Estimating the population impact of a new pediatric influenza vaccination program in England using social media content

Wagner, Moritz; Lampos, Vasileios; Yom-Tov, Elad; Pebody, Richard; Cox, Ingemar Johansson

Published in:
Journal of Medical Internet Research

DOI:
10.2196/jmir.8184

Publication date:
2017

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

Citation for published version (APA):
Wagner, M., Lampos, V., Yom-Tov, E., Pebody, R., & Cox, I. J. (2017). Estimating the population impact of a new pediatric influenza vaccination program in England using social media content. *Journal of Medical Internet Research*, *19*(12), [e416]. https://doi.org/10.2196/jmir.8184

Original Paper

Estimating the Population Impact of a New Pediatric Influenza Vaccination Program in England Using Social Media Content

Moritz Wagner[1,2,3*], MMath; Vasileios Lampos[4*], MSc, PhD; Elad Yom-Tov[5], BSc, MSc, PhD; Richard Pebody[1], MBChB, PhD; Ingemar J Cox[4,6], BSc, PhD

[1]Public Health England, London, United Kingdom

[2]University College London, London, United Kingdom

[3]London School of Hygiene and Tropical Medicine, London, United Kingdom

[4]Department of Computer Science, University College London, London, United Kingdom

[5]Microsoft Research, Herzliya, Israel

[6]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

[*]these authors contributed equally

Corresponding Author:
Moritz Wagner, MMath
Public Health England
61 Colindale Ave
London, NW9 5EQ
United Kingdom
Phone: 44 7539078912
Email: moritz.wagner.16@ucl.ac.uk

Abstract

Background: The rollout of a new childhood live attenuated influenza vaccine program was launched in England in 2013, which consisted of a national campaign for all 2 and 3 year olds and several pilot locations offering the vaccine to primary school-age children (4-11 years of age) during the influenza season. The 2014/2015 influenza season saw the national program extended to include additional pilot regions, some of which offered the vaccine to secondary school children (11-13 years of age) as well.

Objective: We utilized social media content to obtain a complementary assessment of the population impact of the programs that were launched in England during the 2013/2014 and 2014/2015 flu seasons. The overall community-wide impact on transmission in pilot areas was estimated for the different age groups that were targeted for vaccination.

Methods: A previously developed statistical framework was applied, which consisted of a nonlinear regression model that was trained to infer influenza-like illness (ILI) rates from Twitter posts originating in pilot (school-age vaccinated) and control (unvaccinated) areas. The control areas were then used to estimate ILI rates in pilot areas, had the intervention not taken place. These predictions were compared with their corresponding Twitter-based ILI estimates.

Results: Results suggest a reduction in ILI rates of 14% (1-25%) and 17% (2-30%) across all ages in only the primary school-age vaccine pilot areas during the 2013/2014 and 2014/2015 influenza seasons, respectively. No significant impact was observed in areas where two age cohorts of secondary school children were vaccinated.

Conclusions: These findings corroborate independent assessments from traditional surveillance data, thereby supporting the ongoing rollout of the program to primary school-age children and providing evidence of the value of social media content as an additional syndromic surveillance tool.

(J Med Internet Res 2017;19(12):e416) doi:10.2196/jmir.8184

KEYWORDS

health intervention; influenza; vaccination; social media; Twitter

Introduction

Background

In 2012 the Joint Committee on Vaccination and Immunisation recommended the extension of the annual influenza vaccination campaign to include all healthy children aged 2 to 16 years of age in England [1]. This decision was informed by influenza transmission modeling done using an evidence-synthesis approach, showing that vaccination could not only protect the children themselves from infection, but also decrease influenza transmission in the general population. This finding included the indirect protection of at-risk groups, such as people over 65 years of age or those with underlying clinical risk factors [2]. The phased rollout of the live attenuated influenza vaccine (LAIV) program began during the 2013/2014 influenza season. In the first season, the program offered vaccinations to all 2 and 3-year-olds throughout England. A number of geographically distinct pilot regions also offered vaccinations to primary school age children (4-11 years of age) to determine the optimal model of delivery to school-age children. For the 2014/2015 influenza season, the program was extended nationally to offer vaccinations to all 2 to 4-year-olds. Pilot locations were added that offered vaccinations to children either (1) of primary school age (*Primary school*; 4-11 years), (2) the first two years of secondary school age (*Secondary school*, 11-13 years), or (3) both (*Primary and Secondary school*; 4-13 years) to determine optimal models of delivery.

Motivation

Public Health England (PHE) has been using a variety of surveillance systems to assess the overall population impact of the childhood influenza campaign in children of school-age on influenza epidemiology to validate the direct and indirect effects of vaccinating this age group. The pilot locations for 2014/2015 are of particular interest, as the variation in target groups may offer further insights into the optimal strategies for the national rollout. During the 2014/2015 campaign, most influenza indicators through traditional surveillance systems in both targeted and nontargeted age groups demonstrated a significant reduction in pilot areas that offered the vaccine to primary school age children. However, there was little impact in pilot areas, where only two age cohorts of secondary school age children were vaccinated [3]. These surveillance indicators were based on health systems ranging from General Practitioners' consultation rates to excess mortality.

Whilst such results are important in estimating the intervention's effects on health care services, online user-generated information offers a complementary data source that can provide additional insights into the impact of such campaigns on the wider community, including those persons that do not consult the health care system. Our study also highlights the potential value of user-generated information in the absence of routine evaluation systems. Internet-based surveillance systems are being viewed as novel logistically and economically viable developments that offer great potential as an extension of traditional surveillance systems [4]. Recent research efforts have shown that in combination with *machine learning* techniques, data from social media or search engines can be used to accurately estimate disease-related indicators such as influenza-like illness (ILI) rates [5-9]. These technologies provide health monitoring systems with additional, publicly available, and potentially more timely sources of data for syndromic surveillance. Furthermore, compared to traditional surveillance systems, user-generated content may offer insights about a wider range of the population, including the bottom part of the disease population pyramid (ie, those that do not seek medical attention) [10].

For the 2013/2014 pilot areas, in order to provide further evidence of the community-wide effects of vaccinating children with influenza vaccine, Lampos et al made use of online user-generated content in combination with statistical natural language processing techniques to estimate ILI rates in the population [9]. By matching nonvaccinated control areas with pilot areas and using flu-related Twitter posts or Bing search queries from these locations, the impact of the campaign within the *Primary school* age pilot areas was estimated, showing a significant decrease (22% to 33% reduction) in influenza transmission in the general population in these pilot areas compared to corresponding control areas [9]. PHE's estimates also showed evidence of a reduction in influenza transmission in targeted and nontargeted age groups in pilot areas compared to nonpilot areas, based on a variety of influenza indicators during a season dominated by circulation of influenza A(H1N1)pdm09 [11].

