

## Assessing the Impact of a Health Intervention via User-generated Internet Content

Vasileios Lampos · Elad Yom-Tov ·  
Richard Pebody · Ingemar J. Cox

Received: March 12, 2015 / Accepted: June 19, 2015

**Abstract** Assessing the effect of a health-oriented intervention by traditional epidemiological methods is commonly based only on population segments that use healthcare services. Here we introduce a complementary framework for evaluating the impact of a targeted intervention, such as a vaccination campaign against an infectious disease, through a statistical analysis of user-generated content submitted on web platforms. Using supervised learning, we derive a nonlinear regression model for estimating the prevalence of a health event in a population from Internet data. This model is applied to identify control location groups that correlate historically with the areas, where a specific intervention campaign has taken place. We then determine the impact of the intervention by inferring a projection of the disease rates that could have emerged in the absence of a campaign. Our case study focuses on the influenza vaccination program that was launched in England during the 2013/14 season, and our observations consist of millions of geo-located search queries to the Bing search engine and posts on Twitter. The impact estimates derived from the application of the proposed statistical framework support conventional assessments of the campaign.

**Keywords** Gaussian Process · infectious diseases · intervention · search query logs · social media · supervised learning · user-generated content

---

Vasileios Lampos  
Department of Computer Science, University College London  
40 Melton Street, NW1 2FD, UK  
E-mail: v.lampos@ucl.ac.uk (*corr. author*)

Elad Yom-Tov  
Microsoft Research, US

Richard Pebody  
Public Health England, UK

Ingemar J. Cox  
Department of Computer Science  
University College London, UK and University of Copenhagen, Denmark

## 1 Introduction

Infectious diseases are a major concern for public health and a significant cause of death worldwide (Binder et al. 1999; Morens et al. 2004; Jones et al. 2008). Various health interventions, such as improved sanitation, clean water and immunization programs, assist in reducing the risk of infection (Cohen 2000). To monitor infectious diseases as well as evaluate the impact of control and prevention programs, health organizations have established a number of surveillance systems. Typically, these schemes, apart from requiring an established health system, only cover cases that result in healthcare service utilization. Therefore, they are not always able to capture the prevalence of a disease in the general population, where it is likely to be more common (Reed et al. 2009; Briand et al. 2011).

Recent research efforts have proposed various ways for taking advantage of online information to gain a better understanding of offline, real-world situations. Particular interest has been drawn on the modeling of user-generated web content, either in the form of social media text snippets or search engine query logs. Numerous works have provided statistical proof for the predictive capabilities of these resources with applications spreading across the domains of finance (Bollen et al. 2011), politics (O’Connor et al. 2010; Lampos et al. 2013) and healthcare (Ginsberg et al. 2009; Lampos and Cristianini 2010; Culotta 2010). Focusing on the domain of health, the development of models for nowcasting infectious diseases, such as influenza-like illness (ILI),<sup>1</sup> has been a central theme (Milinovich et al. 2014). Initial indications that content from Yahoo’s (Polgreen et al. 2008) or Google’s (Ginsberg et al. 2009) search engine are good ILI indicators, were followed by a series of approaches using the microblogging platform of Twitter as an alternative, publicly available source (Lampos et al. 2010; Signorini et al. 2011; Lamb et al. 2013).

Tracking the prevalence of an infectious disease from Internet activities establishes a complementary and perhaps more sensitive sensor than doctor visits or hospitalizations because it provides access to the bottom of the disease pyramid, i.e., potential cases of infection many of whom may not use the healthcare system. Online data sources do have disadvantages, including noise and ambiguity, and respond not just to changes in disease prevalence, but also to other factors, especially media coverage (Cook et al. 2011; Lazer et al. 2014). Nevertheless, the learning approaches that convert this content to numeric indications about the rate of a disease aim to eliminate most of the aforementioned biases.

The United Kingdom (UK) in an effort to reduce the spread of influenza in the general population has introduced nation-wide interventions in the form of vaccinations. Recognizing that children are key factors in the transmission of the influenza virus (Petrie et al. 2013), a pilot live attenuated influenza vaccine (LAIV) program has been launched in seven geographically discrete areas in

---

<sup>1</sup> ILI is typically defined as the presence of high fever together with cough or sore throat (Monto et al. 2000; Boivin et al. 2000).

England (Table 1) during the 2013/14 influenza season, with LAIV offered to school children aged from 4 to 11 years; this was in addition to offering vaccinations to all healthy children that were 2 or 3 years old. A report, led by one of our co-authors, quantified the impact of the school children LAIV campaign on a range of influenza indicators in pilot compared to non-pilot areas through traditional influenza surveillance schemes (Pebody et al. 2014). However, the sparse coverage of the surveillance system as well as the biases in the population that uses healthcare services, resulted into partially conclusive and not statistically significant outcomes.

In this work, we extend previous ILI modeling approaches from Internet content and propose a statistical framework for assessing the impact of a health intervention. To validate our methodology, we used UK’s 2013/14 pilot LAIV campaign as a case study. Our experimental setup involved the processing of millions of Twitter postings and Bing search queries geo-located in the target vaccinated locations, as well as a broader set of control locations across England. Firstly, we assessed the predictive capacity of various text regression models for inferring ILI rates, proposing a nonlinear method for performing this task based on the framework of Gaussian Processes (Rasmussen and Nickisch 2010), which improved predictions on our data set by a degree greater than 22% in terms of Mean Absolute Error (MAE) as compared to linear regularized regression methods such as the elastic-net (Zou and Hastie 2005). Then, we performed a statistical analysis, to evaluate the impact of the pilot LAIV program. The extracted impact estimates were in line with Public Health England’s (PHE)<sup>2</sup> findings (Pebody et al. 2014), providing both supplementary support for the success of the intervention, and validity evidence for our methodology.

## 2 Data sources

We used two user-generated data sources, namely search query logs from Microsoft’s Bing search engine and Twitter data. In the following paragraphs, we describe the process for extracting textual features from queries or tweets, as well as the additional components of the applied experimental process.

**Feature extraction.** We manually crafted a list of 36 textual markers (or  $n$ -grams) related to or expressing symptoms of ILI by browsing through related web pages (on Wikipedia or health-oriented websites). Then, using these markers as seeds, we extracted a set of frequent, co-occurring  $n$ -grams with  $n \leq 4$ , in a Twitter corpus of approx. 30 million tweets published between February and March 2014 and geo-located in the UK. This expanded the list of markers to a set of  $M = 205$   $n$ -grams (see Supplementary Material, Table S1), which formed the feature space in our experimental process. Overall the number of  $n$ -grams does not reach the quantity explored in previous studies (Ginsberg

---

<sup>2</sup> PHE is an executive agency for the Department of Health in England.

**Table 1** Areas participating in the LAIV program ( $v$ ) and control areas ( $c$ ) with their respective identifiers, population figures and geographical bounding box coordinates.

Areas	id	Population	SW <sup>a</sup>	NE <sup>b</sup>
Bury	$v_1$	186,527	-2.352, 53.550	-2.243, 53.645
Cumbria	$v_2$	498,070	-3.640, 54.042	-2.159, 55.189
Gateshead	$v_3$	199,998	-1.662, 54.914	-1.516, 54.971
Leicester City	$v_{4a}$	NA <sup>c</sup>	-1.216, 52.581	-1.046, 52.692
East Leicestershire	$v_{4b}$	661,575 <sup>d</sup>	-0.891, 52.392	-0.664, 52.978
Rutland	$v_{4c}$	37,606	-0.822, 52.525	-0.428, 52.760
London, Havering	$v_5$	242,080	0.138, 51.487	0.334, 51.632
London, Newham	$v_6$	318,227	-0.021, 51.498	0.098, 51.564
South East Essex	$v_7$	175,798 <sup>e</sup>	0.487, 51.494	1.032, 51.760
Brighton	$c_1$	278,112 <sup>f</sup>	-0.174, 50.807	-0.087, 50.870
Bristol	$c_2$	437,492	-3.118, 51.342	-2.510, 51.544
Cambridge	$c_3$	126,480	0.0774, 52.159	0.191, 52.238
Exeter	$c_4$	121,800	-3.687, 50.566	-3.367, 50.886
Leeds	$c_5$	761,481	-1.800, 53.698	-1.290, 53.946
Liverpool	$c_6$	470,780	-3.019, 53.312	-2.818, 53.475
Norwich	$c_7$	135,893	1.204, 52.555	1.541, 52.685
Nottingham	$c_8$	310,837	-1.247, 52.889	-1.086, 53.019
Plymouth	$c_9$	259,175	-4.303, 50.211	-3.983, 50.531
Sheffield	$c_{10}$	560,085	-1.801, 53.305	-1.325, 53.503
Southampton	$c_{11}$	242,141	-1.564, 50.743	-1.244, 51.063
York	$c_{12}$	202,433	-1.242, 53.799	-0.922, 54.119

<sup>a</sup> Longitude and latitude of the South-West edge of the bounding box.

<sup>b</sup> Longitude and latitude of the North-East edge of the bounding box.

