as the linear model presented above, but with the possibility of using kernels to transform the input into another feature space. In addition μ , β_1 and β_2 are selected to minimize the following: $\min_{\mu, \beta_1, \beta_2} \sum_t V(E(t) - \mu - \beta_1 \hat{E}_c(t) - \beta_2 \hat{E}_w(t)) + \frac{\lambda}{2}(\mu^2 + \beta_1^2 + \beta_2^2)$, where λ is a hyperparameter controlling the penalty for large coefficients, and $V(r)$ is defined as 0 if $|r| < \epsilon$ and otherwise $|r| - \epsilon$. The parameter ϵ controls how precise the prediction has to be before it is treated as correct.

We experiment with an SVR with linear kernel and with a Gaussian kernel defined as: $K(x, x') = \exp(-\gamma \|x - x'\|^2)$, where γ is a hyperparameter.

2.2 Level-0 models

Prediction with clinical data. As level-0 models we use three well-known time-series methods: autoregressive (AR) models [19, 9], ARIMA and Holt Winters (HW).

AR models estimate $\hat{E}(t)$ as: $\hat{E}(t) = \mu + \sum_{i=1}^m \beta_i E(t-i)$ where m is the number of autoregressive terms, μ is the intercept, the β s control the weight that each past observation has on the prediction. AR models assume that future values of E can be predicted by a linear combination of the m most recently observed values of E . With enough autoregressive terms AR models can handle seasonal changes, but not general upwards or downwards trends.

An extension of the AR models are the ARIMA (AutoRegressive Integrated Moving Average) models. In addition to the autoregressive terms these models also include a moving average, which is a weighted sum of the q most recent forecasting errors. Let m denote the number of autoregressive terms and q the number of moving averages; then: $\hat{E}(t) = \mu + \sum_{i=1}^m \beta_i E(t-i) + \sum_{j=1}^q \phi_j \epsilon_{t-j} + \epsilon_t$, where $\epsilon_t = E(t) - \hat{E}(t)$. To handle trend, the original signal E can be differentiated one or more times [2].

HW forecasting is defined by three recursive equations controlling: level, trend and seasonality. HW can forecast time-series with both trend and seasonal changes. Each equation is defined as a weighted sum in which the weight

of historic observations decreases exponentially with time. HW forecasting with level, trend and seasonality is recursively defined as:

$$\begin{aligned} \text{level} \quad & a_t = \alpha(E(t) - s_{t-l}) + (1 - \alpha)(a_{t-1} + b_{t-1}) \\ \text{trend} \quad & b_t = \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1} \\ \text{seasonality} \quad & s_t = \gamma(E(t) - a_t) + (1 - \gamma)s_{t-l} \end{aligned} \tag{1}$$

where l is the length of the season and α , β and γ are the smoothing parameters which control the influence of the historic level, trend and seasonality. Predictions are made by combining level, trend and seasonality: $\hat{E}(t) = a_{t-1} + b_{t-1} + s_{t-l+1}$.

Prediction with web data. As level-0 models we use a linear model, bagging and weighted majority. Our web data consists of time-stamped query frequencies (described in Section 3).

Linear model. Given a collection of n query frequency time-series, denoted Q , we define a simple linear model as: $\hat{E}(t) = \mu + \sum_{i=1}^n \alpha_i Q_i(t)$, where μ and α are coefficients to be estimated. Such a model can be fitted using any of several methods, the most common being ordinary least squares. Another approach is to use LASSO regularization which is commonly used for making predictions using query frequencies [16, 17, 19]. This approach adds an additional constraint to the optimization, namely that sum of the coefficients should also be minimized. The weight of this sum is controlled by the hyperparameter λ . This approach can be used to avoid overfitting and to reduce the coefficients of non-informative features to zero and thereby induce a sparse model. This is a useful property in this context because the collection of queries might contain several non-informative terms.

Bagging. With bagging, we consider the average of the predictions made on subsets of the training data. This helps to reduce variance and overfitting. We generate subsets of the training data by uniformly sampling with replacement n datasets of size m . For each dataset a linear model, as defined above, is fitted using LASSO regularization, where the parameter λ is found using 3-fold cross-validation. The prediction of the ensemble is the average of the n predictions.

