Assessing public health interventions using Web content

Extended Abstract

Vasileios Lampos
University College London
London, UK
v.lampos@ucl.ac.uk

ABSTRACT

Public health interventions are a fundamental tool for mitigating the spread of an infectious disease. However, it is not always possible to obtain a conclusive estimate for the impact of an intervention, especially in situations where the effects are fragmented in population parts that are under-represented within traditional public health surveillance schemes. To this end, online user activity can be used as a complementary sensor to establish alternative measures. Here, we provide a summary of our research on formulating statistical frameworks for assessing public health interventions based on data from social media and search engines (Lampos et al., 2015 [20]; Wagner et al., 2017 [37]). Our methodology has been applied in two real-world case studies: the 2013/14 and 2014/15 flu vaccination campaigns in England, where school-age children were vaccinated in a number of locations aiming to reduce the overall transmission of the virus. Disease models from online data combined with historical patterns of disease prevalence across different areas allowed us to quantify the impact of the intervention. In addition, a qualitative evaluation of our impact estimates demonstrated that they were in line with independent assessments from public health authorities.

1 INTRODUCTION

Data generated directly or indirectly by online users —also simply referred to as user-generated data (UGC)— can reveal a significant amount of information about their offline behaviour and status. In fact, many recent research efforts have leveraged social media content or search engine usage to address interesting questions in a number of domains, ranging from the Social Sciences [1, 8, 12] to Psychology [13, 23, 35] and Health [4, 9, 18].

Drawing our focus on health-oriented applications, one of the most prominent research tasks has been the derivation of Web-based syndromic surveillance models for infectious diseases. Modelling influenza-like illness (ILI) rates was the first successful example [6, 9, 17, 31], followed by other conditions [3, 10, 34], including mental health disorders [2, 4]. Criticisms regarding the accuracy of the original disease models [22, 27] have been resolved in follow-up studies by deploying more elaborate approaches [14, 19, 21]. One of the key motivations behind all the aforementioned works has been the potential of adopting UGC as a complementary sensor to doctor visits or hospitalisations, which are the main sources of information in traditional public health surveillance networks. An other important factor is that online data could provide access to the bottom of a disease pyramid, i.e. cases of infection present within specific demographies that are not well represented otherwise.

In this work, we go beyond disease modelling by proposing a statistical framework for assessing the impact of a health intervention (against an infectious disease) based on online information. Public health interventions, such as improved sanitation, immunisation programmes or, simply, the promotion of health literacy, assist in reducing the risk of various infections [5, 26]. However, the absence of routine evaluation systems for such interventions together with the general deficiencies of the existing disease surveillance schemes (e.g. under-represented parts of the populations), enables only partial assessments, especially in situations where interventions are targeting a seasonal disease that is not characterised by the magnitude of a pandemic.

We evaluate our algorithm against two real-world public health interventions. These are two vaccination campaigns against flu launched in England during 2013/14 (Phase A) and 2014/15 (Phase B). Live attenuated influenza vaccines (LAIV) were administered to school age children in various pilot locations, recognising that children are key factors in the transmission of the influenza virus in the general population [30]. In Phase A, the vaccine was offered to primary school children (4-11 years) only [28], whereas in Phase B it was also offered to children from secondary schools (11-13 years) as well as in an expanded set of locations [29].

Data from Microsoft's search engine, Bing, and the microblogging service of Twitter are used as the main observations for the proposed impact assessment framework. We deploy nonlinear supervised learning techniques using composite Gaussian Process kernels to model the time series of text frequencies in relation to disease rates in the population. We then utilise this disease model to uncover linear relationships between the disease rates in areas of interest during a time period prior to the intervention. Finally, we exploit this relationship to estimate a projection of disease rates to affected areas had the intervention not taken place. Our analysis yields interesting results, indicating that the intervention reduced ILI rates by more than 20% in Phase A locations and by approximately 17% in primary school areas in Phase B. Both estimates that are in agreement with independent assessments by Public Health England (PHE) [28, 29].[1]

2 METHODS

We briefly describe our approach for modelling disease rates from user-generated text and provide an overview of our statistical framework for assessing the impact of a public health intervention.

The estimation of disease rates from online textual information is formulated as a supervised learning task, $f : \mathbf{X} \in \mathbb{R}^{n \times m} \rightarrow \mathbf{y} \in \mathbb{R}^{n}$, where $\mathbf{X}$ represents the frequency of m textual terms over n time intervals, and $\mathbf{y}$ is the disease rate at the same time intervals (as

[1] They are in agreement in principle as direct comparisons are not valid.