<sup>c</sup> Figures for Leicester city alone, which is part of Leicestershire, were not included in (Office for National Statistics, Great Britain 2014a).

<sup>d</sup> This is a figure for the entire Leicestershire.

<sup>e</sup> This is a figure for Southend-on-Sea.

<sup>f</sup> Includes the town of Hove.

et al. 2009; Lampos and Cristianini 2012), although this choice was motivated by the fact that a small set of keywords is adequate for achieving a good predictive performance when modeling ILI from user-generated content published online (Culotta 2013).

**Geographic areas of interest.** We analyzed data that was either geolocated in England as a whole or in specific areas within England. Table 1 lists all the specific locations of interest, dividing them into two categories: the 7 vaccinated areas ( $v_i$ ) where the LAIV program was applied, and the selected 12 control areas ( $c_i$ ) which represent urban centers in England, with considerable population figures, that were distant from all vaccinated areas, and were spread across the geography of the country, to the extent possible. Each area is specified by a geographical bounding box defined by the longitude and latitude of its South-West and North-East edge points.

**User-generated web content.** To perform a more rigorous experimental approach, distinct data sets from two different web sources have been compiled. The first ( $\mathcal{T}$ ) consists of all Twitter posts (tweets) with geo-location enabled and pointing to the region of England from 02/05/2011 to 13/04/2014, i.e., 154 weeks in total. The total number of tweets involved is approx. 308 million, whereas the cumulative appearances of ILI-related  $n$ -grams is approx. 2.2 million. The vaccinated and control areas account for 5.8% and 12.6% of the entire content respectively. The second data set ( $\mathcal{B}$ ) consists of search queries on Microsoft’s web search engine, Bing, from 31/12/2012 to 13/04/2014 (67 weeks in total), geo-located in England. This data set has smaller temporal coverage as compared to Twitter data due to limitations in acquiring past search query logs. The number of queries in  $\mathcal{B}$  is significantly larger than the number of tweets in  $\mathcal{T}$ ;<sup>3</sup> 3.75% of the queries were geo-located in vaccinated areas, 12.53% in control areas, and flu related  $n$ -grams appeared in approx. 7.7 million queries. For all the considered  $n$ -grams (Supplementary Material, Table S1) we extracted their weekly frequency in England as well as in the designated areas of interest. We performed a more relaxed search, looking for content (tweets or search queries) that contains all the 1-gram blocks of an  $n$ -gram.

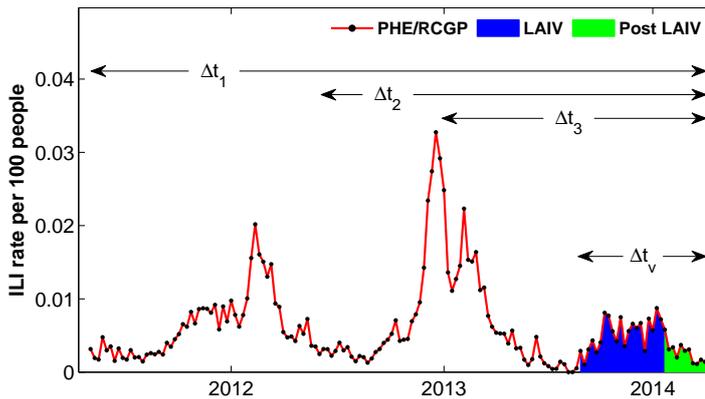
**Official health reports.** For the period covering data set  $\mathcal{T}$ , i.e., 02/05/2011 to 13/04/2014, PHE provided ILI estimates from patient data gathered by the Royal College of General Practitioners (RCGP)<sup>4</sup> in the UK. The estimates represent the number of GP consultations identified as ILI per 100 people for the geographical region of England and their temporal resolution is weekly (Fig. 1).

### 3 Estimating the impact of a healthcare intervention

The proposed methodology consists of two main steps: a) the modeling and prediction of a disease rate proxy from user-generated content as a regression problem, and b) the assessment of the health campaign using a statistical scheme that incorporates the regression models for the disease. Among well studied linear functions for text regression, we also propose a nonlinear technique, where different  $n$ -gram categories (sets of keywords of size  $n$ ) are captured by a different kernel function, as a better performing alternative (see Sections 3.1 and 3.2). The statistical framework for computing the impact of the intervention program is based on a method for evaluating the impact of printed advertisements (Lambert and Pregibon 2008); the method is described in detail in Section 3.3.

<sup>3</sup> The exact number cannot be disclosed as this is sensitive product information.

<sup>4</sup> RCGP has an established sentinel network of general practitioners in England and together with PHE publishes ILI rates on a weekly basis. Summaries of surveillance reports can be found at <http://www.gov.uk/sources-of-uk-flu-data-influenza-surveillance-in-the-uk> (accessed May 31, 2015).



**Fig. 1** Weekly ILI rates for England published by RCGP/PHE, covering three consecutive flu seasons (2011/12, 2012/13 and 2013/14).  $\Delta t$  labels denote the span of the time periods used in our experimental process. The end date for all periods is 13/04/2014, whereas  $\Delta t_1$  commences on 02/05/2011 (154 weeks),  $\Delta t_2$  on 04/06/2012 (97 weeks) and  $\Delta t_3$  on 31/12/2012 (67 weeks).  $\Delta t_v$  represents the effective time period of the LAIV program including a post-vaccination interval, up until the end of the flu season (green color); blue color is used to denote the actual vaccination period (September 2013 to January 2014).

### 3.1 Linear regression models for disease rate prediction

In this supervised learning setting, our observations  $\mathbf{X}$  consist of  $n$ -gram frequencies across time and the responses  $\mathbf{y}$  are formed by official health reports, both focused on a particular geographical region. Using  $N$  weekly time intervals and the  $M$   $n$ -gram features,  $\mathbf{X} \in \mathbb{R}^{N \times M}$  and  $\mathbf{y} \in \mathbb{R}^N$ . Each row of  $\mathbf{X}$  holds the normalized  $n$ -gram frequencies for a week in our data set. Normalization is performed by dividing the number of  $n$ -gram occurrences with the total number of tweets or search queries in the corpus for that week. Previous work performing text regression on social media content suggested the use of regularized linear regression schemes (Lampos and Cristianini 2010; Lampos et al. 2010). Here, we employ two well-studied regularization techniques, namely ridge regression (Hoerl and Kennard 1970) and the elastic-net (Zou and Hastie 2005), to obtain baseline performance rates.

The core element of regularized regression schemes is the minimization of the sum of squared errors between a linear transformation of the observations and the respective responses. In its simplest form, this is expressed by Ordinary Least Squares (OLS):

$$\operatorname{argmin}_{\mathbf{w}, \beta} \sum_{i=1}^N (\mathbf{x}_i \mathbf{w} + \beta - y_i)^2, \quad (1)$$

where  $\mathbf{w} \in \mathbb{R}^M$  and  $\beta \in \mathbb{R}$  denote the regression weights and intercept respectively, and  $y_i \in \mathbb{R}$  is the value of the response variable  $\mathbf{y}$  for a week  $i$ . The regularization of  $\mathbf{w}$  assists in resolving singularities which lead to ill-posed solutions when applying OLS. Broadly applied solutions suggest the penalization

of either the L2 norm (ridge regression) or the L1 norm (lasso) of  $\mathbf{w}$ . Ridge regression (Hoerl and Kennard 1970) is formulated as

$$\operatorname{argmin}_{\mathbf{w}, \beta} \left( \sum_{i=1}^N (\mathbf{x}_i \mathbf{w} + \beta - y_i)^2 + \kappa \sum_{j=1}^M w_j^2 \right), \quad (2)$$

where  $\kappa \in \mathbb{R}^+$  denotes the ridge regression’s regularization term. Lasso (Tibshirani 1996) encourages the derivation of a sparse solution, i.e., a  $\mathbf{w}$  with a number of zero weights, thereby performing feature selection. On a number of occasions, this sparse solution offers a better predictive accuracy than ridge regression (Hastie et al. 2009). However, models based on lasso are shown to be inconsistent in comparison to the true model, when collinear predictors are present in the data (Zhao and Yu 2006). Collinearities are expected in our task, since predictors are formed by time series of  $n$ -gram frequencies and semantically related  $n$ -grams will exhibit a degree of correlation. This is resolved by the elastic-net (Zou and Hastie 2005), an optimization function which merges L1 and L2 norm regularization, maintaining both positive properties of lasso and ridge regression. It is formulated as

$$\operatorname{argmin}_{\mathbf{w}, \beta} \left( \sum_{i=1}^N (\mathbf{x}_i \mathbf{w} + \beta - y_i)^2 + \lambda_1 \sum_{j=1}^M |w_j| + \lambda_2 \sum_{j=1}^M w_j^2 \right), \quad (3)$$

where  $\lambda_1, \lambda_2 \in \mathbb{R}^+$  are the L1 and L2 norm regularization parameters respectively. The Least Angle Regression (LAR) algorithm (Efron et al. 2004) provides an efficient way to compute an optimal lasso or elastic-net solution by exploring the entire regularization path, i.e., all the candidate values for the regularization parameter  $\lambda_1$  in Eq. 3. Parameter  $\lambda_2$  is estimated as a function of  $\lambda_1$ , where  $\lambda_2 = \lambda_1(1 - a)/(2a)$  (Zou and Hastie 2005); we set  $a = 0.5$  in our experiments, a common setting that obtains a 66.6%-33.3% regularization balance between the L1 and L2 norms respectively.

