# Tracking COVID-19 using online search

**Vasileios Lampos[1,*], Simon Moura[1], Elad Yom-Tov[4], Michael Edelstein[5], Maimuna Majumder[6], Yohhei Hamada[3], Molebogeng X. Rangaka[3,7], Rachel A. McKendry[2], and Ingemar J. Cox[1]**

[1]University College London, Department of Computer Science
[2]University College London, London Centre for Nanotechnology
[3]University College London, Institute for Global Health
[4]Microsoft Research
[5]Public Health England
[6]Harvard Medical School
[7]University of Cape Town, Division of Epidemiology and Biostatistics, School of Public Health

[*]Corresponding author, email: **v.lampos@ucl.ac.uk**

**Disclaimer:** The current version considers data up to and including April 14, 2020. The methods and results presented in this **working paper** should be considered as ongoing. The approach as well as the presented outcomes require further cross-checking and development. We would not normally publish work-in-progress, but we do so to potentially assist and collaborate with other groups supporting the response to COVID-19.

**Note:** The most up-to-date versions of this report can be found at github.com/vlampos/covid-19-online-search.

## Introduction

Online search and social media data are routinely used as alternative endpoints for monitoring the nationwide prevalence of infectious diseases, such as influenza[1–7]. Previous work has focused on supervised learning solutions, where *ground truth* information, in the form of historical syndromic surveillance reports, can be used to train machine learning models. However, no sufficient data —in terms of validity, representativeness, and time span— exist to apply such approaches for monitoring the emerging COVID-19 infectious disease pandemic caused by a novel coronavirus (SARS-CoV-2). Therefore, unsupervised, or semi-supervised solutions should be sought, and fully supervised solutions should be used with caution.

Recent outcomes have shown that it is possible to adapt an online search based model for influenza-like illness (ILI) for a source location, where syndromic surveillance data is available, and deploy it to a target location that cannot obtain ground truth information[8]. The accuracy of the target location model depends on identifying the correct search queries and corresponding weights via a transfer learning methodology. In this work, we draw a parallel to previous findings and attempt to develop an unsupervised model for COVID-19 by: (i) carefully choosing search queries that refer to related symptoms as identified by a survey from the National Health Service (NHS) in the United Kingdom (UK), and (ii) weighting them based on their reported ratio of occurrence in people infected by COVID-19. Furthermore, understanding that online searches can also be driven by concern rather than infection, we attempt to minimise this part of the signal by incorporating a basic news media coverage metric. In addition, we propose a transfer learning method for mapping supervised COVID-19 models from a country to another, in an effort to transfer noisy knowledge from areas that are ahead in the epidemic curve. Finally, we conduct a correlation and regression analysis to uncover potentially useful online search queries that could refer to underlying behavioural or symptomatic patterns in relation to confirmed COVID-19 cases. Results are presented for the UK, United States of America (US), Australia, Canada, France, Italy, Greece, and South Africa.

## Data

**Google search.** Google search data is obtained from the Google Health Trends API, a non public API created by Google for research on health-related topics. Data represent daily online search query frequencies for specific areas of interest. Query frequencies are defined as the sum of search sessions that include a target search term divided by the total number of search sessions (for a day and area of interest).[1] We have obtained data from September 30, 2011 to April 14, 2020 for the UK, US,

[1]Google defines a search session as a grouping of consecutive searches by the same user within a short time interval.

