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

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

Kåre Mølbak
Statens Serum Institut
KRM@ssi.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 data from web searches 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 instability can reduce estimation error. We experimentally confirm this on all official children vaccines data used in Denmark between 2011 - 2016.

2. AGGREGATION WITH REGRESSION

To address these challenges, we combine AA with 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 adapt to changes. In cases where: (i) there is more than one data source, e.g. vaccine uptake data and search frequency data, and (ii) when the time series data to be estimated are assumed to be unavailable (near real-time).

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 divided by the birth cohort for that month. To estimate vaccination uptake, we use 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 three fold cross-validation. We also include a parameterized estimation of RT[n] given \( x_t \)

\begin{align*}
\text{Algorithm 1: Adaptive time series estimation} \\
\text{Require: } RT, \eta, i \\
1: W \leftarrow \text{list with weights, initialize to be uniform} \\
2: X \leftarrow \text{list with the first } i \text{ training samples} \\
3: Y \leftarrow \text{list with the first } i \text{ observations of } y \\
4: \hat{Y} \leftarrow \text{empty list of estimations} \\
5: t \leftarrow \text{current time step, starting at } i+1 \\
6: \text{while True do} \\
7: \quad x_t \leftarrow \text{receive new observation from data stream} \\
8: \quad \text{for } n=0 \text{ to } N \text{ do} \\
9: \quad \quad \text{Train } RT[n] \text{ using } X \text{ and } Y \\
10: \quad \quad \hat{Y}_n[n] \leftarrow \text{estimation of } RT[n] \text{ given } x_t \\
11: \quad \text{end for} \\
12: \quad Y[n] \leftarrow \sum_{n=0}^{N} W[n] \cdot \hat{Y}_n[n] \\
13: \quad \text{observed } y \text{ at time } t \\
14: \quad \text{for } n=0 \text{ to } N \text{ do} \\
15: \quad \quad \hat{Y}[t] \leftarrow W[n] \cdot \exp(-\eta (\hat{Y}_n[n] - Y[t])^2) \\
16: \quad \text{end for} \\
17: \quad W \leftarrow \text{normalize } W \\
18: \quad X[t] \leftarrow x_t \\
19: \quad t \leftarrow t+1 \\
20: \text{end while}
\end{align*}

children vaccines in Denmark from January 2011 - June 2016. 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 single term queries provided by [2]. We compare to two baselines: (1) Linear regression with lasso regularization where the hyper-parameter is found using three fold cross-validation on the training data; (2) Linear regression with elastic net regularization where the two hyper-parameters are also selected using three fold cross-validation. We also include a parameterized estimation of RT[n] given \( x_t \)

\begin{align*}
\text{Algorithm 1: Adaptive time series estimation} \\
\text{Require: } RT, \eta, i \\
1: W \leftarrow \text{list with weights, initialize to be uniform} \\
2: X \leftarrow \text{list with the first } i \text{ training samples} \\
3: Y \leftarrow \text{list with the first } i \text{ observations of } y \\
4: \hat{Y} \leftarrow \text{empty list of estimations} \\
5: t \leftarrow \text{current time step, starting at } i+1 \\
6: \text{while True do} \\
7: \quad x_t \leftarrow \text{receive new observation from data stream} \\
8: \quad \text{for } n=0 \text{ to } N \text{ do} \\
9: \quad \quad \text{Train } RT[n] \text{ using } X \text{ and } Y \\
10: \quad \quad \hat{Y}_n[n] \leftarrow \text{estimation of } RT[n] \text{ given } x_t \\
11: \quad \text{end for} \\
12: \quad Y[n] \leftarrow \sum_{n=0}^{N} W[n] \cdot \hat{Y}_n[n] \\
13: \quad \text{observed } y \text{ at time } t \\
14: \quad \text{for } n=0 \text{ to } N \text{ do} \\
15: \quad \quad \hat{Y}[t] \leftarrow W[n] \cdot \exp(-\eta (\hat{Y}_n[n] - Y[t])^2) \\
16: \quad \text{end for} \\
17: \quad W \leftarrow \text{normalize } W \\
18: \quad X[t] \leftarrow x_t \\
19: \quad t \leftarrow t+1 \\
20: \text{end while}
\end{align*}