### 3.2 Disease rate prediction using Gaussian Processes

While the majority of methods for modeling infectious diseases are based on linear solvers (Ginsberg et al. 2009; Lampos et al. 2010; Culotta 2010), there is some evidence that nonlinear methods may be more suitable, especially when features are based on different  $n$ -gram lengths (Lampos 2012). Furthermore, recent studies in natural language processing (NLP) indicate that the usage of nonlinear methods, such as Gaussian Processes (GPs), in machine translation or text regression tasks improves performance, especially in cases where the feature space is not large (Lampos et al. 2014; Cohn et al. 2014). Motivated by these findings, we also considered a nonlinear model for disease prediction formed by a composite GP.

GPs can be defined as sets of random variables, any finite number of which have a multivariate Gaussian distribution (Rasmussen and Williams 2006).

In GP regression, for the inputs  $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^M$  (both expressing rows of the observation matrix  $\mathbf{X}$ ) we want to learn a function  $f: \mathbb{R}^M \rightarrow \mathbb{R}$  that is drawn from a  $\mathcal{GP}$  prior

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \quad (4)$$

where  $\mu(\mathbf{x})$  and  $k(\mathbf{x}, \mathbf{x}')$  denote the mean and covariance (or kernel) functions respectively; in our experiments we set  $\mu(\mathbf{x}) = 0$ . Evidently, the  $\mathcal{GP}$  kernel function is applied on pairs of input  $(\mathbf{x}, \mathbf{x}')$ . The aim is to construct a  $\mathcal{GP}$  that will apply a smooth function on the input space, based on the assumption that small changes in the response variable should also reflect on small changes in the observed term frequencies. A common covariance function that accommodates this is the isotropic Squared Exponential (SE), also known as the radial basis function or exponentiated quadratic kernel, and defined as

$$k_{\text{SE}}(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right), \quad (5)$$

where  $\sigma^2$  describes the overall level of variance and  $\ell$  is referred to as the characteristic length-scale parameter. Note that  $\ell$  is inversely proportional to the predictive relevancy of the feature category on which it is applied (high values of  $\ell$  indicate a low degree of relevance), and that  $\sigma^2$  is a scaling factor. An infinite sum of SE kernels with different length-scales results to another well studied covariance function, the Rational Quadratic (RQ) kernel (Rasmussen and Nickisch 2010). It is defined as

$$k_{\text{RQ}}(\mathbf{x}, \mathbf{x}') = \sigma^2 \left(1 + \frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\alpha\ell^2}\right)^{-\alpha}, \quad (6)$$

where  $\alpha$  is a parameter that determines the relative weighting between small and large-scale variations of input pairs. The RQ kernel can be used to model functions that are expected to vary smoothly across many length-scales. Based on empirical evidence, this kernel was shown to be more suitable for our prediction task.

In the GP framework predictions are conducted using Bayesian<sup>5</sup> integration, i.e.,

$$p(y_* | \mathbf{x}_*, \mathcal{O}) = \int_f p(y_* | \mathbf{x}_*, f) p(f | \mathcal{O}), \quad (7)$$

where  $y_*$  denotes the response variable,  $\mathcal{O}$  the training set and  $\mathbf{x}_*$  the current observation. Model training is performed by maximizing the log marginal likelihood  $p(\mathbf{y} | \mathcal{O})$  with respect to the hyper-parameters using gradient ascent.

Based on the property that the sum of covariance functions is also a valid covariance function (Rasmussen and Nickisch 2010), we model the different  $n$ -gram categories (1-grams, 2-grams, etc.) with a different RQ kernel. The reasoning behind this is the assumption that different  $n$ -gram categories may have varied usage patterns, requiring different parametrization for a proper

<sup>5</sup> Note that it is not strictly Bayesian in the sense that no prior is assumed for each one of the hyper-parameters in the  $\mathcal{GP}$  function.

modeling. Also as  $n$  increases, the  $n$ -gram categories are expected to have an increasing semantic value. The final covariance function, therefore, becomes

$$k(\mathbf{x}, \mathbf{x}') = \left( \sum_{n=1}^C k_{\text{RQ}}(\mathbf{g}_n, \mathbf{g}'_n) \right) + k_{\text{N}}(\mathbf{x}, \mathbf{x}'), \quad (8)$$

where  $\mathbf{g}_n$  is used to express the features of each  $n$ -gram category, i.e.,  $\mathbf{x} = \{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3, \mathbf{g}_4\}$ ,  $C$  is equal to the number of  $n$ -gram categories (in our experiments,  $C = 4$ ) and  $k_{\text{N}}(\mathbf{x}, \mathbf{x}') = \sigma_{\text{N}}^2 \times \delta(\mathbf{x}, \mathbf{x}')$  models noise ( $\delta$  being a Kronecker delta function). The summation of RQ kernels which are based on different sets of features can be seen as an exploration of the first order interactions of these feature families; more elaborate combinations of features could be studied by applying different types of covariance functions (e.g., the Matérn (Matérn 1986)) or an additive kernel (Duvenaud et al. 2011). An extended examination of these and other models is beyond the scope of this work.

Denoting the disease rate time series as  $\mathbf{y} = (y_1, \dots, y_N)$ , the GP regression objective is defined by the minimization of the following negative log-marginal likelihood function

$$\underset{\sigma_1, \dots, \sigma_C, \ell_1, \dots, \ell_C, \alpha_1, \dots, \alpha_C, \sigma_{\text{N}}}{\text{argmin}} \quad ((\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{K}^{-1}(\mathbf{y} - \boldsymbol{\mu}) + \log |\mathbf{K}|), \quad (9)$$

where  $\mathbf{K}$  holds the covariance function evaluations for all pairs of inputs, i.e.,  $(\mathbf{K})_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ , and  $\boldsymbol{\mu} = (\mu(\mathbf{x}_1), \dots, \mu(\mathbf{x}_N))$ . Based on a new observation  $\mathbf{x}_*$ , a prediction is conducted by computing the mean value of the posterior predictive distribution,  $E[y_* | \mathbf{y}, \mathcal{O}, \mathbf{x}_*]$  (Rasmussen and Williams 2006).

### 3.3 Intervention impact assessment

Conventional epidemiology typically assesses the impact of a healthcare intervention, such as a vaccination program, by comparing population disease rates in the affected (target) areas to the ones in non participating (control) areas (Pebody et al. 2014). However, a direct comparison of target and control areas may not always be applicable as comparable locations would need to be represented by very similar properties, such as geography, demographics and healthcare coverage. Identifying and quantifying such underlying characteristics is not something that is always possible or can be resolved in a straightforward manner. We, therefore, determine the control areas empirically, but in an automatic manner, as discussed below.

Firstly, we compute disease estimates ( $\mathbf{q}$ ) for all areas using our input observations (social media and search query data) and a text regression model. Ideally, for a target area  $v$  we wish to compare the disease rates during (and slightly after) the intervention program ( $\mathbf{q}_v$ ) with disease rates that would have occurred, had the program not taken place ( $\mathbf{q}_v^*$ ). Of course, the latter information,  $\mathbf{q}_v^*$ , cannot be observed, only estimated. To do so, we adopt a

methodology proposed for addressing a related task, i.e., measuring the effectiveness of offline (printed) advertisements using online information (Lambert and Pregibon 2008).

Consider a situation where, prior to the commencement of the intervention program, there exists a strong linear correlation between the estimated disease rates of areas that participate in the program ( $v$ ) and of areas that do not ( $c$ ). Then, we can learn a linear model that estimates the disease rates in  $v$  based on the disease rates in  $c$ . Hypothesizing that the geographical heterogeneity encapsulated in this relationship does not change during and after the campaign, we can subsequently use this model to estimate disease rates in the affected areas in the absence of an intervention ( $\mathbf{q}_v^*$ ).

More formally, we first test whether the inferred disease rates in a control location  $c$  for a period of  $\tau = \{t_1, \dots, t_N\}$  days before the beginning of the intervention ( $\mathbf{q}_c^\tau$ ) have a strong Pearson correlation,  $r(\mathbf{q}_v^\tau, \mathbf{q}_c^\tau)$ , with the respective inferred rates in a target area  $v$  ( $\mathbf{q}_v^\tau$ ). If this is true, then we can learn a linear function  $f(w, \beta) : \mathbb{R} \rightarrow \mathbb{R}$  that will map  $\mathbf{q}_c^\tau$  to  $\mathbf{q}_v^\tau$ :

$$\operatorname{argmin}_{w, \beta} \sum_{i=1}^N (q_c^{t_i} w + \beta - q_v^{t_i})^2, \quad (10)$$

where  $q_v^{t_i}$  and  $q_c^{t_i}$  denote weekly values for  $\mathbf{q}_v^\tau$  and  $\mathbf{q}_c^\tau$  respectively. By applying the previously learned function on  $\mathbf{q}_c^*$ , we can predict  $\mathbf{q}_v^*$  using

$$\mathbf{q}_v^* = \mathbf{q}_c^* w + \mathbf{b}, \quad (11)$$

where  $\mathbf{q}_c^*$  denotes the disease rates in the control areas during the intervention program and  $\mathbf{b}$  is a column vector with  $N$  replications of the bias term ( $\beta$ ).