Weighted Majority. We extend the bagging approach to a boosting approach using a weighted majority (WM) algorithm [10]. The WM algorithm works by combining predictions from a collection of models using a weighted average. Each model is associated with its own weight related to its previous predictive performance. If the overall prediction is wrong by a constant ϵ , the weights are updated. The updating works as follows: if the individual prediction of a model has an error larger than ϵ , a new weight is calculated as $w_i = w_i \exp(-\eta)$, where w_i is the weight for model i and η is a hyperparameter controlling the penalty for making wrong predictions. The collection of models we use is identical to the models used for the bagging approach described above.

3. EXPERIMENTAL EVALUATION

Data¹. We evaluate the effectiveness of our approach in predicting vaccination uptake in Denmark for all official

¹We have made all data freely available for research here: www.blinded.for.review.

children vaccines: DiTeKiPol-1, DiTeKiPol-2, DiTeKiPol-3, DiTeKiPol-4, PCV-1, PCV-2, PCV-3, MMR-1, MMR-2(4), MMR-2(12), HPV-1, HPV-2 and HPV-3. We use as clinical data the vaccination uptake recorded by the State Serum Institute, i.e. the total number of vaccines given in a month divided by the number of people expected to be vaccinated that month (based on the size of the monthly birth cohorts published by Statistics Denmark).

We use as web data web search queries that are related to each vaccine. We generate these queries from descriptions of each vaccine in: *www.ssi.dk*, *www.patienthåndbogen.dk*, and *www.min.medicin.dk*. We remove stopwords and collect terms that occur in at least two different descriptions of each vaccine. We treat each term as a query and we submit it to Google Trends using Denmark as the geographical region and with the time period set to January 2011 - September 2015 (only limited coverage of Denmark is available prior to 2011). Only 58 out of 85 queries had enough coverage in Google Trends to return a result. We use these 58 queries.

Training. We use as training data all data which is available prior to the data point being predicted. Hence if we are predicting the vaccination uptake in February 2014 we train on data from January 2011 – January 2014. This means that all models are refitted for each time step. We use monthly time steps. To allow for inference of seasonality, the level-0 models are initialized with 24 months of available data (January 2011 – December 2012) as training data. For the level-1 models we start by using 12 months of data (January 2013 – December 2013).

We evaluate our predictions using the root mean squared error (RMSE), which penalizes large errors more than small.

Our prediction methods are fitted using R packages with default settings at all times, except for the starting point for HW, where we manually select a starting point of the optimization if it cannot be completed with the default value. The AR model is trained using 12 autoregressive terms to capture seasonal variations. For bagging and weighted majority we use as many subsets as there are queries. Each subset contains 10 queries. For the weighted majority we use $\eta = 5$ and $\epsilon = 2$ for all experiments.

Results. Table 1 shows the results when predicting vaccination uptake using either clinical or web data only (with the methods presented in Section 2). “Naive” refers to our naive baseline $\hat{E}(t) = E(t - 1)$. Our methods outperform the naive baseline except for the HPV vaccines. This might be due to an intense debate in Denmark regarding the safety of the vaccine. We see that methods using clinical data outperform the methods using web data for the majority of the vaccines. But interestingly this difference is not very big and for the vaccines DiTeKiPol-3 and DiTeKiPol-4 the methods based on web data perform best. DiTeKiPol-4 is especially interesting since a shortage in 2013 resulted in unusual vaccination behaviour for a few months. When making predictions from web data our two new approaches (bagging and WM) perform best for 9 of the 13 vaccines.

