Institution: University College London		
Unit of Assessment: 11 – Computer Science and Informatics		
Title of case study: Modelling infectious diseases using web search data		
Period when the underpinning research was undertaken: May 2014 – December 2020		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s):	**Role(s) (e.g. job title):**	**Period(s) employed by submitting HEI:**
Dr. Vasileios Lampos Prof. Ingemar J. Cox	Principal Research Fellow Professor	2013 – Present 2002 – Present
Period when the claimed impact occurred: September 2016 – December 2020		
Is this case study continued from a case study submitted in 2014? N		

1. Summary of the impact (indicative maximum 100 words)

The impact of UCL's research on digital epidemiology based on Web search and Twitter data has (i) contributed to the introduction of a national influenza vaccine for children (approximately 5,000,000 per annum), which is estimated to reduce prevalence of influenza in the general population by 20%; (ii) been adopted by PHE as part of the weekly influenza reports, which are used to determine the start and duration of the annual influenza epidemic and hence, underpinned the recommendation for the commencement of prescribing antivirals to those at risk; (iii) been incorporated in PHE's publicly available COVID-19 surveillance reports, which informed the decision making process with regards to COVID19 national policy; and (iv) supported NHS UTLA authorities' regional, tiered response to the pandemic distribution.

2. Underpinning research (indicative maximum 500 words)

Between July 2014 and February 2015, Lampos led a research project at UCL, in collaboration with Google, to help improve machine learning methods for estimating influenza-like illness (ILI) rates from Web search data (**R1**). The UCL-led research was the first to show why Google Flu Trends over or under-estimated ILI rates. The paper (**R1**) went on to propose nonlinear solutions for this machine learning task given the evident nonlinearities in the data. At the time, the proposed solutions provided state-of-the-art performance.

Lampos and Cox subsequently improved feature (search query) selection within this context by using developments in statistical natural language processing (**R2**). That increased the accuracy of ILI rate estimates by between 12% and 28%.

In collaboration with Public Health England (PHE), Lampos and Cox assessed the accuracy of these models with respect to a number of epidemiological indicators, for example, magnitude and date of peak influenza prevalence, rather than common regression accuracy metrics such as mean square error (**R3**). This study corroborated their previous findings, reaffirming the potential value of using such approaches as complementary syndromic surveillance indicators.

To mitigate against situations where either sparse data is available (eg: due to an outage or inaccuracy of a syndromic surveillance system) or only a few training examples exist (eg: due to the lack of a comprehensive way to monitor an infectious disease during previous circulations), Lampos and Cox led a research project that proposed multi-task learning approaches for modelling ILI from web search data at national and subnational level (**R4**). This was also the first attempt to train models for two different countries jointly. Their models were able to improve accuracy by up to 40% when considering a 1-year long training period.

Lampos and Cox also led the development of a transfer learning approach for mapping a model for ILI based on web searches from one country, where historical ILI rates are available, to a country that has little or no such data, ie, where no comprehensive health surveillance system is in place, such as in low- and middle-income countries (LMIC) (**R5**). A variation of this approach was deployed during the first wave of the COVID-19 pandemic to map models from Italy, which had been exposed to a high circulation of COVID-19, to other countries that were in earlier phases of their local epidemics.

Finally, in collaboration with PHE and Microsoft Research, Lampos and Cox led a research project to estimate, for the first time, the effectiveness of a flu vaccination programme using social media and Web search data (**R6**). PHE initiated a pilot live attenuated influenza vaccine (LAIV) programme for school age children in seven geographically discrete areas in England during the 2013/14 influenza season. An analysis based on data from conventional syndromic surveillance systems did not yield statistically significant outcomes for the impact of this vaccination programme. Lampos and Cox then constructed models for the prevalence of ILI based on social media and web search activity in pilot (vaccinated) and control (unvaccinated) regions. The control regions were subsequently used to predict the prevalence of influenza in the pilot regions in the absence of the vaccine. The difference between this prediction and the estimated ILI prevalence in pilot regions was used to infer vaccine effectiveness. Estimates of effectiveness were strongly positive and statistically significant, corroborating the non-statistically significant indicators of prior work. This analysis was repeated for the 2014/15 flu season, yielding similar outcomes.

3. References to the research (indicative maximum of six references)

R1. **Lampos V**, Miller A, Crossan S, Stefansen C (2015). Advances in nowcasting influenza-like illness rates using search query logs. *Nature Scientific Reports* 5 (12760). doi: 10.1038/srep12760

R2. **Lampos V**, Zou B, **Cox IJ** (2017). Enhancing feature selection using word embeddings: The case of flu surveillance. Proc. of the 26th International World Wide Web Conference (WWW '17), pp. 695-704. doi: 10.1145/3038912.3052622

R3. Wagner M, **Lampos V**, **Cox IJ**, Pebody R (2018). The added value of online user-generated content in traditional methods for influenza surveillance. *Nature Scientific Reports* 8 (13963). doi: 10.1038/s41598-018-32029-6

