Time-Series Adaptive Estimation of Vaccination Uptake Using Web Search Queries

Dalum Hansen, Niels; Mølbak, Kåre; Cox, Ingemar J.; Lioma, Christina

Published in:
Proceedings of the 26th International Conference on World Wide Web Companion

DOI:
10.1145/3041021.3054251

Publication date:
2017

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Time-Series Adaptive Estimation of Vaccination Uptake Using Web Search Queries

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

Kåre Mølbak  
Statens Serum Institut  
KRM@ssi.dk

Ingemar J. Cox  
University of Copenhagen  
ingemar.cox@di.ku.dk

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

ABSTRACT

Estimating vaccination uptake is an integral part of ensuring public health. It was recently shown that vaccination uptake can be estimated automatically from web data, instead of slowly collected clinical records or population surveys [2]. All prior work in this area assumes that features of vaccination uptake collected from the web are temporally regular. We present the first ever method to remove this assumption from vaccination uptake estimation: our method dynamically adapts to temporal fluctuations in time series web data used to estimate vaccination uptake. We show our method to outperform the state of the art compared to competitive baselines that use not only web data but also curated clinical data. This performance improvement is more pronounced for vaccines whose uptake has been irregular due to negative media attention (HPV-1 and HPV-2), problems in vaccine supply (DiTeKiPol), and targeted at children of 12 years old (whose vaccination is more irregular compared to younger children).

1. INTRODUCTION AND RELATED WORK

Vaccination programs are an efficient and cost effective method to improve public health. With sufficiently many people vaccinated the population gains herd immunity, meaning the disease cannot spread. Timely actions to avoid drops in vaccination coverage are therefore of great importance. Many countries have no registries of timely vaccination uptake information, but rely for example on yearly surveys. In such countries estimations of near real-time vaccination uptake based solely on web data are valuable. We extend prior work in this area [2], which showed that vaccination uptake can be estimated sufficiently accurately from web search data. Our extension consists of a new estimation method that adapts dynamically to temporal fluctuations in the signal (web search queries in our case) instead of assuming temporal stationarity as in [2]. This contribution is novel within vaccination uptake estimation.

Linear models have been used previously to estimate health events, for instance by combining data from multiple sources with an ensemble of decision trees [5], or, closer to our work, by using query frequencies for influenza like illness [1] or vaccination uptake estimation [2]. These approaches are designed for stationary time-series analysis, i.e. they assume data is generated by a stationary stochastic process. Our motivation is that vaccination uptake often does not follow stationary seasonal patterns. External events such as disease outbreaks, suspicion of adverse effects, or temporary vaccine shortages can alter uptake patterns for shorter or longer periods of time. Hence, while historical data is a good estimator in stable periods, as shown in [2], we reason that adapting the estimation to any unstability can reduce estimation error. We experimentally confirm this on all official children vaccines data used in Denmark between 2011 - 2016.

2. AGGREGATION WITH REGRESSION TREES FOR TIME SERIES ADAPTATION

To address these challenges, we combine AA with regression trees, motivated by recent research showing that random forests outperform ARIMA models on avian influenza prediction [4]. A random forest, i.e. an ensemble of decision trees, is well suited for our problem since it is easy to extend to multiple data sources.

Our method works as follows: We initially generate a set of regression trees. For each time step the ensemble of regression trees is retrained based on the initial set of trees and a weighted sum is used to make the estimation. AA is used to continuously update the weights of each tree. Each regression tree is trained based on a set of features and training samples. For each tree a feature set is drawn with replacement from the complete feature set. Training samples are selected based on time-relative indices, where index 0 corresponds to the current time step. The indices are uniformly drawn with replacement from the interval [0; s], where s is a window size. We use trees with different window sizes to account for stationarity and non-stationarity of the signal.

Our adaptive vaccination estimation algorithm is shown in Algorithm 1, where $\eta$ is the learning rate, RT a set of $N$ regression trees, $\tau$ the amount of initial training data and $y$ the vaccination uptake.

3. EVALUATION

To facilitate direct comparison, we evaluate our method on the same data as [2]: monthly vaccination uptake of all official
Vaccination uptake is defined as the total number of people vaccinated in a month divided by the birth cohort for that month. To estimate vaccination uptake, we use frequencies of web search queries extracted from Google Trends. We use the exact same frequencies of single term queries provided by [2].

We compare to two baselines: (1) Linear regression with lasso regularization where the hyper-parameter is found using threefold cross-validation on the training data; (2) Linear regression with elastic net regularization where the two hyper-parameters are selected using threefold cross-validation. We also include for reference two upper bounds corresponding to the best score reported in [2] when using (i) only web data, and (ii) web data combined with clinical data. These scores are not theoretical upper bounds, but just the best scores across all methods evaluated in [2]. We treat them as performance upper bounds because they do not correspond to any individual method, but to the best score per vaccine across all methods reported in [2]. Neither baselines or upper bounds account for time-series adaptation, i.e. they all assume data stationarity.

The initial number of training samples, $i$, is set to 24. All algorithms are evaluated in a leave-one-out fashion, where all data prior to the data point being estimated is used for training. For our algorithm (ATSE) a parameter search is performed by randomly sampling from the following intervals: Window size interval 1-46, number of features derived from vaccination data 0-45, number of regression trees 500-10000, $\eta$ between 0.001-0.25.

Table 1 displays the root mean squared error (RMSE) between the estimated vaccination uptake and the real vaccination uptake for all methods. Our method yields the overall best performance compared to the baselines (it outperforms all baselines for 8 out of 12 vaccines). Our method also outperforms the upper bounds of [2] (any of the two) for 6 vaccines. This supports our reasoning that adapting the estimation to temporal fluctuations is a better strategy than assuming data stationarity. Our method yields the strongest performance improvements for HPV-1, HPV-2, MMR-2(12) and DiTeKiPol.

All of these vaccines have temporally irregular uptake patterns, as explained next. HPV-1 and HPV-2 have been subject to a heavy media debate in Denmark, which might have forced people to postpone the initial vaccination, hence introducing irregularities in the signal. For PCV vaccines there have been no noted irregularities in their uptake patterns, which explains the slight drop in performance by our method compared to the baselines.

5. REFERENCES