Algorithm 1 Assessing the impact of a health intervention using online user-generated data [20]

Input: $\mathbf{X}$ (user-generated data), $\mathbf{y}$ (disease rates), $\mathcal{T}$ (target locations where the intervention was applied), C (control locations; no intervention), Δt_r (pre-intervention time period), Δt_α (intervention time period), $\rho_{\min}$ (Pearson correlation threshold)
Output: θ (percentage of impact), ϵ_θ (confidence intervals), S_θ (statistical significance)

Train a model f that estimates disease rates from user-generated data during Δt_r, $f\colon \mathbf{X} \rightarrow \mathbf{y}$
Derive all location subsets $\mathcal{T}_s$, C_s of $\mathcal{T}$, C respectively
Compute disease rates $\mathbf{y}_{\mathcal{T}_s}$, $\mathbf{y}_{C_s}$ during Δt_r using f
Compute all pairwise Pearson correlations, $\mathbf{r}_{\mathcal{T}_s, C_s}$, between the time series of $\mathbf{y}_{\mathcal{T}_s}$ and $\mathbf{y}_{C_s}$
for all pairs between $\mathcal{T}_s$ and C_s **do**
    if $r_{i,j} \geq \rho_{\min}$ **then** ▷ i, j refer to elements of $\mathcal{T}_s$, C_s respectively
        During Δt_r, train a model h_{ij} that estimates the disease rates of a subset of target locations from a subset of control locations, $h_{ij}\colon \mathbf{y}_{C_{sj}} \rightarrow \mathbf{y}_{\mathcal{T}_{si}}$
        Use f to estimate disease rates in C_{sj} during Δt_α based on user-generated data, $\mathbf{y}_c$
        Use h_{ij} and $\mathbf{y}_c$ to project disease rates in $\mathcal{T}_{si}$ from the ones in C_{sj} during Δt_α, $\mathbf{y}^c_\tau$
        Use f to estimate disease rates in $\mathcal{T}_{si}$ during Δt_α based on user-generated data, $\mathbf{y}_\tau$
        Estimate the impact of the intervention at $\mathcal{T}_{si}$ as $\theta_i = \frac{\mu(\mathbf{y}_\tau) - \mu(\mathbf{y}^c_\tau)}{\mu(\mathbf{y}^c_\tau)}$
        Use bootstrapped impact estimates, θ^b_i, to estimate confidence intervals for θ_i, ϵ_{θ_i} (.025 and .975 quantiles)
        if $|\theta_i| > 2\sigma(\theta^b_i)$ **then**
            Consider the impact estimate θ_i as statistically significant, $S_{\theta_i} = 1$
        else
            $S_{\theta_i} = 0$
        end if
    end if
end for

obtained by a public health authority). Provided that nonlinear models tend to outperform linear ones in text regression tasks [16, 19, 32], we composed and applied a Gaussian Process (GP) kernel for capturing the structure of our observations. GPs are defined as random variables any finite number of which have a multivariate Gaussian distribution. GP methods aim to learn a function $f\colon \mathbb{R}^m \rightarrow \mathbb{R}$ that is specified through a mean and a covariance (or kernel) function, i.e. $f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$, where $\mathbf{x}$ and $\mathbf{x}'$ (both $\in \mathbb{R}^m$) denote rows of the input matrix $\mathbf{X}$; for a detailed description of GPs, we refer the reader to [33]. By setting $\mu(\mathbf{x}) = 0$, a common practice in GP modelling, we just learn the hyper-parameters of the kernel. We define the following abstract kernel formulation:

$$k(\mathbf{x}, \mathbf{x}') = \left(\sum_{z=1}^{Z} k_\tau(\mathbf{g}_z, \mathbf{g}'_z) \right) + k_\nu(\mathbf{x}, \mathbf{x}'), \tag{1}$$

where k_τ can be any compatible GP kernel in the literature (we use the Rational Quadratic and the Matérn covariance functions in [20] and [37] respectively) that is applied on Z categories (or clusters) of textual features,[2] and k_ν captures noise.