Table 2 shows the results for the ensemble predictions using clinical and web data. Except for the three HPV vaccines the ensemble approaches outperform any of the methods using only one data source. We see that when using an SVR with a Gaussian kernel as level-1 model we obtain the best results, i.e. 7/13 lowest RMSE. When comparing within the methods using an SVR with a Gaussian kernel the HW+WM is the best performing method. The most im-

	Naive	Clinical data			Web data			
		HW	AR12	ARIMA	WM	B	L	O
MMR-1	20.704	18.149	18.606	15.574	16.609	16.597	16.605	30.387
MMR-2 (4)	20.582	13.110	16.566	16.284	15.841	15.635	15.500	29.288
MMR-2 (12)	20.637	19.592	20.600	18.726	21.631	20.815	21.112	31.897
HPV-1	8.080	11.291	11.192	9.871	13.474	14.320	12.701	11.547
HPV-2	8.704	12.522	12.806	11.276	18.154	18.025	18.423	15.404
HPV-3	6.579	9.161	13.958	9.418	24.239	23.494	23.074	17.317
DiTeKiPol-1	14.091	6.700	5.185	5.097	8.067	8.058	8.069	15.913
DiTeKiPol-2	17.693	7.520	8.030	8.064	10.003	9.941	9.951	20.082
DiTeKiPol-3	17.884	17.596	20.936	19.459	17.160	17.160	17.158	30.424
DiTeKiPol-4	21.676	26.103	21.676	23.385	15.414	15.535	15.934	33.888
PCV-1	13.323	6.897	6.394	6.623	7.745	7.797	7.845	14.014
PCV-2	17.533	7.266	8.845	8.353	9.679	9.796	9.770	16.027
PCV-3	18.405	7.877	7.781	7.634	10.410	10.364	10.368	15.582

Table 1: RMSE of predictions with only clinical or web data. WM: weighed majority, B: Bagging, L: linear model w. LASSO and O: linear model w. OLS. Blue: lowest RMSE per vaccine. Bold: better than naive.

provements are obtained when combining predictions based on web data with either predictions from HW or AR12.

4. CONCLUSIONS

We presented a method that uses ensemble learning to combine clinical and web-mined time-series data to make predictions about future vaccination uptake (how many people are likely to get vaccinated at some point in the near future). As clinical data we used official registries of vaccines in Denmark. As web data we used query frequencies collected from Google Trends. We created those queries by extracting terms from publicly available descriptions of the vaccines on the web. Experiments using all officially recommended children vaccines in Denmark for the period January 2011 – September 2015 showed that for 10/13 vaccines our ensemble learning methods that combined clinical with web data for prediction outperformed predictions using either clinical or web data alone. Though the combination yields the lowest overall error, it is worth noticing that when only using web data we are able to get predictions with an error only slightly worse than for the predictions made using only clinical data. This indicates the potential usefulness of web data, such as query frequencies, to predict vaccination uptake in countries where there is no national vaccination registry.

References

- [1] E. H. Chan, V. Sahai, C. Conrad, and J. S. Brownstein. Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLoS Negl Trop Dis*, 5(5):e1206, 2011.
- [2] C. Chatfield. *The analysis of time series: An introduction*. CRC press, 2013.
- [3] R. Chunara, J. R. Andrews, and J. S. Brownstein. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene*, 86(1):39–45, 2012.
- [4] L. M. Gargano, N. L. Herbert, J. E. Painter, J. M. Sales, C. Morfaw, K. Rask, D. Murray, R. DiClemente, and J. M. Hughes. Impact of a physician recommendation and parental immunization attitudes on receipt or intention to receive adolescent vaccines. *Human vaccines & immunotherapeutics*, 9(12):2627–2633, 2013.
- [5] L. M. Gargano, N. L. Underwood, J. M. Sales, K. Seib, C. Morfaw, D. Murray, R. J. DiClemente, and J. M. Hughes. Influence of sources of information about influenza vaccine on parental attitudes and adolescent vaccine receipt. *Human vaccines & immunotherapeutics*, 2015.