Table 1 displays the root mean squared error (RMSE) between the estimated vaccination uptake and the real vaccination uptake for vaccines with known irregularities in their uptake patterns. We have been subject to a heavy media debate in Denmark, with a subsequent drop in vaccinations. MMR-2(12) denotes the second MMR vaccine targeted 12 year-olds. As children grow, parents are less likely to follow the recommended vaccination schedule and fluctuations correlated with measles outbreaks are observed, thus making the time series less stationary. Lastly, in recent years there have been problems obtaining a sufficient supply of certain DiTeKiPol vaccines 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.

### 4. CONCLUSION

We presented an automatic method for near real time estimation of health events using web search query data. Our method combines an Aggregation Algorithm (AA) to automatically reduce estimation error in changing environments with regression trees. We applied our method to estimate vaccination uptake in all official Danish children vaccines, following [2], and showed that our approach overall outperformed strong baselines that assumed data to be temporally regular. Our method was particularly strong estimating uptake for vaccines with known irregularities in their usage, such as HPV-1, HPV-2, MMR-2(12) and DiTeKiPol.

This work confirms recent findings that vaccination uptake can be automatically estimated only from web data, and further extends this area by accounting for irregular uptake patterns.

Table 1: Estimation error when estimating vaccination uptake from web search queries with our method (ATSE), Lasso (LASS), Elastic Net (EN), and the two performance upper bounds of [2] with web search (UBW) and web search and clinical data (UBWC). Bold marks best (excluding upper bounds). Asterisk marks better or equal to any upper bound.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HPV-1</td>
<td>14.6</td>
<td>13.8</td>
<td>10.0*</td>
<td>11.5</td>
<td>9.3</td>
</tr>
<tr>
<td>HPV-2</td>
<td>15.9</td>
<td>16.1</td>
<td>10.1</td>
<td>15.4</td>
<td>8.7</td>
</tr>
<tr>
<td>MMR-1</td>
<td>12.9</td>
<td>12.9</td>
<td>12.0*</td>
<td>16.5</td>
<td>14.9</td>
</tr>
<tr>
<td>MMR-2(4)</td>
<td>15.5</td>
<td>14.7</td>
<td>14.2</td>
<td>12.4</td>
<td>12.3</td>
</tr>
<tr>
<td>MMR-2(12)</td>
<td>21.7</td>
<td>21.4</td>
<td>16.0*</td>
<td>20.8</td>
<td>16.5</td>
</tr>
<tr>
<td>DiTeKiPol-1</td>
<td>16.2</td>
<td>16.2</td>
<td>10.8</td>
<td>8.0</td>
<td>4.6</td>
</tr>
<tr>
<td>DiTeKiPol-2</td>
<td>14.1</td>
<td>14.2</td>
<td>12.4</td>
<td>9.9</td>
<td>7.1</td>
</tr>
<tr>
<td>DiTeKiPol-3</td>
<td>10.8</td>
<td>11.1</td>
<td>10.0*</td>
<td>17.1</td>
<td>16.4</td>
</tr>
<tr>
<td>DiTeKiPol-4</td>
<td>13.7*</td>
<td>14.3*</td>
<td>14.4*</td>
<td>15.4</td>
<td>14.4</td>
</tr>
<tr>
<td>PCV-1</td>
<td>7.5</td>
<td>7.8</td>
<td>10.0</td>
<td>7.7</td>
<td>5.2</td>
</tr>
<tr>
<td>PCV-2</td>
<td>9.6*</td>
<td>9.5</td>
<td>10.0</td>
<td>9.6</td>
<td>6.4</td>
</tr>
<tr>
<td>PCV-3</td>
<td>9.4*</td>
<td>9.5</td>
<td>10.1*</td>
<td>10.3</td>
<td>6.6</td>
</tr>
</tbody>
</table>

Funded by IBM and the Danish Agency for Science & Higher Education.

### 5. REFERENCES