Aim

The work in this paper applies the same statistical framework as Lampos et al [9] (with a slightly improved supervised learning approach) on Twitter data for the influenza season of 2014/2015. We aim to assess the impact of influenza vaccine pilot trials in school age children on influenza transmission in those pilot areas. The 2014/2015 season was dominated by circulation of influenza A(h3N2) and influenza B. In addition, we examined the impact of vaccinating different target populations, specifically primary and/or secondary school-age children, on influenza rates in the general population. This analysis provides further insights into the most effective strategies for reducing community-wide influenza transmission. This work also aims to reevaluate the hypothesis that a statistical framework based on online user-generated content can form a valid source for more fine-grained influenza surveillance tasks, such as estimating the impact of a targeted intervention. We repeated the analysis for the 2013/2014 LAIV campaign that was previously studied in Lampos et al [9], but with revised pilot and control areas, for consistency with our study for the 2014/2015 season.

Methods

Data Sources

Two data sources were used for the experiments: geo-located Twitter posts related to ILI and official ILI rates provided by the Royal College of General Practitioners (RCGP) [12], the latter defining the *ground truth*. In addition, boundary data and population estimates from the Office for National Statistics (ONS) [13,14] were used to map the vaccine pilot and control areas.

Twitter Data

The Twitter data consisted of all exactly geo-located Twitter posts in England from August 29, 2011 to August 30, 2015, which comprise approximately 1% of all tweets made by users in England. This number is a rough estimate based on approximately 20% of the United Kingdom population using Twitter, with 33% of active users assumed to be posting 5 tweets per day [15]. Our dataset consists of 350,000 geo-located tweets per day on average. As in Lampos et al [9], the same initial list of 36 *n*-grams (phrases with *n* words) related to ILI was created manually. Then, based on frequent cooccurrence with this list in the Twitter time series data, a set of 217 *n*-grams was extracted (n<5; see Multimedia Appendix 1).

The RCGP ILI rates used for model learning were only available on a weekly basis, so frequency rates of this set of *n*-grams for a period of 7 days prior to any given day were computed, and formed the explanatory variables. To estimate the impact on the pilot areas, *n*-gram frequencies of tweets geo-located in the chosen pilot and control areas during the intervention period were used.

Official Health Reports

Weekly ILI estimates were provided by the RCGP, a sentinel network of approximately 100 practices in England, which covers a registered population of approximately 1 million persons [12]. These ILI estimates represent the weekly incidence rate of ILI cases/consultations per 100,000 patients registered with eligible practices during that week [12]. The data used cover the period from August 29, 2011 to August 30, 2015 for England.

Pilot and Control Areas

A total number of 140 local authorities implemented vaccinations as part of the pilot program. To create a suitable list of pilot areas for the impact assessment, these areas were combined on a county level, where possible. This list included a large amount of *Secondary school* pilot areas (37), so only the most populated ones were considered, whilst ensuring an even geographical distribution throughout the country. The geographical distribution and the areas' population sizes were defined using ONS boundary data and population estimates of England, respectively [13,14]. Of the 7 *Primary and Secondary school* pilot areas, 3 were eliminated due to small size or because they were enclosed within another pilot area. Pilot areas involving special schools were ignored, as these included only a small number of schools and were thus unlikely to provide any significant community-wide benefits. This preprocessing resulted in 6 *Primary school*, 4 *Primary and Secondary school*, and 7 *Secondary school* pilot areas.

A list of eligible control locations was chosen according to the following criteria: appropriate distance from pilot areas, a moderate population size, and a plausible geographical spread. These criteria resulted in a list of 16 control areas. Nonoverlapping boundary rectangles represented by their North-East and South-West corners were created around the chosen pilot and control areas. The geographical distribution of the pilot and control areas is shown in Figure 1. Table 1 lists the pilot areas considered for this study. For a full list of control and pilot areas, see Multimedia Appendix 2.

Figure 1. Geographical distribution of the pilot and control areas chosen for the study with their corresponding boundary boxes. Control areas with red boxes have a distance of at least 10 km to any pilot area. The “Secondary” and “Primary and Secondary” pilot areas that were excluded from the study are shown without boundary boxes and in a lighter shade of blue and green, respectively. Contains National Statistics and OS data, Crown copyright and database right.