Two metrics are used to quantify the difference between the actual estimated disease rates ( $\mathbf{q}_v$ ) and the projected ones had the campaign not taken place ( $\mathbf{q}_v^*$ ). The first metric,  $\delta_v$ , expresses the absolute difference in their mean values

$$\delta_v = \bar{\mathbf{q}}_v - \bar{\mathbf{q}}_v^*, \quad (12)$$

and the second one,  $\theta_v$ , measures their relative difference

$$\theta_v = \frac{\bar{\mathbf{q}}_v - \bar{\mathbf{q}}_v^*}{\bar{\mathbf{q}}_v^*}. \quad (13)$$

We refer to  $\theta_v$  as the impact percentage of the intervention. A successful campaign is expected to register significantly negative values for  $\delta_v$  and  $\theta_v$ .

Confidence intervals (CIs) for these metrics can be derived via bootstrap sampling (Efron and Tibshirani 1994). By sampling with replacement the regression's residuals  $\mathbf{q}_c^\tau - \hat{\mathbf{q}}_c^\tau$  in Eq. 10 (where  $\hat{\mathbf{q}}_c^\tau$  is the fit of the training data  $\mathbf{q}_v^\tau$ ) and then adding them back to  $\hat{\mathbf{q}}_c^\tau$ , we create bootstrapped estimates for the mapping function  $f(w, \beta)$ . We additionally sample with replacement  $\mathbf{q}_v$  and  $\mathbf{q}_c$ , before applying the bootstrapped function on them. This process is repeated 100,000 times and an equivalent number of estimates for  $\delta_v$  and  $\theta_v$

is computed. The CIs are derived by the .025 and .975 quantiles in the distribution of those estimates. Provided that the distribution of the bootstrap estimates is unimodal and symmetric, we assess an outcome as statistically significant, if its absolute value is higher than two standard deviations of the bootstrap estimates (similarly to [Lambert and Pregibon \(2008\)](#)).

## 4 Results

In the following sections, we apply the previously described framework to assess the UK’s pilot school children LAIV campaign based on user-generated Internet data. First, we evaluate the aforementioned regression methods that provide a proxy for ILI via the modeling of Bing and Twitter content geo-located in England. As ‘ground truth’ in these experiments, we use ILI rates (see Fig. 1) published by the RCGP/PHE. We then use the best performing regression model in the framework for estimating the impact of the vaccination campaign.

### 4.1 Predictive performance for ILI inference methods

We have applied a set of inference methods, starting from simple baselines (ridge regression) to more advanced regularized regression models (elastic-net), including a nonlinear function based on a composite GP. We evaluate our results by performing a stratified 10-fold cross validation, creating folds that maintain a similar sample distribution in the relatively short time-span covered by our input observations. To allow a better interpretation of the results, we used two standard performance metrics, the Pearson correlation coefficient ( $r$ ), which is not always indicative of the prediction accuracy, and the MAE between predictions ( $\hat{\mathbf{y}}$ ) and ‘ground truth’ ( $\mathbf{y}$ ). For  $N$  predictions of a single fold, MAE is defined as

$$\text{MAE}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|, \quad (14)$$

being expressed in the same units as the predictions. Then, the average  $r$  and MAE on the 10 folds are computed together with their corresponding standard deviations.

Given that the extracted tweets had a more extended temporal coverage compared to the search queries, we have performed experiments on the following data sets: (a) Twitter data for the period  $\Delta t_1 = 154$  weeks, from 02/05/2011 to 13/04/2014, a time period that encompasses three influenza seasons, (b) search query log data from Bing for the period  $\Delta t_2 = 67$  weeks, from 31/12/2012 to 13/04/2014, and (c) Twitter data for the same period  $\Delta t_3$ . All data sets are considering content geo-located in England and the respective time periods are depicted on Fig. 1. The latter data set (c) permits a better comparison between Twitter and Bing data.

**Table 2** Performance of ILI estimators for England under all investigated models and data sets ( $\mathcal{T}$ : Twitter,  $\mathcal{B}$ : Bing) based on a 10-fold cross validation.  $\mu(r)$  and  $\mu(\text{MAE})$  denote the average Pearson correlation and average MAE (the latter is multiplied by  $10^3$ ) between predictions and response data in the 10 folds; parentheses contain the standard deviation of the mean. Row result pairs with an asterisk (\*) have a statistically significant difference in their mean performance, whereas column result pairs with a dagger (†) do not.

	Ridge Regression		Elastic-Net		GP-kernel	
	$\mu(r)$	$\mu(\text{MAE}) \times 10^3$	$\mu(r)$	$\mu(\text{MAE}) \times 10^3$	$\mu(r)$	$\mu(\text{MAE}) \times 10^3$
$\mathcal{T}, \Delta t_1$	.640 (.112)	3.074 (.497)	.718 (.206)	2.828 (.809)*	.845 (.062)	2.196 (.477)*
$\mathcal{T}, \Delta t_3$	.698 (.181)	4.084 (.879)	.744 (.137)	3.198 (.137)*	.924 (.053)	1.999 (.763)*,†
$\mathcal{B}, \Delta t_3$	.814 (.103)	2.963 (.638)	.867 (.067)	2.564 (.677)*	<b>.952 (.041)</b>	<b>1.598 (.504)*,†</b>

Table 2 enumerates the derived performance figures. For all three data sets, the GP-kernel method performs best. Due to its larger time span, the experiment on Twitter data published during  $\Delta t_1$  provides a better picture for assessing the learning performance of each of the applied algorithms. There, the two dominant models, i.e., the elastic-net and the GP-kernel, have a statistically significant difference in their mean performance, as indicated by a two-sample  $t$ -test ( $p = .0471$ ); this statistically significant difference is replicated in all experiments ( $p < .005$ ) indicating that the GP model handles the ILI inference task better. Bing data provide a better inference performance as compared to Twitter data from the same time period ( $\mu(r) = .952$ ,  $\mu(\text{MAE}) = 1.598 \times 10^{-3}$ ), but in that case the difference in performance between the two sources is not statistically significant at the 5% level ( $p = 0.1876$ ). The usefulness of incorporating different  $n$ -gram categories and not just 1-grams has also been empirically verified (see Appendix B, Table 5). Experiments, where Bing and Twitter data were combined (by feature aggregation or different kernels), indicated a small performance drop. However, this cannot form a generalized conclusion as it may be a side effect of the data properties (format, time-span) we were able to work with. We leave the exploration of more advanced data combinations for future work.

#### 4.2 Assessing the impact of the LAIV campaign

Taking into account the results presented in the previous section, we rely on the best performing GP-kernel model for estimating an ILI proxy. For both Twitter and Bing, we have used ILI models trained on all data geo-located in England (time frames  $\Delta t_1$  and  $\Delta t_3$  apply respectively). After learning a generic model for England, we then use it to infer ILI rates in specific locations.<sup>6</sup>

To assess the impact of the LAIV campaign, we first need to identify control areas with estimated ILI rates that are strongly correlated to rates in the target vaccinated locations before the start of the LAIV program (Table 1 lists all the considered areas). As the strains of influenza virus may vary between distant time periods (Smith et al. 2004), invalidating our hypothesis for geographical

<sup>6</sup> This decision is also enforced by the lack of ground truth for specific locations.

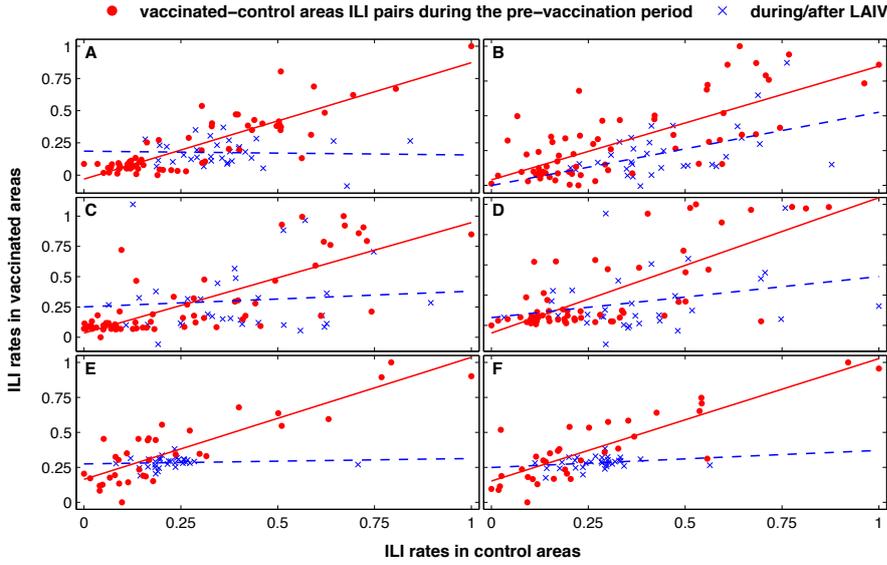
**Table 3** Statistically significant estimates of the LAIV program’s impact on the vaccinated areas using Twitter ( $\mathcal{T}$ ) or Bing ( $\mathcal{B}$ ) data. Column  $r(v, c)$  holds the top discovered Pearson correlations between the modeled ILI rates in vaccinated target areas ( $v$ ) and the corresponding controls ( $c$ ) during the designated pre-vaccination periods.  $\delta_v$  ( $\times 10^3$ ) and  $\theta_v$  denote the estimated mean difference and mean impact percentage on ILI rates respectively (parentheses include their bootstrap confidence intervals).