Our methodology for assessing the intervention's impact, influenced by the work presented in [15], will utilise the above disease rate model. It is presented in detail in Alg. 1. Assume that there is a set of target areas $\mathcal{T}$, where the intervention is applied, and a set of control areas C, where the intervention has no effect. We firstly compute disease rate estimates for all areas as well as all possible subsets of them ($\mathcal{T}_s$, C_s) from UGC. Ideally, for a target area we wish to compare the disease rates during (and slightly after) the intervention with disease rates that would have occurred, had the intervention not taken place. Of course, the latter information can only be estimated. Focusing on target-control area pairs with strong linear correlations ($\geq \rho_{\min} = .6$) in historical disease rates prior to the intervention (Δt_r), we hypothesise that this relationship would have been maintained in the absence of an intervention. Therefore, we can learn a linear model (h) that estimates the disease rates in a target area based on the disease rates of a control area with data prior to the intervention. Then, we can use this model to project disease rates in a target area during the intervention period (Δt_α), but had the intervention not taken place. Finally, we can quantify the impact of the intervention by computing the relative percentage of difference (θ) between the actual estimated disease rates (from UGC) and the projected ones. Confidence intervals for θ can be derived via bootstrap sampling [7], and in particular by both sampling (with replacement) the linear regression's residuals (from h) as well as the input data. 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.

[2] We use $Z = 4$ categories of textual features based on the number of tokens (1 to 4).

3 RESULTS AND DISCUSSION

We first provide a brief overview of the data sets used in our analysis. We then summarise the outcomes of the intervention's impact assessment in both vaccination campaigns (Phase A and B). Finally, we propose potential directions for future research.

3.1 Data Sets

For the 2013/14 vaccination campaign (Phase A), we considered 7 target and 12 control areas (see Table 1 in [20]). We extracted

Table 1: Impact estimates (disease reduction rates) for super-sets of locations in England participating in vaccination programmes as estimated by online user-generated data. Estimates in bold were assessed as statistically significant.

Phase	Data Source	Target Locations ($\mathcal{T}$)	Num. of Control Locations (C)	$r(\mathcal{T}, C)$	Disease Reduction Rate % (θ)
A (2013/14)	Twitter	All locations	8	.86	**-32.72** (−47.43, −15.62)
	Bing	All locations	7	.87	**-21.71** (−32.12, −9.12)
B (2014/15)	Twitter	All locations	10	.89	−4.51 (−25.72, 22.61)
	Twitter	Primary school cohort	8	.71	**-16.97** (−30.09, −2.42)
	Twitter	Secondary school cohort	7	.83	1.41 (−19.40, 28.40)
	Twitter	Primary & secondary school cohort	7	.84	−0.30 (−16.71, 19.36)

308 million tweets (May, 2011 to April, 2014), 2.2 million of which contained flu-related n-grams.[3] We additionally obtained search query data (December, 2012 to April, 2014) for a smaller time period due to user privacy regulations, which contained approx. 7.7 million flu-related queries. As the campaign expanded in 2014/15 to include more locations (Phase B) and different school-age children groups, the number of target locations increased to 17 (6 primary, 7 secondary, and 4 primary and secondary school cohorts), and 16 control areas were deployed (see Table 1 in [37]). For this period, we extracted 520 million tweets geolocated in England (August, 2011 to August, 2015). This analysis did not use any search engine data. Historical ILI rates at a national level for England were obtained from the Royal College of General Practitioners, representing the number of ILI cases per 100,000 people from 2011 to 2015.

3.2 Intervention Impact Assessment

A GP, as described in Section 2, was used for modelling ILI rates from UGC since it outperformed linear alternatives, namely ridge regression [11] and elastic net [39]. Using a 10-fold cross validation, the mean absolute error (MAE) for the Twitter-based model during Phase A was equal to 2.2 (per 100,000 people) with an average Pearson correlation of $r = .85$, whereas the model used in Phase B (trained and tested on more data) resulted to a MAE of 2.4 and $r = .84$. The model trained on Bing data (Phase A) outperformed other models on average (MAE = 1.6, $r = .95$), but at the same time was tested on a significantly shorter time span.[4]

To assess the impact of the LAIV campaign, we first needed to identify control areas with estimated ILI rates that were strongly correlated to rates in the target vaccinated locations before the start of the intervention. In doing so for Phase A (2013/14), we looked for correlated areas in a pre-vaccination period that included the previous flu season only (2012/13). The reason for this was that the strains of influenza virus may vary between distant time periods [36] and thus, disease rates may be non homogeneous. For Phase B (2014/15), however, we could not anymore use the previous flu season to establish relationships, given that the Phase A campaign had already violated the assumed geographical homogeneity for 2013/14. Thus, we resided to using the period 2011/13[5] based on the fact that the circulated flu strains were not characterised by any significant anomalies. Nevertheless, that resulted in less robust estimates as indicated by our bootstrap sampling analysis (which yielded many of them as not statistically significant) and, taking into account the one-year gap between training and applying, perhaps less accurate projections as well.