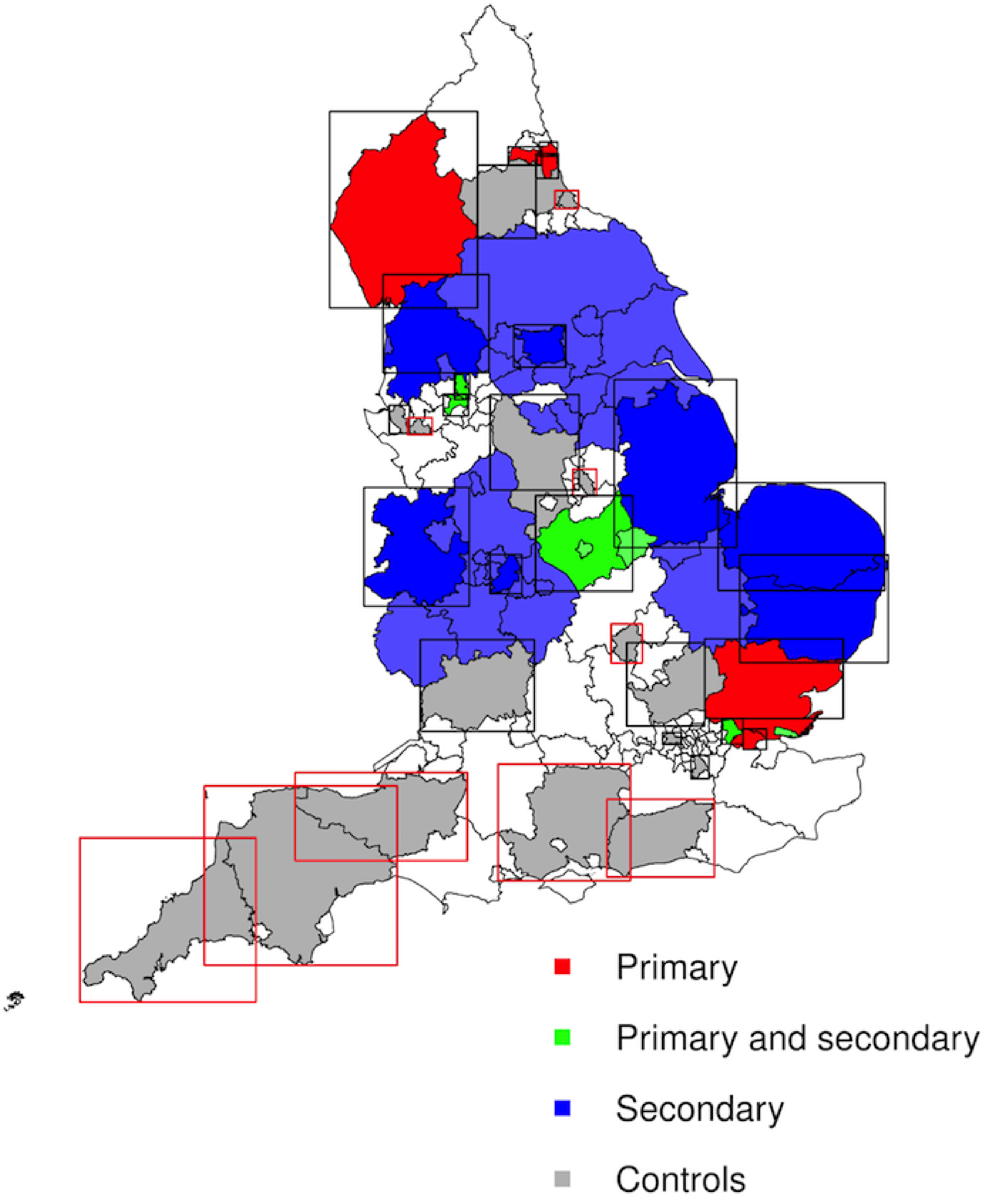


Table 1. Pilot areas considered for this study during the 2014/2015 LAIV program with their respective population size [14] and geographical boundary rectangle corner coordinates. Pilot areas that were also used or have partial overlap with the ones used in the 2013/2014 LAIV program are highlighted in italics.

Location	Pilot	Population	North-East corner[a]	South-West corner[b]
Thurrock	Primary school	163,270	51.568, 0.551	51.448, 0.334
Gateshead	*Primary school*	*200,505*	*54.984, -1.510*	*54.878, -1.853*
South Tyneside	Primary school	148,740	55.011, -1.352	54.928, -1.536
Sunderland	Primary school	276,889	54.944, -1.346	54.799, -1.569
Cumbria	*Primary school*	*497,874*	*55.189, -2.159*	*54.040, -3.641*
Essex	*Primary school*	*1,431,953*	*52.093, 1.297*	*51.632, -0.020*
Lancashire	Secondary school	1,184,735	54.240, -2.045	53.667, -3.085
Birmingham	Secondary school	1,101,360	52.609, -1.729	52.381, -2.034
Norfolk	Secondary school	877,710	52.993, 1.745	52.355, 0.154
Leeds	Secondary school	766,399	53.946, -1.290	53.699, -1.800
Suffolk	Secondary school	738,512	52.550, 1.769	51.932, 0.340
Lincolnshire	Secondary school	731,516	53.616, 0.358	52.640, -0.821
Shropshire	Secondary school	310,121	52.998, -2.233	52.306, -3.236
Bury	*Primary and Secondary school*	*187,474*	*53.667, -2.234*	*53.512, -2.383*
Salford	Primary and Secondary school	242,040	53.542, -2.245	53.416, -2.490
Havering	*Primary and Secondary school* [c]	*245,974*	*51.632, 0.334*	*51.484, 0.138*
Leicestershire	*Primary and Secondary school*	*667,905*	*52.948, -0.664*	*52.392, -1.598*

[a] Longitude and latitude of the North-East edge of the bounding box

[b] Longitude and latitude of the South-West edge of the bounding box

[c] The secondary school program in Havering included the year 7 cohorts only (11-12 years)

Statistical Framework

The following sections provide a brief outline of the statistical framework that was implemented. Apart from a slightly improved supervised learning approach, this framework is based on the work by Lampos et al [9], in which it is described and validated in more detail. The method consists of first learning a nonlinear regression model to estimate ILI rates from *n*-grams based on user-generated content (tweets in this case). Thereafter, by making use of inferred ILI rates in matched pilot and control regions, a linear modeling approach was applied to assess the potential impact of the intervention in the pilot areas.

Estimating Disease Rates Using a Gaussian Process

The majority of techniques used to acquire infectious disease estimates from user-generated data involve the use of linear regression models [16-18]. Lampos et al showed that nonlinear methods can improve model performance, especially when working with a smaller feature space consisting of varying *n*-gram sizes [8]. The authors proposed the use of Gaussian Processes (GPs) to model ILI rates and successfully applied these to Twitter, Google, and Bing data [8,9]. See below for details of the GP model used in this study.

Let $\mathbf{X} \in \mathbb{R}^{N \times M}$ be the observation matrix with *N* weeks and *M* frequency rates of *n*-gram features. Then given inputs $\mathbf{x},\mathbf{x}' \in \mathbb{R}^M$ (representing rows of **X**), a GP can be defined as a statistical distribution for which any finite linear combination of samples is normally distributed and is written as:

$$f(\mathbf{x}) \sim \mathrm{GP}\left(\mu(\mathbf{x}), k(\mathbf{x},\mathbf{x}')\right)$$

Here $\mu(\mathbf{x})$ and $k(\mathbf{x},\mathbf{x}')$ represent the mean and covariance function (or kernel), respectively [19]. By assuming that $\mu(\mathbf{x})=0 \forall$ $i=1,\ldots, N$, the distribution is entirely determined by its covariance function. As our core kernel, the sum of two differently parameterized Matérn functions (k_M) [20], with degrees of freedom $v=3/2$ was found to be the most suitable for estimating ILI rates from Twitter data:

$$k_M(\mathbf{x},\mathbf{x}') = \sum_{m=1}^{2} \sigma_m{}^2 \left(1 + \frac{\sqrt{3}\|\mathbf{x}-\mathbf{x}'\|_2}{l_m}\right) \exp\left(-\frac{\sqrt{3}\|\mathbf{x}-\mathbf{x}'\|_2}{l_m}\right)$$

where σ_m represents the overall level of variance and l_m a characteristic length scale. Assuming that different *n*-gram sizes may vary in their usage and are likely to have a more concise semantic interpretation with an increasing *n*, we model them with different kernels. The fact that the sum of covariance functions forms a valid covariance function in itself allows for this and we have:

$$k(\mathbf{x},\mathbf{x}') = \left(\sum_{n=1}^{C} k_M(\mathbf{g}_n, \mathbf{g}'_n)\right) + k_{SE}(\mathbf{x},\mathbf{x}') + k_N(\mathbf{x},\mathbf{x}')$$

where $\mathbf{g}_n$ represents the features that belong to each *n*-gram category and C=3 is the number of *n*-gram categories (3-grams and 4-grams are merged in this particular model). To model noise, we use the sum of a squared exponential:

$$\mathrm{k}_{\mathrm{SE}}(\mathbf{x},\mathbf{x}') = \sigma_{\mathrm{SE}}{}^2\exp\left(-\ \|\mathbf{x}-\mathbf{x}'\|_2^2/2l^2\right)$$

and a noise function:

$$\mathrm{k}_{\mathrm{N}}(\mathbf{x}_i,\mathbf{x}_j) = \sigma_{\mathrm{N}}{}^2\delta(\mathbf{x},\mathbf{x}')$$

(δ is a Kronecker delta function), as defined in [19].