Data	Targets ( $v$ )	Controls ( $c$ )	$r(v, c)$	$\delta_v \times 10^3$	$\theta_v$ (%)
$\mathcal{T}$	all	$c_1 - c_3,$ $c_5 - c_8, c_{10}$	<b>.861</b>	<b>-2.503</b> (-4.110,-1.043)	<b>-32.772</b> (-47.428,-15.621)
$\mathcal{T}$	$v_5, v_6$	$c_1 - c_4, c_6,$ $c_7, c_{12}$	.738	-1.727 (-2.523,-0.942)	-30.453 (-41.751,-17.516)
$\mathcal{T}$	$v_2$	$c_1, c_3, c_4,$ $c_7 - c_9, c_{11}$	.769	-1.181 (-2.274,-0.094)	-21.060 (-37.136,-1.821)
$\mathcal{T}$	$v_6$	$c_1, c_3, c_4, c_6$	.738	-1.633 (-2.782,-0.521)	-30.436 (-46.742,-10.627)
$\mathcal{B}$	all	$c_1, c_2,$ $c_4 - c_7, c_{11}$	<b>.866</b>	<b>-1.929</b> (-3.249,-0.707)	<b>-21.705</b> (-32.120,-9.116)
$\mathcal{B}$	$v_5, v_6$	$c_4 - c_7, c_{11}$	.848	-2.811 (-4.073,-1.566)	-28.372 (-36.717,-17.943)
$\mathcal{B}$	$v_3$	$c_7$	.618	-3.737 (-6.908,-0.878)	-30.246 (-44.624,-9.174)

homogeneity across the considered flu seasons, we look for correlated areas in a pre-vaccination period that includes the previous flu season only (2012/2013). For Twitter data, this is from June, 2012 to August, 2013 (all inclusive), whereas for Bing data, given their smaller temporal coverage, the period was from January to August, 2013 (all inclusive). To determine the best control areas, an exhaustive search is performed comparing the correlation between vaccinated and control areas, for all individual areas and supersets of them.

Table 3 presents results of location pairs with an ILI proxy rate correlation of  $\geq .60$  during the pre-vaccination period, for which we have computed statistically significant impact estimates ( $\delta_v$  and  $\theta_v$ ), together with bootstrap confidence intervals (see Appendix B, Table 6 for all the results, including statistical significance metrics). The vaccinated areas or supersets of them include the London borough of Newham ( $v_6$ ), Cumbria ( $v_2$ ), Gateshead ( $v_3$ ), both London boroughs (Havering and Newham —  $v_5, v_6$ ) as well as a joint representation of all areas ( $v_1 - v_7$ ). The best correlations between vaccinated and control areas we were able to discover were: a)  $r = .866$  ( $p < .001$ ) for all vaccinated locations based on the Bing data, and b)  $r = .861$  ( $p < .001$ ) again for all vaccinated areas, but based on Twitter data. Note that in these two cases the optimal controls differed per data set, but had a substantial intersection of areas ( $c_1 - c_3, c_6, c_7$ ).

Fig. 2 depicts the linear relationships between the six most correlated location pairs of Table 3. To ease interpretation, the range of the axes has been normalized from 0 to 1. Red dots denote data pairs prior to the vaccination program and blue crosses denote pairs during and after the vaccination period (from October 2013 up until the end of the 2013/14 flu season, i.e., 13/04/2014). We observe that there is a linear, occasionally noisy, relationship between pairs of points prior to the vaccination, and between pairs of points during and after the vaccination. The slope of the best fit line is different for

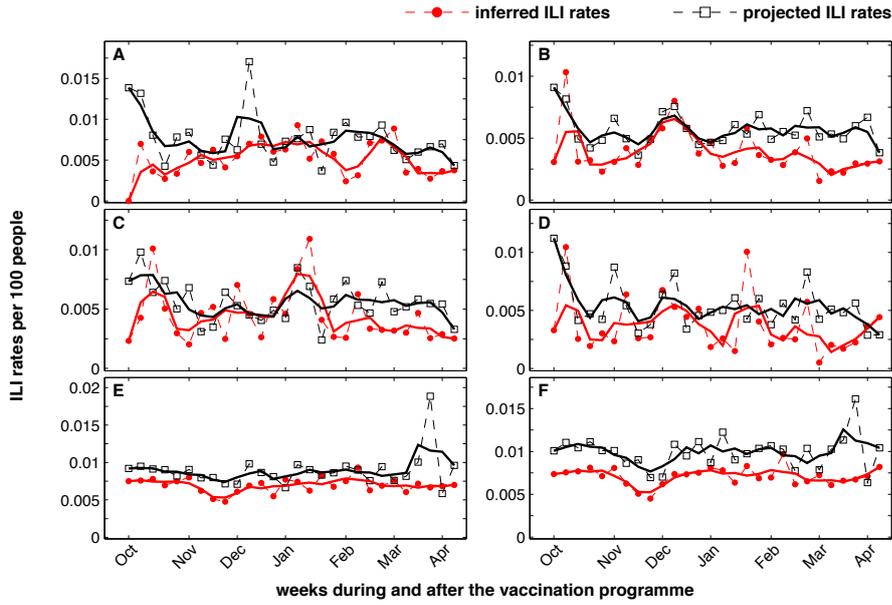


**Fig. 2** Linear relationship between the ILI rates in vaccinated areas and their respective controls during the pre-vaccination period (red dots) and during the LAIV program up until the end of the 2013/14 flu season (blue crosses). Axes are normalized from 0 to 1 to assist a better visualization across the cases and data sets; the red-solid and blue-dashed lines denote the least squares fits for the corresponding location pairs before and during the LAIV program respectively. **A:** All vaccinated areas with controls  $c_1 - c_3, c_5 - c_8$  and  $c_{10}$  ( $\mathcal{T}$ ). **B:** London areas with controls  $c_1 - c_4, c_6, c_7$  and  $c_{12}$  ( $\mathcal{T}$ ). **C:** Cumbria with controls  $c_1, c_3, c_4, c_7 - c_9$  and  $c_{11}$  ( $\mathcal{T}$ ). **D:** London borough of Newham with controls  $c_1, c_3, c_4$  and  $c_6$  ( $\mathcal{T}$ ). **E:** All vaccinated areas with controls  $c_1, c_2, c_4 - c_7$  and  $c_{11}$  ( $\mathcal{B}$ ). **F:** London areas with controls  $c_4 - c_7$  and  $c_{11}$  ( $\mathcal{B}$ ).

the two time periods. In particular, the slope during and after the vaccination period is consistently less than the slope before the vaccination, indicating that ILI rates in the target regions (y-axis) have reduced in comparison with the control regions (x-axis) during and after the vaccination.

The linear mappings between control and vaccinated areas before the vaccination are used to project ILI rates in the vaccinated areas during and slightly after the LAIV program. Fig. 3 depicts these estimates (same layout as in Fig. 2), showing a comparative plot of the proxy ILI rates (estimated using Twitter or Bing data) versus the projected ones; to allow for a better visual comparison, a smoothed time series is also displayed (3-point moving average). Referring to the moving average curves, we observe that it is almost always true that the projected ILI rates estimated from the control areas are higher than the proxy ILI rates estimated directly from Twitter or Bing. This indicates that the primary school children vaccination program may have assisted in the reduction of ILI in the pilot areas.