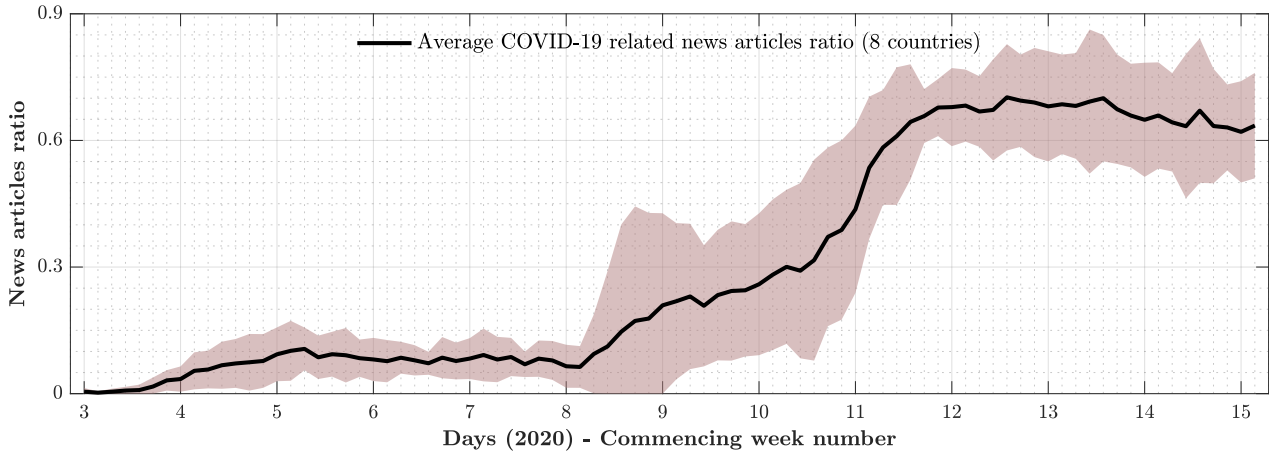


**Figure 1.** Average daily news articles ratio about COVID-19 across all countries in our analysis and corresponding confidence intervals (two standard deviations above and below the mean).

Australia, Canada, France, Italy, Greece, and South Africa. The list of search terms is determined by COVID-related symptoms and keywords. For each country, we mainly used queries in its native language(s).[2]

**News media volume.** We are using an extensive global news corpus to extract news media coverage trends for COVID-19 in all the countries of our study. This is estimated by counting the proportion of articles mentioning a COVID-19 related term. In particular, daily counts of total news media articles, and the subset that included at least one relevant keyword anywhere in the body of the text were collected from the MediaCloud database[3] via national corpora for the UK (93), US (225), Australia (61), Canada (79), France (360), Italy (178), Greece (75), and South Africa (135), where in the parentheses we state number of media sources considered per country. These counts were collected from September 30, 2019 through April 14, 2020, based on the following keywords:

- 'κορονοϊός', 'κορονοϊού', 'κορωνοϊός', 'κορωνοϊού', 'κορωνοϊοί', 'κορονοϊοί', 'covid', 'covid-19', 'covid 19', 'covid19', 'coronavirus', and 'ncov' for Greece, and
- 'covid', 'covid-19', 'covid 19', 'covid19', 'coronavirus', 'ncov' for the rest of the countries.

Figure 1 depicts the average daily ratio across all countries, as soon as it started being above zero (beginning of week 3, 2020), with two standard deviations as confidence intervals. There exists a distinctive pattern of (exponential) increase and, more recently, of a slowly decreasing trend. In addition, we observe a certain variability across locations and/or time periods, that adds to the potential value of this signal.

**COVID-19 symptoms.** We used data from the NHS first few hundred (FF100) symptom questionnaire based on people who have contracted SARS-CoV-2. FF100 provides a probability for each identified symptom.

**Aggregated confirmed COVID-19 cases.** The numbers of confirmed COVID-19 cases on a daily basis for the UK and England are obtained from PHE.[4] For the rest of the locations we obtain daily confirmed cases data from the European Centre for Disease Prevention and Control (ECDC).[5]

## Methods

**Unsupervised symptom-based online search model for COVID-19.** We generate $k$ symptom-based search query groups using the $k$ identified symptoms from the FF100 NHS questionnaire for COVID-19 ($k = 18$). In a separate model, we also consider two additional groups one referring to the symptom of anosmia,[6] and another that includes specific COVID-19 terminology, i.e. the "covid-19" keyword itself among others. Query groups may include different wordings for the same symptom or queries with minor grammatical differences (especially for queries in Greek and French). If a symptom is

[2]The search queries used for the results presented will be shared at a later time.
[3]MediaCloud, mediacloud.org
[4]Available at gov.uk/government/publications/covid-19-track-coronavirus-cases
[5]Available at ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide
[6]Anosmia is the loss of the sense of smell.

represented by more than one search query, then we obtain the total frequency (sum) across these queries. Query group time series are smoothed using a harmonic mean over the past 14 days (see Eq. 11 for a definition of the harmonic mean), and any trends across the entire period of the analysis are removed using linear detrending. We then apply a min-max normalisation to the frequency time series of each query group to obtain a balanced representation between more and less frequent searches. We divide our data into two periods of interest, the current one (from September 30, 2019 until April 14, 2020) and a historical one (from September 30, 2011 to September 29, 2019). The corresponding data sets are denoted by $\mathbf{X} \in \mathbb{R}_{\geq 0}^{N_1 \times k}$ and $\mathbf{H} \in \mathbb{R}_{\geq 0}^{N_2 \times k}$, where $N_1$, $N_2$ represent the different numbers of days in the current and historical data, respectively. We use the symptom conditional probability distribution from the FF100 to assign weights ($\mathbf{w} \in \mathbb{R}_{\geq 0}^{k}$) to each query category, and compute weighted time series ($\mathbf{x} = \mathbf{X}\mathbf{w}$, $\mathbf{h} = \mathbf{H}\mathbf{w}$), which are subsequently divided by the sum of $\mathbf{w}$ (weighted average). For the historical data, we divide their time span into yearly periods, and compute an average time series trend, $\mathbf{h}_\mu$, using two standard deviations as upper and lower confidence intervals.