GP regression involves minimizing the negative log-marginal likelihood function:

$$\underset{\sigma_{1_1},\ldots,\sigma_{1_C},l_{1_1},\ldots,l_{1_C},\sigma_{2_1},\ldots,\sigma_{2_C},l_{2_1},\ldots,l_{2_C},\sigma_{\mathrm{N}},\sigma_{\mathrm{SE}},l}{\operatorname{argmin}} \left((\mathbf{y}-\boldsymbol{\mu})^T\mathbf{K}(\mathbf{y}-\boldsymbol{\mu}) + \log|\mathbf{K}|\right)$$

where $\mathbf{y}$ denotes the ILI rates time-series, $(\mathbf{K})_{ij}=\mathrm{k}(\mathbf{x}_\mathrm{i},\mathbf{x}_\mathrm{j})$ and $\boldsymbol{\mu}=(\mu(\mathbf{x}_1),\ldots,\mu(\mathbf{x}_N))$. Once the model is learnt, newly observed feature frequency rates $\mathbf{x}_*$ result in new ILI rate estimates $\mathbf{y}_*$ by computing $\mathrm{E}[\mathbf{y}_*|\mathbf{y},\Omega,\mathbf{x}_*]$, the mean of the posterior predictive distribution. The performance of the model was measured using a 10-fold cross validation (random temporal splits) on the training set, using the average Pearson correlation (*r*) and the mean absolute error (MAE).

Estimating the Impact of the LAIV Program

Once the GP model was trained, the impact of the LAIV campaign in pilot areas could be estimated using the methodology outlined in Lampos et al, Section 3.3 [9], which we briefly describe here as well.

Given a set of pilot and control areas, *n*-gram frequencies of Twitter posts geo-located in those areas are extracted for a period before and during the intervention. ILI rate estimates can then be computed for all areas and supersets of areas using a pretrained GP model and we denote these with $\mathbf{q}_\mathrm{v}$ and $\mathbf{q}_\mathrm{c}$ for pilot and control areas, respectively. By looking at these ILI estimates for a number of weeks, $\tau=\{\ t_1,\ldots,\ t_N\ \}$, prior to the intervention, control and pilot locations with similar influenza activity can be matched based on a strong Pearson correlation, $r(\mathbf{q}_\mathrm{v}^\tau,\mathbf{q}_\mathrm{c}^\tau)$. Assuming a linear relationship in ILI rates between locations with similar influenza activity, a linear regression model can be learnt using $\mathbf{q}_\mathrm{v}^\tau$ and $\mathbf{q}_\mathrm{c}^\tau$ (ie, the ILI estimates prior to the intervention in the various matched area pairs):

$$q_v^{t_i} = q_c^{t_i}\omega + \beta + \varepsilon_i$$

where $\omega,\beta,\varepsilon_i$ denote the regression's weight and intercept, and independent, zero-centered noise, respectively. Using $\mathbf{q}_\mathrm{c}$, the ILI estimates in the control areas during the intervention, this linear model can then predict the hypothetical ILI rates in pilot locations during the intervention had the intervention not taken place:

$$\mathbf{q}_\mathrm{v}^* = \mathbf{q}_\mathrm{c}\omega + \mathbf{b}$$

where $\mathbf{b}\in\mathbb{R}^\mathrm{N}$ with $(\mathbf{b})_\mathrm{k}=\beta\forall k=1,\ldots,N$.

Comparing these hypothetical ILI rates to the ILI rates estimated by the GP model during the intervention allows the impact of the campaign to be estimated. The following measures were applied:

$$\delta_v = \overline{\mathbf{q}_\mathrm{v}} - \overline{\mathbf{q}_\mathrm{v}^*}$$

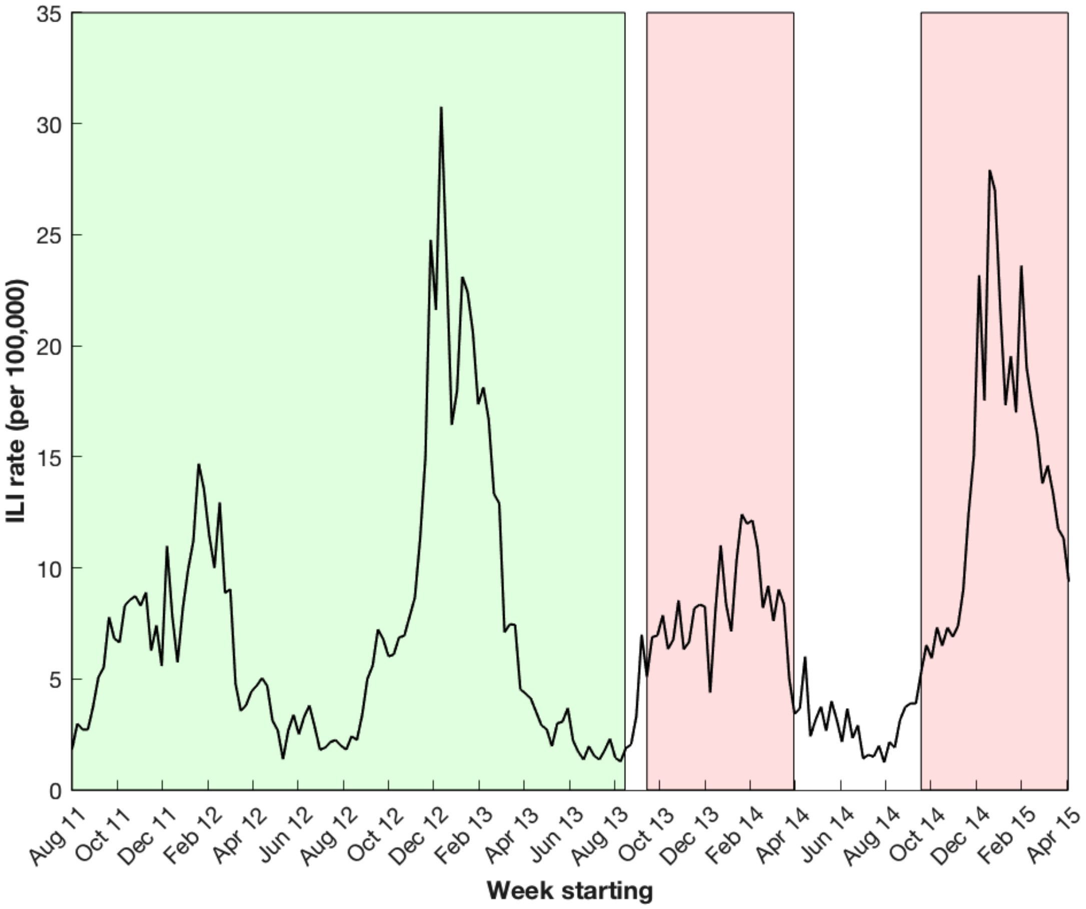
$$\theta_v = \left(\overline{\mathbf{q}_\mathrm{v}} - \overline{\mathbf{q}_\mathrm{v}^*}\right)/\overline{\mathbf{q}_\mathrm{v}^*}$$