The time period used for evaluating the LAIV program includes the weeks starting from 30/09/2013 and ending at 13/04/2014 (28 weeks in total), i.e., the time frame covering the actual campaign (up to January, 2014) plus the



**Fig. 3** Modeled ILI rates inferred via user-generated content ( $q_v$ , red dots) in comparison with projected ILI rates ( $q_v^*$ , black squares) during the LAIV program and up until the end of the influenza season. The projection represents an estimation of the ILI rates that would have appeared, had the LAIV program not taken place. The solid lines (3-point moving average) represent the general trends of the actual data points (dashed lines) to allow for a better visual comparison. **A:** All vaccinated areas ( $\mathcal{T}$ ). **B:** London areas ( $\mathcal{T}$ ). **C:** Cumbria ( $\mathcal{T}$ ). **D:** London borough of Newham ( $\mathcal{T}$ ). **E:** All vaccinated areas ( $\mathcal{B}$ ). **F:** London areas ( $\mathcal{B}$ ).

weeks up until the end of the flu season (see Fig. 1). The bootstrap estimates for both impact metrics ( $\delta_t$  and  $\theta_t$ ) provide confidence intervals as well as a measure for testing the statistical significance of an outcome. Given that the distribution of the bootstrap estimates appears to be unimodal and symmetric (see Appendix B, Fig. 4), an outcome is considered as statistically significant, if it is smaller than two standard deviations of the bootstrap sample. The statistically significant impact estimates (Table 3) indicate a reduction of ILI rates, with impact percentages ranging from  $-21.06\%$  to  $-32.77\%$ . Interestingly, the estimated impact for the London areas is in a similar range for both Bing and Twitter data ( $-28.37\%$  to  $-30.45\%$ ).

#### 4.3 Sensitivity of impact estimates

Our analysis so far has been based on the linear relationship between vaccinated locations and only the top-correlated set of controls. To assess the sensitivity of our results to the choice of control regions, we repeated each impact estimation experiment for all control regions (sets of  $c_1$  through  $c_{12}$ ) found to have a correlation score (with a target area) greater or equal to 95%

**Table 4** Sensitivity assessment of LAIV campaign’s impact estimates (cases are aligned with Table 3). The mean values of the impact metrics using Twitter ( $\mathcal{T}$ ) or Bing ( $\mathcal{B}$ ) data are computed using the top hundred controls with a linear correlation greater or equal to 95% of the best correlation. Controls with non statistically significant estimates have been filtered out. Our sensitivity metric,  $\Delta\theta_v$ , denotes the percentage of difference between  $\mu(\theta_v)$  and the original  $\theta_v$  estimate (see Table 3).

Data set	Targets	# Controls	$\mu(r(v, c))$	$\mu(\delta_v) \times 10^3$	$\mu(\theta_v)$ (%)	$\Delta\theta_v$ (%)
$\mathcal{T}$	all	100	.841 (0.007)	-2.506 (0.234)	-32.740 (2.066)	<b>0.10</b>
$\mathcal{T}$	$v_5, v_6$	79	.703 (0.011)	-1.532 (0.148)	-27.918 (1.955)	8.32
$\mathcal{T}$	$v_2$	8	.744 (0.015)	-1.236 (0.111)	-21.793 (1.516)	<b>3.48</b>
$\mathcal{T}$	$v_6$	32	.705 (0.013)	-1.340 (0.218)	-26.277 (3.149)	13.66
$\mathcal{B}$	all	46	.854 (0.003)	-1.382 (0.369)	-16.417 (3.590)	24.36
$\mathcal{B}$	$v_5, v_6$	100	.841 (0.002)	-1.448 (0.212)	-16.899 (1.827)	40.44
$\mathcal{B}$	$v_3$	2	.607 (0.016)	-3.229 (0.719)	-27.120 (4.421)	10.34

of the best correlation. In the case where the number of controls exceeded 100, we used the top-100 correlated controls. Considering only statistically significant results, we computed the mean  $\delta_v$  and  $\theta_v$  (and their corresponding standard deviations) on the outcomes for all the applicable controls. We also measured the percentage of difference in  $\theta_v$  ( $\Delta\theta_v$ ) compared to the most highly correlated control (reported in Table 3) and used it as our sensitivity metric. Table 4 enumerates the derived averaged impact and sensitivity estimates, together with the number of applicable controls per case. Generally, we observe that results stemming from Twitter data are less sensitive (0.10% – 13.7%) to changes in control regions as compared to Bing data (10.3% – 40.3%). The most consistent estimate (from Table 3) is the one indicating a  $-32.77\%$  impact on the vaccinated areas as a whole based on Twitter data, with  $\Delta\theta_v$  equal to just 0.1%.

## 5 Related work

User-generated web content has been used to model infectious diseases, such as influenza-like illness (Milinovich et al. 2014). Coined as “infodemiology” (Eysenbach 2006), this research paradigm has been first applied on queries to the Yahoo engine (Polgreen et al. 2008). It became broadly known, after the launch of the Google Flu Trends (GFT) platform (Ginsberg et al. 2009). Both modeling attempts used simple variations of linear regression between the frequency of specific keywords (e.g., ‘flu’) or complete search queries (e.g., ‘how to reduce fever’) and ILI rates reported by syndromic surveillance. In the latter case, the feature selection process, i.e., deciding which queries to include in the predictive model, was based on a correlation analysis between query frequency and published ILI rates (Ginsberg et al. 2009). However, GFT has been criticized as in several occasions its publicly available outputs exhibited significant deviation from the official ILI rate reports (Cook et al. 2011; Olson et al. 2013; Lazer et al. 2014).

Research has also considered content coming from the social platform of Twitter as a publicly available alternative to access user-generated information. Regression models, either regularized (Lampos and Cristianini 2010; Lampos et al. 2010) or based on a smaller set of features (Culotta 2010), were used to infer ILI rates. Qualitative properties of the H1N1 pandemic in 2009 have been investigated through an analysis of tweets containing specific keywords (Chew and Eysenbach 2010) as well as a more generic modeling (Signorini et al. 2011); in the latter work support vector regression (Cristianini and Shawe-Taylor 2000) was used to estimate ILI rates. Bootstrapped regularized regression (Bach 2008) has been applied to make the feature selection process more robust (Lampos and Cristianini 2012); the same method has been applied to infer rainfall rates from tweets, indicating some generalization capabilities of those techniques. Furthermore, proof has been provided that for Twitter content a small set of keywords can provide an adequate prediction performance (Culotta 2013). Other studies, focused on unsupervised models that applied NLP methods in order to identify disease oriented tweets (Lamb et al. 2013) or automatically extract health concepts (Paul and Dredze 2014).

In this paper, we base our ILI modeling on previous findings, but apart from relying on a linear model, we also investigate the performance of a nonlinear multi-kernel GP (Rasmussen and Williams 2006). GPs have been applied in a number of fields, ranging from geography (Oliver and Webster 1990) to sports analytics (Miller et al. 2014). Recently, they were also used – as a better performing alternative – in NLP tasks such as the annotation modeling for machine translation (Cohn and Specia 2013), text regression (Lampos et al. 2014), and text classification (Preoțiuc-Pietro et al. 2015), where various multi-modal features were combined in one learning function. To the best of our knowledge, there has been no previous work aiming to model the impact of a health intervention through user-generated online content. This evaluation is usually conducted by an analysis of the various epidemiological surveillance outputs, if they are available (Pebody et al. 2014; Matsubara et al. 2014). The core methodology (and its statistical properties) on which we based our impact analysis has been proposed by Lambert and Pregibon (2008).

## 6 Discussion

We presented a statistical framework for transforming user-generated content published on web platforms to an assessment of the impact of a health-oriented intervention. As an intermediate step, we proposed a kernelized nonlinear GP regression model for learning disease rates from  $n$ -gram features. Assuming that an ILI model trained on a national level represents sufficiently smaller parts of the country, we used it as our ILI scoring tool throughout our experiments. Focusing on the theme of influenza vaccinations (Osterholm et al. 2012; Baguelin et al. 2012), especially after the H1N1 epidemic in 2009 (Smith et al. 2009), we measured the impact of a pilot primary school LAIV program introduced in England during the 2013/14 flu season. Our experimental re-

sults are in concordance with independent findings from traditional influenza surveillance measurements (Pebody et al. 2014). The derived vaccination impact assessments resulted in percentages (per vaccinated area or cumulatively) ranging from  $-21.06\%$  to  $-32.77\%$  based on the two data sources available.

The results from Twitter data, however, demonstrated less sensitivity across similar controls as compared to Bing data, suggesting a greater reliability. To that end, the most reliable impact estimate from the processed tweets regarded an aggregation of all vaccinated locations and was equal to  $-32.77\%$ . PHE’s own impact estimates looked at various end-points, comparing vaccinated to all non vaccinated areas, and ranged from  $-66\%$  based on sentinel surveillance ILI data to  $-24\%$  using laboratory confirmed influenza hospitalizations; albeit, these numbers represent different levels of severity or sensitivity, and notably none of these computations yielded statistical significance (Pebody et al. 2014). Thus, we cannot use them as a directly comparable metric, but mostly as a qualitative indication that the vaccination campaign is likely to have been effective.

A legitimate question is whether our analysis can yield one number that quantifies the intervention’s impact. This is a difficult undertaking given that no definite ground truth exists to allow for a proper verification. In addition, our estimations are based on models trained on syndromic surveillance data, which themselves may lack some specificity, hence not forming a solid gold standard. Interestingly, for the three distinct areas, where our method delivered statistically significant impact estimates based on Twitter data, i.e., Havering ( $-41.21\%$ ; see Appendix B, Table 6), Newham ( $-30.44\%$ ) and Cumbria ( $-21.06\%$ ), there exists a clear analogy with the reported level of vaccine uptake —  $63.8\%$ ,  $45.6\%$  and  $35.8\%$  respectively — as published by PHE (Pebody et al. 2014); a similar pattern is evident in the Bing data. This observation provides further support for the applied methodology.