R4. Zou B, **Lampos V**, **Cox IJ** (2018). Multi-task learning improves disease models from Web search. Proc. of the 27th International World Wide Web Conference (WWW '18), pp. 87-96. doi: 10.1145/3178876.3186050

R5. Zou B, **Lampos V**, **Cox IJ** (2019). Transfer learning for unsupervised influenza-like illness models from online search data. Proc. of the 28th International Web Conference (WWW '19), pp. 2505-2516. doi: 10.1145/3308558.3313477

R6. **Lampos V**, Yom-Tov E, Pebody R, **Cox IJ** (2015). Assessing the impact of a health intervention via user-generated Internet content. *Data Mining and Knowledge Discovery* 29 (5), pp. 1434-1457. doi: 10.1007/s10618-015-0427-9

4. Details of the impact (indicative maximum 750 words)

Prior to COVID-19, the UK Government Cabinet Office Risk Register identified pandemic influenza as highly likely to occur and to have the greatest adverse impact on the country. Lampos and Cox led the development of an online tool, *Flu Detector* (also known as "i-senseFlu") that displays estimates of flu rates for England based on web search data on a daily basis. The output from *Flu Detector* was incorporated in PHE's suite of syndromic surveillance methods during the 2017/18 influenza season for the first time and has been used by PHE in all subsequent years. To the best of our knowledge, this is the first system of its kind that has been formally adopted by a national health agency worldwide. Estimates from this tool have been included in PHE's weekly reports on influenza. Data in these reports are used to determine the timing and duration of an influenza epidemic and associated health recommendations, eg commencement of prescription antiviral drugs to the elderly. In a paper jointly authored with

PHE, Lampos and Cox demonstrated that surveillance based on Web searches could give an earlier indicator (one to two weeks) of the onset of an influenza epidemic or pandemic (**R5**) compared to a traditional network of sentinel doctors (coordinated by the Royal College of General Practitioners) that report the fraction of patients presenting at practices with influenza-like illness.

According to PHE, "*(Flu Detector) provides a vital early indicator of changes in influenza activity in the community. This was vital, for example, in the 2019-20 influenza season when activity began earlier than usual and Flu Detector was one of the first systems to detect this. This resulted in an early Chief Medical Officer (CMO) instruction to allow prescribing of antivirals in the community*" (**S1**).

Children are a major vector (spreader) of influenza in the community. PHE's models of the spread of influenza suggested that vaccination of children would lead to about a 20% reduction in influenza in the (unvaccinated) community. In collaboration with PHE, Lampos and Cox's assessment in (**R2**) supported this hypothesis. A WHO expert now leading the European response to COVID-19 and formerly leading the Influenza and Other Respiratory Pathogens team at PHE writes: "*Our traditional surveillance metrics and these novel indicators based on non-traditional epidemiological data were used as evidence for introducing an influenza vaccination programme across all primary schools in the UK for years 2-7. Influenza vaccinations for children became a national policy in 2015/16. NHS England currently recommends and offers a free influenza vaccine for all children between 2 and 7 years of age (over 5,000,000 children).*" (**S2**) PHE has estimated that this reduces the prevalence of influenza in the general population (not just vaccinated children) by 20% (**S3**).

During the first wave of the COVID-19 epidemic in the UK, PHE requested that the UCL team construct a model for COVID-19 based on their previous work on influenza. Lampos led the development of this model which exploits many of the ideas developed for influenza surveillance together with several novel and significant differences. A Consultant Epidemiologist at Public Health England (PHE) and lead of the COVID-19 surveillance cell writes that "*This system … provided an essential surveillance system for early indication of changes in COVID-19 activity. This was particularly notable at the start of the pandemic when the COVID-19 Google search surveillance system gave us one of our earliest indicators that the national lockdown was successfully reducing COVID-19 activity*" (**S1**).

The estimates of this model are sent to PHE on a weekly basis and are included in their publicly available COVID-19 surveillance report. The former digital surveillance lead for COVID-19 at PHE, writes these reports were "*used by policy makers at the national level to make decisions on outbreak management policy. In addition to appearing in a written form, the data is presented and discussed in a range of national level situational reports meetings attended by senior staff from Public Health England and the Department of Health and has undoubtedly informed the decision making process with regards to COVID19 national policy*" (**S4**).

In addition, the data provided by Lampos and Cox in collaboration with Microsoft "*contributed to the provision and interpretation of localised (at Local Authority level) COVID19-related search engine data (using Microsoft Bing data) provided to Public health England regional teams, which supported the detection of localised COVID19 clusters. This has become increasingly important as a tier-based approach to COVID management, based on the local epidemiology has been adopted in England*" (**S4**).

5. Sources to corroborate the impact (indicative maximum of 10 references)

S1. Consultant Epidemiologist, Immunisation and Countermeasures Division, Public Health England
S2. WHO, formerly Public Health England

S3. Flu vaccination 2020 to 2021 – Programme briefing for schools, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/902790/Flu_vaccination_programme_briefing_for_school_team_and_headteachers.pdf
S4. Bar-Ilan University, Israel, formerly Public Health England