**Minimising the effect of news media using confirmed COVID-19 cases.** On any given day the proportion of news articles about the COVID-19 pandemic is $m \in [0,1]$, and the weighted score of symptom-related online searches (see previous paragraph) is equal to $g$; we can apply a min-max normalisation so that $g \in [0,1]$ as well. We hypothesise that $g$ incorporates two signals based on infected ($g_p$) and concerned ($g_c$) users, respectively, i.e.

$$g = g_p + g_c \,. \tag{1}$$

Then, there exists a constant $\gamma \in [0,1]$ such that

$$g_p = \gamma g \,, \tag{2}$$

and

$$g_c = (1-\gamma) g \,. \tag{3}$$

We apply ordinary least squares (OLS) regression to learn a mapping from $g$ and $m$ to the actual number of confirmed infections, $d$, per day. For a meaningful interpretation of the regression's weights, $d$ is also min-max normalised, i.e. such that $d \in [0,1]$. In particular, at each day, we use the previous $N$ days (including the current one) to optimise

$$\underset{a_1,a_2}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} (d_i - a_1 g_i - a_2 m_i)^2 \,, \tag{4}$$

where $a_1$ and $a_2 \in \mathbb{R}$ denote the weights of the online search and news signals, respectively.

If $a_1 > 0$ and $a_2 < 0$, we can then hypothesise that the negative component coming from the media ($a_2 m$) is approximately equal to the unwanted component of the online search signal that is related to concern, i.e.

$$a_1 g_c \approx -a_2 m \,. \tag{5}$$

Solving this for $\gamma$, we get

$$\gamma = 1 + \frac{a_2 m}{a_1 g} \,. \tag{6}$$

Now, if $a_1 > 0$ and $a_2 > 0$, we can adjust for the relative contribution from the media by directly solving the equation $d = a_1 g + a_2 m = \gamma g$, which results to

$$\gamma = a_1 + a_2 \frac{m}{g} \,. \tag{7}$$

In the rare case that $a_2$ is set to a positive number that is close to zero (i.e. $a_2 \leq .01$), we set $\gamma = 1$, as the impact coming from the news media signal is negligible. If $a_1 \leq 0$, our current approach does not attempt to interpret this further, and therefore we also set $\gamma = 1$, meaning that we consider the signal from the online search data in its entirety. Valid values for $\gamma$ are thresholded so that $\gamma$ is always in $[0,1]$.

Using the above approach, we can learn a different $\gamma$ per day, and use $g_p = \gamma g$ as our unsupervised (or semi-supervised in this case) online search signal, attempting to minimise the impact of news in a dynamic fashion.

**Minimising the effect of news media using autoregression.** The previously described method requires a form of ground truth in order to be applicable, i.e. confirmed COVID-19 cases that are represented by variable $d$. However, confirmed cases may not be a population representative statistic given that in most countries tests are not yet conducted at the community level. To alleviate the effect of using potentially inaccurate information, we also obtain an estimate for $\gamma$ using only the time series of $g$ (online search score) and $m$ (news media COVID-19 ratio).[7]

The rationale of this approach is similar to the logic behind a Granger causality test[9]. First, we train a linear autoregressive (AR) model for forecasting the online search score at a time point (day) $t$, $g_t$, using its previous values; this is denoted by AR$(g)$. We also train a linear AR model with the same forecasting target, but an expanded space of observations that includes current ($m_t$) and previous (e.g. $m_{t-1}$) values of the news articles ratio; this is denoted by AR$(g,m)$. We then use the relative error difference of the two models in forecasting $g_t$, as our $\gamma$ for time point $t$. In particular, we first solve

$$\arg\min_{\mathbf{w},b_1} \frac{1}{N}\sum_{t=1}^{N}(g_t - w_1 g_{t-1} - w_2 g_{t-2} - b_1)^2\,, \tag{8}$$

to obtain a pair of weights ($\mathbf{w}$) and an intercept term ($b_1$) for AR$(g)$. We use 2 lags (past values) to keep the complexity of the task tractable given the small amount of samples at our disposal ($N$). We then solve

$$\arg\min_{\mathbf{w},\mathbf{v},b_2} \frac{1}{N}\sum_{t=1}^{N}(g_t - w_1 g_{t-1} - w_2 g_{t-2} - v_1 m_t - v_2 m_{t-1} - v_3 m_{t-2} - b_2)^2\,, \tag{9}$$