OLS													
	HW+WM	HW+B	HW+L	HW+O	AR12+WM	AR12+B	AR12+L	AR12+O	ARIMA+WM	ARIMA+B	ARIMA+L	ARIMA+O	
MMR-1	15.190	15.476	15.187	16.842	17.697	17.572	17.457	18.151	16.296	16.492	15.968	17.877	
MMR-2 (4)	12.875	13.497	13.349	13.121	16.305	16.108	16.195	16.032	18.872	16.220	21.085	15.871	
MMR-2 (12)	18.082	17.650	17.541	17.221	18.711	19.523	18.762	18.909	18.469	19.409	18.961	18.693	
HPV-1	10.552	10.435	10.960	11.516	9.377	9.690	10.130	10.281	10.348	10.080	9.992	10.080	
HPV-2	12.743	12.923	14.384	12.191	11.883	12.201	13.240	10.503	10.708	10.655	10.279	9.220	
HPV-3	8.743	8.771	10.231	9.987	11.321	11.063	12.110	12.151	9.893	9.237	9.818	9.918	
DiTeKiPol-1	6.416	6.875	6.477	5.498	4.835	4.831	4.829	5.082	6.072	5.690	5.625	5.584	
DiTeKiPol-2	9.094	8.216	8.967	7.956	7.686	7.343	8.116	8.019	9.461	8.989	15.485	9.057	
DiTeKiPol-3	18.478	17.891	18.410	16.662	17.529	18.225	17.550	18.439	17.168	17.227	17.137	19.076	
DiTeKiPol-4	15.812	17.977	17.495	19.860	17.290	19.891	16.537	19.849	24.391	45.079	36.403	24.220	
PCV-1	7.042	5.783	5.716	5.391	6.201	5.950	5.830	5.973	10.785	6.569	9.174	6.169	
PCV-2	8.317	8.236	9.283	7.553	10.135	8.670	10.401	8.284	8.395	8.399	9.681	8.330	
PCV-3	7.345	7.436	8.199	7.759	6.825	7.014	7.108	6.736	7.931	8.364	8.240	7.670	
SVR linear													
	HW+WM	HW+B	HW+L	HW+O	AR12+WM	AR12+B	AR12+L	AR12+O	ARIMA+WM	ARIMA+B	ARIMA+L	ARIMA+O	
MMR-1	16.541	15.969	15.478	16.905	17.298	17.252	17.399	18.496	19.406	16.793	16.857	18.204	
MMR-2 (4)	12.648	12.981	12.388	12.952	16.148	15.373	16.663	15.095	14.969	16.101	15.419	15.620	
MMR-2 (12)	17.906	18.054	17.872	17.374	19.302	19.020	18.248	19.075	19.123	18.193	19.195	17.731	
HPV-1	10.530	10.649	10.922	11.146	10.486	10.494	10.688	10.657	10.331	10.789	10.810	10.448	
HPV-2	13.489	12.344	12.130	12.425	10.147	10.467	11.984	11.062	9.738	9.906	10.482	8.727	
HPV-3	8.903	8.260	8.501	10.674	11.732	11.718	13.049	12.617	9.806	10.001	9.484	10.411	
DiTeKiPol-1	6.926	5.993	5.739	6.500	4.740	4.758	4.626	5.077	5.090	5.354	5.287	5.507	
DiTeKiPol-2	9.808	9.476	9.006	9.100	8.476	8.563	10.503	7.734	9.938	8.480	9.872	9.591	
DiTeKiPol-3	22.546	22.599	22.546	17.433	21.402	21.925	21.154	19.003	22.244	21.319	22.104	19.568	
DiTeKiPol-4	16.694	37.909	17.357	14.461	15.922	16.405	16.124	16.164	22.072	26.591	18.984	17.609	
PCV-1	7.351	7.186	6.175	6.198	5.282	6.143	6.335	5.510	6.765	7.413	6.840	6.830	
PCV-2	7.689	7.946	15.613	7.955	9.029	8.879	9.104	8.823	8.794	11.662	14.559	9.024	
PCV-3	7.648	8.261	8.388	7.784	6.904	6.758	6.994	6.633	9.649	9.384	9.058	8.491	
SVR Gaussian													
	HW+WM	HW+B	HW+L	HW+O	AR12+WM	AR12+B	AR12+L	AR12+O	ARIMA+WM	ARIMA+B	ARIMA+L	ARIMA+O	
MMR-1	14.928	16.694	16.355	16.198	17.770	18.207	18.117	16.927	16.703	17.490	17.569	17.560	
MMR-2 (4)	14.377	12.870	14.094	13.122	15.709	14.780	15.024	15.468	14.973	15.109	16.625	15.838	
MMR-2 (12)	18.007	17.446	18.972	16.530	17.945	19.115	18.406	18.176	18.385	18.553	19.625	19.041	
HPV-1	10.748	11.606	11.918	11.289	11.002	10.902	11.403	9.130	11.664	10.684	11.505	9.987	
HPV-2	13.513	11.958	12.304	12.097	10.789	12.376	9.964	12.483	10.537	11.249	11.715	10.961	
HPV-3	12.889	13.784	14.775	13.352	13.016	14.521	15.222	14.374	12.818	12.172	12.147	12.682	
DiTeKiPol-1	5.204	5.008	5.098	5.331	5.486	5.467	5.486	6.227	5.725	6.393	5.577	6.615	
DiTeKiPol-2	7.927	8.017	8.172	9.661	7.149	7.443	7.818	8.078	9.009	9.870	9.848	8.721	
DiTeKiPol-3	16.639	16.448	16.433	17.275	18.650	18.442	19.009	18.355	18.380	17.545	18.298	18.962	
DiTeKiPol-4	15.616	14.877	15.246	16.543	16.865	16.038	16.687	15.653	15.938	15.647	15.923	15.932	
PCV-1	5.256	5.808	5.769	5.664	5.450	6.358	6.363	6.614	6.724	7.160	6.611	7.091	
PCV-2	6.463	7.665	7.366	7.470	8.811	7.450	6.952	9.026	8.672	9.062	8.991	9.148	
PCV-3	7.121	7.665	8.396	7.798	9.022	9.556	10.008	7.871	8.616	8.148	9.527	8.176	