where $\overline{\mathbf{q}}$ denotes the mean value of $\mathbf{q}$. Thus, δ_v and θ_v measure the absolute and relative mean impact of the intervention, respectively. Confidence intervals for these measures are produced using bootstrap sampling [21]. This calculation involves sampling with replacement the residuals ε_i of the linear regression, adding them to the fitted values, and then running the linear model for these, which produces estimates for β and ω. These values are then applied to a sampled (with replacement) set of $\mathbf{q}_\mathrm{v}$ and $\mathbf{q}_\mathrm{c}$. Repeating this procedure 100,000 times creates sets of estimates for δ_v and θ_v from which we can derive confidence intervals using the 0.025 and 0.975 quantiles, provided that their distributions are unimodal and symmetric. Results are considered statistically significant if absolute values are higher than two standard deviations of the bootstrap estimates [9,22].

Results

We present an assessment of the impact of the childhood LAIV campaign during the 2013/2014 and 2014/2015 influenza seasons based on the previously described methodology. The GP model was trained on RCGP ILI rates in England and Figure 2 shows the RCGP ILI rates used, with the preintervention correlation period and the two impact assessment periods highlighted.

Figure 2. Weekly influenza-like illness (ILI) rate (per 100,000) provided by the Royal College of General Practitioners (RCGP) in England with the pre-intervention correlation period highlighted in green and the two impact assessment periods (2013/14 and 2014/15 influenza seasons) highlighted in red.



Performance of the Supervised Model for Estimating ILI Rates

A GP regression model was trained using weekly Twitter data geo-located in England from August 29, 2011 to August 30, 2015 and the corresponding RCGP ILI rates. Based on a 10-fold cross validation, an average Pearson correlation r=0.84 with a standard deviation of 0.08 and average MAE of 2.42 (weekly ILI rate per 100,000 people) with a standard deviation of 0.52 were measured. This approach is in line with the performance of the GP model used in the previous impact assessment [9].

Impact Estimates of the LAIV Program

Using the GP model trained on a national level (England), ILI rates for the chosen pilot locations were estimated. This was done for individual pilot locations, the set of all pilot locations, and sets of pilot locations in which the same cohorts were vaccinated (ie, *Primary school*, *Secondary school*). An exhaustive search of all possible combinations of control areas was performed. These combinations of control locations were matched to the sets of pilot locations during a period prior to the start of the LAIV campaign (August 29, 2011 to September 1, 2013) based on similar influenza activity, as measured by Pearson correlation. The 2013/2014 influenza season is not included in this correlation phase, as this involved the vaccination of 2 and 3-year-olds nationally and a number of primary school age pilot areas, which could change the linear relationship between certain control and pilot locations. For each pilot area and set of pilot areas, the most highly correlated combination of control areas was used to then estimate the impact of the LAIV campaign for the 2014/2015 influenza season. There is some overlap with the pilot areas of the previous influenza season, so the same analysis was redone for the 2013/2014 season (in this case with a different set of control areas) so results could be compared to previous studies [9,11].

Table 2 and Table 3 show the results for individual pilot locations, and sets of them for the 2014/2015 and 2013/2014 influenza season, respectively. For each area, the tables include the Pearson correlation r, the mean and 95% confidence intervals of 100,000 bootstrap estimates of the absolute and relative mean impact δ_v and θ_v during the intervention period, the number of control areas chosen n (c), and the size of the population targeted in the pilot *Pop* (v) and matched collection of control *Pop* (c) areas. The distribution of the bootstrap estimates was assessed graphically and seemed unimodal. Thus, statistically significant results are based on absolute values being higher than two standard deviations of the bootstrap estimates and are highlighted in italics. In addition, a significant preintervention correlation was necessary for reliable impact estimates, which we defined as being a Pearson correlation >0.60, as was done in the previous study [9].

Table 2. Estimates of the impacts of LAIV pilot program during the 2014/2015 influenza season in individual pilot locations and supersets of them. For each area considered, the precampaign Pearson correlation r with chosen control areas, the mean and 95% confidence intervals of the absolute and relative mean impact δ_V and θ_V during the intervention period, the number of control areas chosen *n(c)*, and the size of the population targeted in the chosen vaccination *Pop(v)* and control *Pop(c)* areas are presented. Statistically significant results are highlighted in italics.