A summary of the overall impact assessments is provided in Table 1, where outcomes in bold are statistically significant. During Phase A, both data sets (Twitter and Bing) point to significant reductions of disease rates, i.e. from −21.06% (Bing) to −32.77% (Twitter) on average. A subsequent sensitivity analysis (see Table 4 in [20]), where more than one control areas were used to project disease rates indicated that results from Twitter were generally more robust, with the overall impact estimate (−32.77%) being the most consistent one. PHE's own impact estimates compared vaccinated to all non vaccinated areas, and ranged from −66% based on sentinel surveillance ILI data to −24% using laboratory confirmed influenza hospitalisations. Note though that these numbers represent different levels of severity or sensitivity, and notably none of these computations was statistically significant [28]. As a further evaluation point, we observed an analogy between the actual level of vaccine uptake and the estimated impact from our end for a number of areas.

In Phase B, our analysis indicated that areas where primary school children were vaccinated benefited the most with an estimated θ of −16.97%. However, for the current implementation of the secondary school only vaccination programme, there was no clear evidence of any population wide effect. Both these conclusions are in line with findings of previous studies and complement traditional surveillance sources in exhibiting community wide effects of the LAIV pilot campaign [28, 29].

3.3 Future Work

Our approach faces common limitations of research efforts based on unstructured user-generated text. Better methods that automate the semantic interpretation of language can be deployed to derive more accurate results. In fact, in follow-up works, we have proposed techniques that are capable of combining the text statistics

[3] We used approximately 200 n-grams, listed in the supplementary material of [20].
[4] A more detailed performance evaluation is provided in Section 4.1 of [20].
[5] Includes two flu seasons from August, 2011 to August, 2013.

(e.g. frequency time series) with a word embedding representation [21, 24, 25, 38]. A further, perhaps more significant limitation, is that the entirety of this work relies on the existence of ground truth. Knowing historical disease rates is essential in order to train a disease model from UGC. However, this may not be possible for places with less established healthcare systems or for new infectious diseases. In addition, even when syndromic surveillance can provide estimates for the prevalence of a disease, it is very likely that these will incorporate demographic biases, carrying them over to any supervised model. Thus, there is a necessity to establish unsupervised disease indicators from UGC. This is a harder problem as it will be difficult to evaluate solutions and one will need to account for the specific demographic biases of the online users in order to produce any viable conclusion. Nevertheless, ongoing work will focus on resolving these issues as well as investigating the framework's applicability in assessing different types of a public health intervention.

ACKNOWLEDGMENTS

This work presented in this extended abstract has been supported by the grant EP/K031953/1 (EPSRC, "i-sense").