Table 2: RMSE of ensemble predictions (clinical and web data). Blue: lowest RMSE per vaccine. Bold: lower RMSE than for the individual ensemble components in Table 1.

- [6] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.
- [7] R. J. Hyndman and Y. Khandakar. Automatic time series for forecasting: The forecast package for R. Technical report, Monash University, Department of Econometrics and Business Statistics, 2007.
- [8] B. J. Kelly, A. E. Leader, D. J. Mittermaier, R. C. Hornik, and J. N. Cappella. The HPV vaccine and the media: How has the topic been covered and what are the effects on knowledge about the virus and cervical cancer? *Patient education and counseling*, 77(2):308–313, 2009.
- [9] D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parasite of Google flu: Traps in big data analysis. *Science*, 343(14 March), 2014.
- [10] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.
- [11] K. Ma, W. Schaffner, C. Colmenares, J. Howser, J. Jones, and K. Poehling. Influenza vaccinations of young children increased with media coverage in 2003. *Pediatrics*, 117(2):e157–e163, 2006.
- [12] B. Mason and P. Donnelly. Impact of a local newspaper campaign on the uptake of the measles mumps and rubella vaccine. *Journal of epidemiology and community health*, 54(6):473, 2000.
- [13] D. J. McIver and J. S. Brownstein. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. *PLoS Comput Biol*, 10(4):e1003581, 2014.
- [14] M. J. Paul, M. Dredze, and D. Broniatowski. Twitter improves influenza forecasting. *PLoS currents*, 6, 2014.
- [15] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448, 2008.
- [16] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput Biol*, 11(10):e1004513, 2015.
- [17] M. Santillana, E. O. Nsoesie, S. R. Mekaru, D. Scales, and J. S. Brownstein. Using clinicians’ search query data to monitor influenza epidemics. *Clinical Infectious Diseases*, 59(10):1446–1450, 2014.
- [18] M. J. Smith, S. S. Ellenberg, L. M. Bell, and D. M. Rubin. Media coverage of the measles-mumps-rubella vaccine and autism controversy and its relationship to MMR immunization rates in the United States. *Pediatrics*, 121(4):e836–e843, 2008.
- [19] S. Yang, M. Santillana, and S. Kou. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478, 2015.
- [20] Q. Yuan, E. O. Nsoesie, B. Lv, G. Peng, R. Chunara, and J. S. Brownstein. Monitoring influenza epidemics in China with search query from Baidu. *PLoS one*, 8(5):e64323, 2013.