Pilot area	r [a]	δ_V [b]	θ_V [c]	*n* (*c*)[d]	*Pop* (*v*)[e]	*Pop* (c)[f]
All vaccinated	0.89	-0.50 (-2.77 to 1.99)	-4.51 (-25.72 to 22.61)	10	9,772,977	5,066,069
All "Primary school"	*0.71*	*-1.15 (-2.19 to -0.15)*	*-16.97 (-30.09 to -2.42)*	*8*	*2,719,231*	*2,371,367*
All "Primary and Secondary school"	0.84	-0.06 (-1.50 to 1.43)	-0.30 (-16.71 to 19.36)	6	1,097,419	2,174,854
All "Primary school" and "Primary and Secondary school"	0.85	-1.35 (-3.37 to 0.66)	-13.01 (-30.54 to 7.31)	9	4,062,624	3,601,377
All "Secondary school"	0.83	0.06 (-1.58 to 1.90)	1.41 (-19.40 to 28.40)	7	5,710,353	4,038,921
Cumbria ("Primary school")	0.59	0.04 (-0.24 to 0.33)	1.07 (-5.75 to 8.17)	7	497,874	3,999,608
Essex ("Primary school")	0.68	-0.32 (-1.13 to 0.51)	-5.91 (-20.56 to 10.58)	8	1,431,953	3,199,730
Gateshead ("Primary school")	*0.59*	*-0.39 (-0.74 to -0.04)*	*-8.46 (-15.56 to -1.02)*	*4*	*200,505*	*1,551,060*
South Tyneside ("Primary school")	*0.34*	*0.25 (0.03 to 0.52)*	*6.82 (0.81 to 14.07)*	*3*	*148,740*	*1,697,971*
Sunderland ("Primary school")	0.54	0.12 (-0.05 to 0.32)	3.20 (-1.38 to 8.38)	3	276,889	1,119,136
Thurrock ("Primary school")	0.32	0.04 (-0.14 to 0.23)	1.01 (-3.56 to 6.24)	3	163,270	753,563
Bury ("Primary and Secondary school")	0.32	-0.11 (-0.37 to 0.12)	-2.60 (-8.94 to 3.13)	2	187,474	893,813
Leicestershire ("Primary and Secondary school")	0.81	0.32 (-0.70 to 1.38)	4.97 (-10.01 to 21.22)	6	667,905	2,756,865
Salford ("Primary and Secondary school")	0.67	0.40 (-0.20 to 1.01)	8.45 (-3.96 to 22.00)	7	242,040	4,183,184
Havering ("Primary and Secondary school"-year 7)	0.48	-0.03 (-0.35 to 0.31)	-0.55 (-8.23 to 7.79)	4	245,974	1,742,705
Birmingham ("Secondary school")	0.79	0.53 (-0.27 to 1.34)	10.36 (-4.86 to 27.21)	10	1,101,360	5,435,742
Lancashire ("Secondary school")	0.65	0.18 (-0.78 to 1.13)	3.45 (-13.41 to 21.40)	8	1,184,735	3,463,060
Leeds ("Secondary school")	0.63	0.54 (-0.40 to 1.51)	10.81 (-7.41 to 30.98)	7	766,399	2,731,293
Lincolnshire ("Secondary school")	0.66	-0.29 (-0.78 to 0.19)	-6.09 (-16.20 to 4.25)	6	731,516	1,737,168
Norfolk ("Secondary school")	0.71	-0.12 (-0.60 to 0.35)	-2.31 (-11.55 to 7.25)	6	877,710	2,784,394
Shropshire ("Secondary school")	0.35	0.13 (-0.13 to 0.39)	3.30 (-3.18 to 9.71)	6	310,121	2,833,659
Suffolk ("Secondary school")	0.59	0.10 (-0.34 to 0.53)	2.24 (-7.54 to 12.35)	5	738,512	2,015,339

[a] r: The precampaign Pearson correlation with the chosen aggregation of control areas

[b] δ_V: The absolute difference in the mean ILI rate during the intervention period

[c] θ_V: The relative difference in the mean ILI rate during the intervention period

[d] *n* (*c*): The number of aggregated control areas chosen

[e] *Pop* (*v*): The size of the population targeted in the chosen vaccination areas

[f] *Pop* (*c*): The size of the population targeted in the chosen aggregation of control areas

Table 3. Estimates of the impacts of the LAIV pilot program during the 2013/2014 influenza season in individual pilot locations and supersets of these locations. For each area considered, the precampaign Pearson correlation *r* with chosen control areas, the mean and 95% confidence intervals of the absolute and relative mean impact δ_V and θ_V during the intervention period, the number of control areas chosen *n(c)*, and the size of the population targeted in the chosen vaccination *Pop(v)* and control *Pop(c)* areas are presented. Statistically significant results are highlighted in italics.

Pilot area	*r* [a]	δ_V [b]	θ_V [c]	*n* (*c*) [d]	*Pop* (*v*) [e]	*Pop* (c) [f]
All vaccinated (Primary school)	*0.82*	*-1.03 (-2.00 to -0.10)*	*-13.77 (-25.01 to -1.45)*	*9*	*3,231,685*	*3,601,377*
Leicestershire (Primary school)	0.81	-0.28 (-1.02 to 0.47)	-4.44 (-15.93 to 7.95)	6	667,905	2,756,865
Essex (Primary school)	0.68	0.34 (-0.30 to 1.12)	7.45 (-6.41 to 24.32)	8	1,431,953	3,199,730
Gateshead (Primary school)	0.59	0.38 (-0.06 to 0.85)	9.11 (-1.40 to 20.76)	4	200,505	1,551,060
Cumbria (Primary school)	0.59	0.36 (-0.00 to 0.75)	9.12 (-0.07 to 19.11)	7	497,874	3,999,608
Havering (Primary school)	0.48	0.15 (-0.19 to 0.52)	3.80 (-4.99 to 13.43)	4	245,974	1,742,705
Bury (Primary school)	0.32	-0.09 (-0.34 to 0.14)	-2.40 (-8.44 to 3.64)	2	187,474	893,813

[a] *r*: The precampaign Pearson correlation with the chosen aggregation of control areas

[b] δ_V: The absolute difference in the mean ILI rate during the intervention period

[c] θ_V: The relative difference in the mean ILI rate during the intervention period

[d] *n* (*c*): The number of aggregated control areas chosen

[e] *Pop* (*v*): The size of the population targeted in the chosen vaccination areas

[f] *Pop* (*c*): The size of the population targeted in the chosen aggregation of control areas

For the 2014/2015 influenza season, correlations ranged from 0.32 to 0.89, and pilot areas with larger populations tend to have more control areas, larger populations of control areas, and higher Pearson correlations. The only significant impact was observed in the *Primary school* age pilot areas, for which the results suggest that during the 2014/2015 influenza season the mean ILI rate was reduced by 16.97% (95% CI 2.42-30.09). For the individual locations, Gateshead and South Tyneside did show significant results, but their precampaign correlations were 0.59 and 0.34, respectively; both were less than the predefined threshold of 0.60, which makes their impact estimates possibly less reliable.

The correlations for the 2013/2014 influenza season ranged from 0.32 to 0.82, and whilst none of the individual locations demonstrated significant results, all pilots together estimated a statistically significant impact of a 13.77% (95% CI 1.45-25.01) reduction in the mean ILI rate during that season. Note that for the 2013/2014 season, the primary school-age vaccination was the only program implemented across all pilot areas.

Discussion

Principal Results

By using social media content to assess the impact of the childhood influenza pilot program in England in 2013/2014 and 2014/2015, statistically significant results suggest a reduction in the mean ILI rate of approximately 17% (Table 2, row 2, column 4) across all ages in *Primary school* age pilot areas only during the 2014/2015 influenza season and 14% (Table 3, row 1, column 4) in the aggregation of *Primary school* age vaccinated areas during the 2013/2014 influenza season.