Understanding the properties of the underlying population behind each disease surveillance metric is instrumental. First of all, the demographics (age, social class) of people who use a social media tool, a web search engine, or visit healthcare facilities may vary. For example, we know that  $51\%$  of the UK-based Twitter users are relatively young (15-34 years old), whereas only an  $11\%$  of them is 55 years or older (Ipsos MORI 2014). On the other hand, non-adults or the elderly are often responsible for the majority of doctor visits or hospital admissions (O’Hara and Caswell 2012). In addition, the relative volume of the aforementioned inputs also varies. We estimate that Twitter users in our experiments represent at most  $0.24\%$  of the UK population, whereas Bing has a larger penetration (approx.  $4.2\%$ ; see Appendix A for details). On the other side, in an effort to draw a comparable statistic, a 5-year study (2006-2011) on a household-level community cohort in England indicated that only  $17\%$  of the people with confirmed influenza are medically attended (Hayward et al. 2014). An other study estimated that 7,500 ( $0.01\%$ ) hospitalizations occurred due to the second and strongest wave of the 2009 H1N1 pandemic in England, when the percentage of the population being symptomatic was approx.  $2.7\%$  (Presanis et al. 2011). It is, therefore, a valid activity to seek complementary

ways, sensors or population samples for quantifying infectious diseases or the success of a healthcare intervention campaign.

Our method accesses a different segment of the population compared to traditional surveillance schemes, given that Internet users provide a potentially larger sample compared to the people seeking medical attention. The caveat is that user-generated content will be more noisy, thus, less reliable compared to doctor reports, and that it will entail certain biases. However, it can be advantageous, when data from traditional epidemiological sources are sparse, e.g., due to a mild influenza season, but also useful in other settings, where either traditional surveillance schemes are not well established or a more geographically focused signal is required. Despite the fact that our case study focuses on influenza, the proposed framework can potentially be adapted for estimating the impact of different health intervention scenarios. Future work should be focused on improving the various components of such frameworks as well as in the design of experimental settings that can provide a more rigorous evaluation ability.

**Acknowledgements** This work has been supported by the EPSRC grant EP/K031953/1 (“Early-Warning Sensing Systems for Infectious Diseases”). The authors would also like to acknowledge the Royal College of General Practitioners in the UK (in particular Simon de Lusignan) and Public Health England for providing ILI surveillance data.

## References

- Bach FR (2008) Bolasso: Model Consistent Lasso Estimation Through the Bootstrap. In: Proc. of the 25th International Conference on Machine Learning, pp. 33–40
- Baguelin M, Jit M, Miller E, Edmunds WJ (2012) Health and economic impact of the seasonal influenza vaccination programme in England. *Vaccine* 30(23):3459–3462
- Binder S, Levitt AM, Sacks JJ, Hughes JM (1999) Emerging Infectious Diseases: Public Health Issues for the 21st Century. *Science* 284(5418):1311–1313
- Boivin G, Hardy I, Tellier G, Maziade J (2000) Predicting Influenza Infections during Epidemics with Use of a Clinical Case Definition. *Clin Infect Dis* 31(5):1166–1169
- Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. *J Comp Sci* 2(1):1–8
- Briand S, Mounts A, Chamberland M (2011) Challenges of global surveillance during an influenza pandemic. *Public Health* 125(5):247–256
- Chew C, Eysenbach G (2010) Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS ONE* 5(11):e14,118
- Cohen ML (2000) Changing patterns of infectious disease. *Nature* 406(6797):762–767
- Cohn T, Specia L (2013) Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation. In: Proc. of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 32–42
- Cohn T, Preotiu-Pietro D, Lawrence N (2014) Gaussian Processes for Natural Language Processing. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics: Tutorials, pp. 1–3
- Cook S, Conrad C, Fowlkes AL, Mohebbi MH (2011) Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *PLoS ONE* 6(8):e23,610
- Cristianini N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press

- Culotta A (2010) Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In: Proc. of the 1st Workshop on Social Media Analytics, pp. 115–122
- Culotta A (2013) Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Lang Resour Eval* 47(1):217–238
- Duvenaud DK, Nickisch H, Rasmussen CE (2011) Additive Gaussian Processes. In: Advances in Neural Information Processing Systems 24, pp. 226–234
- Efron B, Tibshirani RJ (1994) An Introduction to the Bootstrap. CRC press
- Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least Angle Regression. *Ann Stat* 32(2):407–499
- Eysenbach G (2006) Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In: AMIA Annual Symposium Proceedings, vol 2006, pp. 244–248
- Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–1014
- Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning. Springer
- Hayward AC, Fragaszy EB, Birmingham A, Wang L, Copas A, Edmunds WJ, et al. (2014) Comparative community burden and severity of seasonal and pandemic influenza: results of the Flu Watch cohort study. *Lancet Respir Med* 2(6):445–454
- Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12:55–67
- Ipsos MORI (2014) MediaCT Tech Tracker Q1. Tech. Rep.
- Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, et al. (2008) Global trends in emerging infectious diseases. *Nature* 451(7181):990–993
- Lamb A, Paul MJ, Dredze M (2013) Separating Fact from Fear: Tracking Flu Infections on Twitter. In: Proc. of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies, pp. 789–795
- Lambert D, Pregibon D (2008) Online Effects of Offline Ads. In: Proc. of the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising, pp. 10–17
- Lampos V (2012) Detecting Events and Patterns in Large-Scale User Generated Textual Streams with Statistical Learning Methods. Ph.D. Thesis, University of Bristol
- Lampos V, Cristianini N (2010) Tracking the flu pandemic by monitoring the Social Web. In: Proc. of the 2nd International Workshop on Cognitive Information Processing, pp. 411–416
- Lampos V, Cristianini N (2012) Nowcasting Events from the Social Web with Statistical Learning. *ACM Trans Intell Syst Technol* 3(4):72:1–72:22
- Lampos V, De Bie T, Cristianini N (2010) Flu Detector: Tracking Epidemics on Twitter. In: Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 599–602
- Lampos V, Preoțiuc-Pietro D, Cohn T (2013) A user-centric model of voting intention from Social Media. In: Proc. of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 993–1003
- Lampos V, Aletras N, Preoțiuc-Pietro D, Cohn T (2014) Predicting and Characterising User Impact on Twitter. In: Proc. of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 405–413
- Lazer D, Kennedy R, King G, Vespignani A (2014) The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343(6176):1203–1205
- Leetaru K, Wang S, Cao G, Padmanabhan A, Shook E (2013) Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* 18(5)
- Matérn B (1986) Spatial Variation. Springer
- Matsubara Y, Sakurai Y, van Panhuis WG, Faloutsos C (2014) FUNNEL: Automatic Mining of Spatially Coevolving Epidemics. In: Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 105–114
- Milinovich GJ, Williams GM, Clements ACA, Hu W (2014) Internet-based surveillance systems for monitoring emerging infectious diseases. *Lancet Infect Dis* 14(2):160–168
- Miller A, Bornn L, Adams R, Goldsberry K (2014) Factorized Point Process Intensities: A Spatial Analysis of Professional Basketball. In: Proc. of the 31th International Conference on Machine Learning, pp. 235–243