REFERENCES

[1] E. Bakshy, S. Messing, and L. A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.
[2] A. Benton, M. Mitchell, and D. Hovy. 2017. Multitask Learning for Mental Health Conditions with Limited Social Media Data. In *Proc. of EACL '17*. 152–162.
[3] E. H. Chan, V. Sahai, C. Conrad, and J. S. Brownstein. 2011. Using Web Search Query Data to Monitor Dengue Epidemics: A New Model for Neglected Tropical Disease Surveillance. *PLOS Negl. Trop. Dis.* 5, 5 (2011), e1206.
[4] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. 2013. Predicting Depression via Social Media. In *Proc. of ICWSM '13*. 128–137.
[5] M. L. Cohen. 2000. Changing patterns of infectious disease. *Nature* 406, 6797 (2000), 762–767.
[6] A. Culotta. 2010. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In *Proc. of the Workshop on Social Media Analytics*. 115–122.
[7] B. Efron and R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC press.
[8] H. Gil de Zúñiga, N. Jung, and S. Valenzuela. 2012. Social Media Use for News and Individuals' Social Capital, Civic Engagement and Political Participation. *J. Comput. Mediat. Commun.* 17, 3 (2012), 319–336.
[9] J. Ginsberg, M. H. Mohebbi, R. S. Patel, et al. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.
[10] J. Gomide, A. Veloso, W. Meira, Jr., V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira. 2011. Dengue Surveillance Based on a Computational Model of Spatio-temporal Locality of Twitter. In *Proc. of WebSci '11*. 1–8.
[11] A. E. Hoerl and R. W. Kennard. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12 (1970), 55–67.
[12] M. Kosinski, D. Stillwell, and T. Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci.* 110, 15 (2013), 5802–5805.
[13] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci.* 111, 24 (2014), 8788–8790.
[14] A. Lamb, M. J. Paul, and M. Dredze. 2013. Separating Fact from Fear: Tracking Flu Infections on Twitter. In *Proc. of NAACL '13*. 789–795.
[15] D. Lambert and D. Pregibon. 2008. Online Effects of Offline Ads. In *Proc. of the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising*. 10–17.
[16] V. Lampos, N. Aletras, D. Preoţiuc-Pietro, and T. Cohn. 2014. Predicting and Characterising User Impact on Twitter. In *Proc. of EACL '14*. 405–413.
[17] V. Lampos and N. Cristianini. 2010. Tracking the flu pandemic by monitoring the Social Web. In *Proc. of CIP '10*. 411–416.
[18] V. Lampos and N. Cristianini. 2012. Nowcasting Events from the Social Web with Statistical Learning. *ACM Trans. Intell. Syst. Technol.* 3, 4 (2012), 1–22.
[19] V. Lampos, A. C. Miller, S. Crossan, and C. Stefansen. 2015. Advances in nowcasting influenza-like illness rates using search query logs. *Sci. Rep.* 5, 12760 (2015).
[20] V. Lampos, E. Yom-Tov, R. Pebody, and I. J. Cox. 2015. Assessing the impact of a health intervention via user-generated Internet content. *Data Min. Knowl. Discov.* 29, 5 (2015), 1434–1457.
[21] V. Lampos, B. Zou, and I. J. Cox. 2017. Enhancing Feature Selection Using Word Embeddings: The Case of Flu Surveillance. In *Proc. of WWW '17*. 695–704.
[22] D. Lazer, R. Kennedy, G. King, and A. Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343, 6176 (2014), 1203–1205.
[23] A. M. Manago, T. Taylor, and P. M. Greenfield. 2012. Me and my 400 friends: The anatomy of college students' Facebook networks, their communication patterns, and well-being. *Dev. Psychol.* 48, 2 (2012), 369–380.
[24] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proc. of the ICLR, Workshop Track*. 1–12.
[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in NIPS 26*. 3111–3119.
[26] D. Nutbeam. 2000. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. *Health Promot. Int.* 15, 3 (2000), 259–267.
[27] D. R. Olson, K. J. Konty, M. Paladini, et al. 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 (2013), e1003256.
[28] R. Pebody 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. *Eurosurveillance* 19, 22 (2014), 20823.
[29] R. Pebody et al. 2015. 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. *Eurosurveillance* 20, 39 (2015), 30029.
[30] J. G. Petrie et al. 2013. Influenza Transmission in a Cohort of Households with Children: 2010-2011. *PLoS ONE* 8, 9 (2013), e75339.
[31] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. 2008. Using Internet Searches for Influenza Surveillance. *Clin. Infect. Dis.* 47, 11 (2008), 1443–1448.
[32] D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras. 2015. Studying User Income through Language, Behaviour and Affect in Social Media. *PLoS ONE* 10, 9 (2015), e0138717.
[33] C. E. Rasmussen and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press.
[34] F. Rohart, G. J. Milinovich, S. M. R. Avril, K.-A. Lê Cao, S. Tong, and W. Hu. 2016. Disease surveillance based on Internet-based linear models: an Australian case study of previously unmodeled infection diseases. *Sci. Rep.* 6, 38522 (2016).
[35] H. A. Schwartz et al. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE* 8, 9 (2013), e73791.
[36] D. J. Smith et al. 2004. Mapping the antigenic and genetic evolution of influenza virus. *Science* 305, 5682 (2004), 371–376.
[37] M. Wagner, V. Lampos, E. Yom-Tov, R. Pebody, and I. J. Cox. 2017. Estimating the Population Impact of a New Pediatric Influenza Vaccination Program in England Using Social Media Content. *J. Med. Internet Res.* 19, 12 (2017), e416.
[38] B. Zou, V. Lampos, R. Gorton, and I. J. Cox. 2016. On Infectious Intestinal Disease Surveillance Using Social Media Content. In *Proc. of the 6th International Conference on Digital Health*. 157–161.
[39] H. Zou and T. Hastie. 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* 67, 2 (2005), 301–320.