Comparison With Prior Work

Both impact estimates are in line with results from independent studies by PHE that used traditional surveillance systems [3,11]. For the 2014/2015 season, however, the impact results are generally lower than expected with only a few statistically significant results. For example, it was expected that the *Primary and Secondary school* or the combined set of *Primary school* and *Primary and Secondary school* pilot locations would yield significant impacts, as they included a similar program to that in the *Primary school* pilot areas. Looking at the boundary boxes in more detail (Figure 1) shows that of the 4 *Primary and Secondary school* pilot areas, Leicestershire and Salford both include substantial parts of nonpilot areas, which is likely to have biased their results and underestimated effect sizes. The lack of statistically significant results across all individual locations is possibly due to the sparsity of the Twitter data available. For example, the individual *Primary school* pilot areas did not yield statistically significant impact estimates (with the exception of Gateshead and South Tyneside, which did show significant results, but their preintervention correlations were below the 0.60 threshold), whilst the aggregation of all *Primary school* areas did.

The previous study by Lampos et al implemented a similar approach using Twitter and Bing data to assess the impact of the LAIV pilots during the 2013/2014 influenza season [9]. This study estimated the impact to be approximately 33% for the aggregation of all pilot locations based on Twitter data, which is more than double what was found in this study. The discrepancy between these results is most likely due to two factors. First, the pilot areas used for the 2013/2014 season in the present study are slightly larger than those in the previous one, as some of the reused pilot areas have been expanded. This issue particularly applies to the boundary boxes for Leicestershire and Essex, as the previous study only included parts of these areas. Second, apart from one control area (Liverpool), most of the previous control areas were part of the 2014/2015 pilot program, and thus not reusable. New control areas were therefore selected, which may explain the discrepancy in impact estimates. Nevertheless, given that both studies exhibited a significant impact, the methodology produces qualitatively consistent results for the same influenza season, even when using a different set of control and pilot areas.

Conclusions

There is a strong indication that the primary school age vaccination program has the potential to be an effective strategy in reducing influenza transmission in the general population. This notion supports the ongoing rollout of the campaign for primary school children. For a secondary school-only vaccination program offering the vaccine to just two-year cohorts (and not to all children of secondary age), there is no clear evidence of any population-wide effect. Both of these conclusions are in line with findings from previous studies and complement traditional surveillance sources in exhibiting community-wide effects of the LAIV pilot campaign [3,9,11,23].

Most current influenza surveillance schemes rely on established health systems. Although these schemes provide important information on health care-related burden of disease and potential reductions due to vaccine impact, several provide less direct insight into community-wide transmission. User-generated content from social media offers rapid access to a larger range of the population, which has the potential of including a wider community (ie, including those that do not seek medical attention) and thus offers a valuable complementary source for the surveillance and evaluation of public health programs.

Limitations

There are several potential limitations in this study. Work is still needed to refine the methods used to deal with issues such as noise, model and data biases, and the fact that estimates from user-generated content are not directly based on actual ILI cases. More advanced natural language processing techniques may deliver more accurate results [24]. The choice of control areas requires further refinement; we are seeking an even geographical distribution as well as an adequate distance from pilot areas to avoid regional biases, and to isolate the potential impact observed in pilot areas, respectively. Furthermore, the methodology is highly dependent on the quantity and type of user-generated data that is available, as this determines the accuracy and interpretation of the ILI rate estimates. The majority of Twitter users, for example, are between the ages of 15-44 years with a higher proportion situated in urban/suburban areas [25]. This factor may skew results towards illness in certain demographic groups. The current framework conducts ILI rate modeling by training on syndromic surveillance data (from RCGP), such that biases that are found there are also passed onto the models. Furthermore, even if these biases can be avoided, there is an issue that no definite ground truth exists to allow for a proper verification.

Future Work

Future work could aim at moving towards unsupervised models that do not depend on traditional surveillance sources for training purposes. These models could produce their own, independent ILI indicators based solely on user-generated content with the potential of being able to tap into the bottom part of the disease population pyramid [10]. Inference of the demographics of users, such as age [26], socioeconomic status [27,28], or severity of disease [29] could be another focus of forthcoming work. Pebody et al showed that for both influenza seasons the impact of the pilot program was lower as influenza end-points of infection became more severe, which is an insight that the current modeling framework is unable to pick up on [3,11]. With suitable data access in the future, this framework has the potential of assessing the impact of intervention programs whose uptake is variable. The applicability of this framework extends beyond influenza, but across a number of health interventions, thereby allowing for a timely and potentially cost-effective complementary to the collection of traditional surveillance data.

Acknowledgments

This work has been funded by the EPSRC through i-sense (*Early-Warning Sensing Systems for Infectious Diseases*; EP/K031953/1). We would also like to thank the RCGP for providing anonymized and aggregate syndromic surveillance data.

Authors' Contributions

VL, IJC, RP, EY-T, and MW conceived the general concept of this research; EY-T provided the Twitter data; VL designed the models; VL and MW performed the experiments; MW and VL wrote the paper; all coauthors reviewed and commented on the manuscript, and approved the final version.

Multimedia Appendix 1

The list of the 217 *n*-grams used as features in our predictive models for ILI rates.

[PDF File (Adobe PDF File), 50KB - rehab_v4i2e13_app1.pdf]

Multimedia Appendix 2

A table of the pilot and control areas chosen with their respective population size, distance to closest pilot areas, and geographical boundary rectangle corner coordinates.

[PDF File (Adobe PDF File), 76KB - jmir_v19i12e416_app2.pdf]