- Monto A, Gravenstein S, Elliott M, Colopy M, Schweinle J (2000) Clinical signs and symptoms predicting influenza infection. *Arch Intern Med* 160(21):3243–3247
- Morens DM, Folkers GK, Fauci AS (2004) The challenge of emerging and re-emerging infectious diseases. *Nature* 430(6996):242–249
- O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In: Proc. of the 4th International AAAI Conference on Weblogs and Social Media, pp. 122–129
- Office for National Statistics, Great Britain (2013) Internet Access – Households and Individuals 2013. Tech. Rep.
- Office for National Statistics, Great Britain (2014a) Annual Mid-year Population Estimates. Tech. Rep.
- Office for National Statistics, Great Britain (2014b) Internet Access – Households and Individuals 2014. Tech. Rep.
- O'Hara B, Caswell K (2012) Health Status, Health Insurance, and Medical Services Utilization: 2010. *Curr Pop Rep* pp. 70–133
- Oliver MA, Webster R (1990) Kriging: a method of interpolation for geographical information systems. *Int J Geogr Inf Syst* 4(3):313–332
- Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L (2013) Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. *PLoS Comput Biol* 9(10):e1003256
- Osterholm MT, Kelley NS, Sommer A, Belongia EA (2012) Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis. *Lancet Infect Dis* 12(1):36–44
- Paul MJ, Dredze M (2014) Discovering Health Topics in Social Media Using Topic Models. *PLoS ONE* 9(8):e103408
- Pebody RG, Green HK, Andrews N, Zhao H, Boddington N, et al. (2014) 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. *Euro Surveill* 19(2)
- Petrie JG, Ohmit SE, Cowling BJ, Johnson E, Cross RT, et al. (2013) Influenza Transmission in a Cohort of Households with Children: 2010–2011. *PLoS ONE* 8(9):e75339
- Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA (2008) Using Internet Searches for Influenza Surveillance. *Clin Infect Dis* 47(11):1443–1448
- Preotiuc-Pietro D, Lamos V, Aletras N (2015) An analysis of the user occupational class through Twitter content. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics
- Presanis AM, Pebody RG, Paterson BJ, Tom BDM, Birrell PJ, et al. (2011) Changes in severity of 2009 pandemic A/H1N1 influenza in England: a Bayesian evidence synthesis. *BMJ* 343
- Rasmussen CE, Nickisch H (2010) Gaussian Processes for Machine Learning (GPML) Toolbox. *J Mach Learn Res* 11:3011–3015
- Rasmussen CE, Williams CKI (2006) Gaussian Processes for Machine Learning. MIT Press
- Reed C, Angulo FJ, Swerdlow DL, Lipsitch M, Meltzer MI, Jernigan D, Finelli L (2009) Estimates of the prevalence of pandemic (H1N1) 2009. *Emerg Infect Dis*
- Signorini A, Segre AM, Polgreen PM (2011) The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE* 6(5):e19467
- Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, et al. (2004) Mapping the antigenic and genetic evolution of influenza virus. *Science* 305(5682):371–376
- Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, et al. (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 459:1122–1125
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J Roy Stat Soc B Met* 58(1):267–288
- Zhao P, Yu B (2006) On Model Selection Consistency of Lasso. *J Mach Learn Res* 7:2541–2563
- Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J Roy Stat Soc B Met* 67(2):301–320

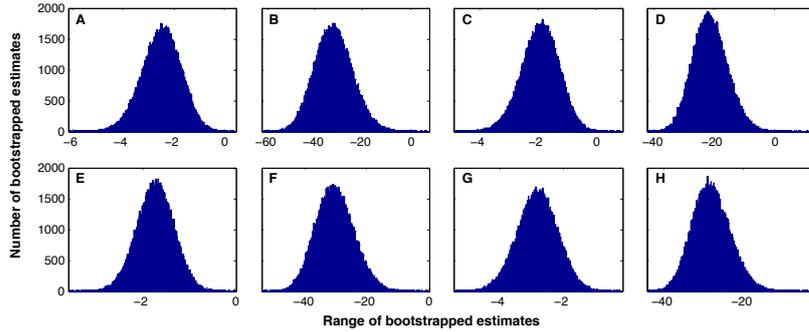
## Appendix A: Twitter and Bing populations in the UK

Knowing that Twitter users represented approx. 13%-15% of the UK population in the years 2012 to 2014 (Ipsos MORI 2014) and that only 1.6% of these users tend to enable the exact geo-location feature (Leetaru et al. 2013), we can estimate that the Twitter data in our experiments represents at most 0.24% of the population. Bing data have a larger penetration, estimated to be around 4.2% by combining the search tool's market share (approx. 5%) and the percentage of households with Internet access in the UK (Office for National Statistics, Great Britain 2013, 2014b).

## Appendix B: Supplemental outputs

**Table 5** Performance figures for ILI prediction under the GP-kernel model using Twitter data and different sets of  $n$ -gram features. The experimental setting corresponds to the results presented in the first row of Table 2.

	$\mu(r)$	$\mu(\text{MAE}) \times 10^3$
GP-kernel (1-grams)	.739 (.177)	2.405 (0.732)
GP-kernel (2-3-4-grams)	.836 (.068)	2.246 (0.389)



**Fig. 4** Histograms of bootstrap estimates for  $\delta_v$  ( $\times 10^3$ ) and  $\theta_v$  (%) on the vaccinates areas. **A, B:** All areas using  $\mathcal{T}(\delta_v, \theta_v)$ . **C, D:** All areas using  $\mathcal{B}(\delta_v, \theta_v)$ . **E, F:** London areas using  $\mathcal{T}(\delta_v, \theta_v)$ . **G, H:** London areas using  $\mathcal{B}(\delta_v, \theta_v)$ .

**Table 6** Estimates of the LAIV program’s impact on the vaccinated areas using Twitter ( $\mathcal{T}$ ) or Bing ( $\mathcal{B}$ ) data. Column  $r(v, c)$  holds the top discovered Pearson correlations between the modeled ILI rates in vaccinated target areas ( $v$ ) and the corresponding controls ( $c$ ) before the beginning of the vaccination program.  $\delta_v$  ( $\times 10^3$ ) and  $\theta_v$  denote the estimated mean difference and mean impact percentage on ILI rates respectively (parentheses include their bootstrap confidence intervals). In addition, columns  $\epsilon_\delta$  and  $\epsilon_\theta$  hold numbers equal to two standard deviations of the corresponding bootstrap estimates. Results indicating non statistically significant outcomes are in italics.

Data	$v$	$c$	$r(v, c)$	$\delta_v \times 10^3$	$\epsilon_\delta$	$\theta_v$ (%)	$\epsilon_\theta$
$\mathcal{T}$	all	$c_1 - c_3,$ $c_5 - c_8, c_{10}$	.861	-2.503 (-4.110,-1.043)	1.566	-32.772 (-47.428,-15.621)	16.176
$\mathcal{T}$	$v_5, v_6$	$c_1 - c_4, c_6,$ $c_7, c_{12}$	.738	-1.727 (-2.523,-0.942)	0.804	-30.453 (-41.751,-17.516)	12.382
$\mathcal{T}$	$v_1$	$c_1 - c_3, c_6,$ $c_{12}$	.278	-0.353 (-1.630,0.947)	<i>1.313</i>	-7.545 (-31.590,22.386)	<i>27.578</i>
$\mathcal{T}$	$v_2$	$c_1, c_3, c_4,$ $c_7 - c_9, c_{11}$	.769	-1.181 (-2.274,-0.094)	1.114	-21.060 (-37.136,-1.821)	18.085
$\mathcal{T}$	$v_3$	$c_1, c_4, c_{10},$ $c_{12}$	.614	-0.665 (-2.051,0.783)	<i>1.441</i>	-13.369 (-37.566,17.284)	<i>27.985</i>
$\mathcal{T}$	$v_4$	$c_1, c_2, c_5,$ $c_7 - c_9, c_{11}$	.725	-0.523 (-1.821,0.793)	<i>1.340</i>	-8.879 (-28.867,14.539)	<i>22.219</i>
$\mathcal{T}$	$v_5$	$c_2, c_4, c_{11},$ $c_{12}$	.420	-2.614 (-3.689,-1.514)	1.107	-41.211 (-54.125,-25.506)	14.621
$\mathcal{T}$	$v_6$	$c_1, c_3, c_4,$ $c_6$	.738	-1.633 (-2.782,-0.521)	1.149	-30.436 (-46.742,-10.627)	18.482
$\mathcal{T}$	$v_7$	$c_1, c_2,$ $c_5 - c_8, c_{12}$	.452	-0.431 (-1.664,0.846)	<i>1.278</i>	-7.238 (-26.343,15.179)	<i>21.114</i>
$\mathcal{B}$	all	$c_1, c_2,$ $c_4 - c_7, c_{11}$	.866	-1.929 (-3.249,-0.707)	1.294	-21.705 (-32.120,-9.116)	11.696
$\mathcal{B}$	$v_5, v_6$	$c_4 - c_7, c_{11}$	.848	-2.811 (-4.073,-1.568)	1.276	-28.372 (-36.717,-17.943)	9.556
$\mathcal{B}$	$v_1$	$c_3, c_{11}$	.427	-0.173 (-1.494,0.802)	<i>1.190</i>	-1.919 (-14.504,9.919)	<i>12.668</i>
$\mathcal{B}$	$v_2$	$c_1, c_7, c_8,$ $c_{11}$	.454	0.507 (-2.605,2.767)	<i>2.737</i>	0.051 (-23.018,31.590)	<i>28.083</i>
$\mathcal{B}$	$v_3$	$c_7$	.618	-3.737 (-6.908,-0.878)	3.066	-30.246 (-44.624,-9.174)	18.020
$\mathcal{B}$	$v_4$	$c_1, c_3, c_4$	.583	-1.079 (-4.684,2.702)	<i>3.754</i>	-9.954 (-35.680,30.285)	<i>33.903</i>
$\mathcal{B}$	$v_5$	$c_5, c_7, c_9,$ $c_{12}$	.512	-3.641 (-6.651,-1.021)	2.871	-33.380 (-48.302,-12.111)	18.447
$\mathcal{B}$	$v_6$	$c_1, c_4$	.754	-1.972 (-4.248,0.159)	<i>2.247</i>	-21.434 (-37.184,2.233)	20.131
$\mathcal{B}$	$v_7$	$c_1, c_4, c_8,$ $c_{11}, c_{12}$	.675	0.798 (-1.368,2.777)	<i>4.537</i>	7.136 (-27.149,56.222)	<i>42.840</i>