References

1. Joint Committee on Vaccination and Immunisation. 2011 Oct 5. Minute of the meeting held on Wednesday 5 October 2011 URL: http://webarchive.nationalarchives.gov.uk/20120907090205/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@ab/documents/digitalasset/dh_133598.pdf [accessed 2017-06-15] [WebCite Cache ID 6rEtwOgu3]
2. Baguelin M, Flasche S, Camacho A, Demiris N, Miller E, Edmunds WJ. Assessing optimal target populations for influenza vaccination programmes: an evidence synthesis and modeling study. PLoS Med 2013 Oct;10(10):e1001527 [FREE Full text] [doi: 10.1371/journal.pmed.1001527] [Medline: 24115913]
3. Pebody RG, Green HK, Andrews N, Boddington NL, Zhao H, Yonova I, et al. Uptake and impact of vaccinating school age children against influenza during a season with circulation of drifted influenza A and B strains, England, 2014/15. Euro Surveill 2015;20(39):30029 [FREE Full text] [doi: 10.2807/1560-7917.ES.2015.20.39.30029] [Medline: 26537222]
4. Milinovich GJ, Williams GM, Clements AC, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis 2014 Feb;14(2):160-168. [doi: 10.1016/S1473-3099(13)70244-5] [Medline: 24290841]
5. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009 Feb 19;457(7232):1012-1014. [doi: 10.1038/nature07634] [Medline: 19020500]
6. Lampos V, Cristianini N. Nowcasting events from the social web with statistical learning. ACM Trans Intell Syst Technol 2012 Sep 01;3(4):1-22. [doi: 10.1145/2337542.2337557]
7. Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. PLoS Curr 2014;6:1 [FREE Full text] [doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117] [Medline: 25642377]
8. Lampos V, Miller AC, Crossan S, Stefansen C. Advances in nowcasting influenza-like illness rates using search query logs. Sci Rep 2015 Aug 03;5:12760 [FREE Full text] [doi: 10.1038/srep12760] [Medline: 26234783]
9. Lampos V, Yom-Tov E, Pebody R, Cox IJ. Assessing the impact of a health intervention via user-generated Internet content. Data Min Knowl Disc 2015 Jul 2;29(5):1434-1457. [doi: 10.1007/s10618-015-0427-9]
10. Gibbons CL, Mangen MJ, Plass D, Havelaar AH, Brooke RJ, Kramarz P, Burden of Communicable diseases in Europe (BCoDE) consortium. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. BMC Public Health 2014 Feb 11;14:147 [FREE Full text] [doi: 10.1186/1471-2458-14-147] [Medline: 24517715]
11. Pebody R, Green H, Andrews N, Zhao H, Boddington N, Bawa Z. Uptake and impact of a new live attenuated influenza vaccine programme in England: early results of a pilot in primary school-age children, 2013/14 influenza season. Eurosurveillance 2014;19(22):20823. [doi: 10.2807/1560-7917.es2014.19.22.20823]
12. Correa A, Hinton W, McGovern A, van VJ, Yonova I, Jones S, et al. Royal College of General Practitioners Research and Surveillance Centre (RCGP RSC) sentinel network: a cohort profile. BMJ Open 2016 Apr 20;6(4):e011092 [FREE Full text] [doi: 10.1136/bmjopen-2016-011092] [Medline: 27098827]
13. Office for National Statistics. UK Data Service. 2011. 2011 Census: boundary data (England and Wales) data collection URL: http://census.ukdataservice.ac.uk/get-data/boundary-data.aspx [WebCite Cache ID 6rEuELJCJ]
14. Office for National Statistics. Annual mid-year population estimates. 2015. URL: http://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/2015-06-25 [accessed 2017-06-15] [WebCite Cache ID 6rEvoMShY]
15. Rose K. Rosemcgrory.co.uk. 2016. UK Social Media Statistics for 2016 URL: http://www.rosemcgrory.co.uk/2016/01/04/social-media-statistics-2016/ [accessed 2017-06-15] [WebCite Cache ID 6rEw81MAf]
16. Lampos V, Cristianini N. Tracking the influenza pandemic by monitoring the social web. In: 2nd International Workshop on Cognitive Information Processing. 2010 Presented at: 2nd International Workshop on Cognitive Information Processing; 2010; Elba Island, Italy p. 441-416. [doi: 10.1109/CIP.2010.5604088]
17. Lampos V, De Bie T, Cristianini N. Flu Detector - tracking epidemics on Twitter. LNCS 2010;6323:599-602. [doi: 10.1007/978-3-642-15939-8_42]
18. Culotta A. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. Lang Resources & Evaluation 2012 May 13;47(1):217-238. [doi: 10.1007/s10579-012-9185-0]
19. Rasmussen C, Williams C. Gaussian processes for machine learning. Cambridge, MA: MIT Press; 2006.
20. Matern B. Spatial variation. Berlin: Springer-Verlag; 1986.
21. Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
22. Lambert D, Pregibon D. Online effects of offline ads. 2008 Presented at: Proceedings of the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising - ADKDD '08; 2008; Las Vegas, Nevada, USA. [doi: 10.1145/1517472.1517474]
23. Public Health England. 2014. Surveillance of influenza and other respiratory viruses in the United Kingdom: winter 2014 to 2015 URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/429617/Annualreport_March2015_ver4.pdf [accessed 2017-06-15] [WebCite Cache ID 6rEyjgzcr]
24. Lampos V, Zou B, Cox I. Enhancing feature selection using word embeddings. 2017 Presented at: Proceedings of the 26th International Conference on World Wide Web; 2017; Perth, Australia p. 695-704. [doi: 10.1145/3038912.3052622]
25. Duggan M, Ellison N, Lampe C, Lenhart A, Madden M. Pew Research Center: Internet, Science & Tech. 2015. Demographics of Key Social Networking Platforms Internet URL: https://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/#twitter [accessed 2017-09-11] [WebCite Cache ID 6tOVyGVDv]

26. Rao D, Yarowsky D, Shreevats A, Gupta M. Classifying latent user attributes in Twitter. 2010 Presented at: Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents; 2010; Toronto, ON, Canada p. 37-44. [doi: 10.1145/1871985.1871993]
27. Preoţiuc-Pietro D, Lampos V, Aletras N. An analysis of the user occupational class through Twitter content. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015 Presented at: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics; 2015; Beijing, China p. 1754-1764 URL: http://www.aclweb.org/anthology/P15-1169
28. Lampos V, Aletras N, Geyti J, Zou B, Cox I. Inferring the socioeconomic status of social media users based on behaviour and language. 2016 Presented at: Proceedings of the 38th European Conference on Information Retrieval; 2016; Padua, Italy p. 689-695. [doi: 10.1007/978-3-319-30671-1_54]
29. Yom-Tov E, Johansson-Cox I, Lampos V, Hayward AC. Estimating the secondary attack rate and serial interval of influenza-like illnesses using social media. Influenza Other Respir Viruses 2015 Jul;9(4):191-199 [FREE Full text] [doi: 10.1111/irv.12321] [Medline: 25962320]

Abbreviations

GP: Gaussian Process
ILI: influenza-like illness
LAIV: live attenuated influenza vaccine
MAE: mean absolute error
ONS: Office for National Statistics
PHE: Public Health England
RCGP: Royal College of General Practitioners

Edited by N Collier, M Conway, N Limsopatham, V Lampos, A Culotta; submitted 15.06.17; peer-reviewed by K Saha, J Lei, C Basch, P Wark; comments to author 01.09.17; revised version received 15.09.17; accepted 02.10.17; published 21.12.17

Please cite as:
Wagner M, Lampos V, Yom-Tov E, Pebody R, Cox IJ
Estimating the Population Impact of a New Pediatric Influenza Vaccination Program in England Using Social Media Content
J Med Internet Res 2017;19(12):e416
URL: http://www.jmir.org/2017/12/e416/
doi:10.2196/jmir.8184
PMID:29269339