to obtain the weights ($[\mathbf{w};\mathbf{v}]$) and an intercept term ($b_2$) for AR$(g,m)$. Using both models, we forecast the next (unseen) value of $g$, which following the notation in the equations above is $\hat{g}_{t+1}$, and compute the absolute error from its known true value, $g_{t+1}$. This yields errors, $\varepsilon_1$ and $\varepsilon_2$ for AR$(g)$ and AR$(g,m)$, respectively. If $\varepsilon_1 < \varepsilon_2$, then the news media signal does not help to improve the accuracy of AR$(g)$, and hence we assume that it does not affect the online searches. Otherwise, we estimate its effect to be represented by

$$\gamma = \frac{\varepsilon_2}{\varepsilon_1}\,. \tag{10}$$

After obtaining a time series of $\gamma$'s for the all days in our analysis, we smooth each one of them using a harmonic mean over the values of the previous 6 days (or 7 days including the day of focus).

**Transferring supervised COVID-19 models to different countries.** Previous work has shown that it is possible to transfer a model for seasonal flu, based on online search query frequency time series, from one country that has access to historical syndromic surveillance data to another that has not[8]. Here, we adapt this method to transfer a model for COVID-19 from a source country where the disease spread has progressed significantly to a target country that is still in earlier stages of the epidemic curve. The rationale for this is that a supervised model based on data from the source country might be able to capture the disease dynamics better. The steps and data transformations that are required to apply this technique are detailed below.

Search query frequency time series are denoted by $\mathbf{S} \in \mathbb{R}_{\geq 0}^{M \times n_S}$ and $\mathbf{T} \in \mathbb{R}_{\geq 0}^{M \times n_T}$, for the source and target countries respectively; $M$ denotes the number of days considered, and $n_S$, $n_T$ the number of queries for the two locations. As these time series are quite volatile for some locations in our study, something that does not help in cross-location mapping of the data, we have smoothed them using a harmonic query frequency mean based on a window of the $D$ past days. More specifically, a smoothed search query frequency $s_i$ for a day $i$ is equal to:

$$s_i = \frac{1}{\sum_{p=1}^{D}\frac{1}{p}}\sum_{p=1}^{D}\frac{x_{i-p+1}}{p}\,, \tag{11}$$

where $x_{(\cdot)}$ denotes the raw (non smoothed) search query frequency.

We train an elastic net model on data from the source location[10], similarly to previous work on ILI[4,6,11]. In particular, we solve the following optimisation task

$$\arg\min_{\mathbf{w},b}\left(\|\mathbf{y} - \mathbf{S}\mathbf{w} - b\|_2^2 + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2^2\right)\,, \tag{12}$$

where $\mathbf{y} \in \mathbb{R}^M$ denotes the daily number of confirmed COVID-19 cases in the source location, $\lambda_1$, $\lambda_2 \in \mathbb{R}_{>0}$ are the $\ell_1$- and $\ell_2$-norm regularisation parameters, and $\mathbf{w} \in \mathbb{R}^{n_S}$, $b \in \mathbb{R}$ denote the query weights and regression intercept, respectively. Prior to

[7] In the current version of the report, we show results only for this news debiasing approach.

deploying elastic net, we apply a min-max normalisation on both $\mathbf{S}$ and $\mathbf{y}$. We fix the ratio of $\lambda$'s, and then train $q$ models for different values of $\lambda_1$. All different regression models represented by the columns of $\mathbf{W} \in \mathbb{R}^{n_S \times q}$, and the elements of $\mathbf{b} \in \mathbb{R}^q$, are used as an ensemble for a more inclusive transfer that combines various source models with different sparsity levels.

To generate an equivalent feature space for the target location (same dimensionality, similar feature attributes), we first identify query group pairs between the source and the target location using the symptom categories in the NHS FF100 questionnaire. We map a source query to the target query from the same symptom category that maximises their Pearson correlation based on their frequency time series. To do this more effectively, prior to computing correlations, we shift the data by $z$ days (looking at a maximum window of 60 days backwards or forwards) so that the average correlation between search query frequencies in $\mathbf{S}'$ and $\mathbf{T}$ are maximised; $\mathbf{S}'$ here denotes a subset of $\mathbf{S}$ that includes only the search queries that have been assigned a non zero weight by the elastic net (Eq. 12). If no target search query exists for a certain symptom category, we use the best correlated one from all target queries available (irrespectively of the symptom category) as its mapping. After this process, we end up with a subset $\mathbf{Z} \in \mathbf{R}^{M \times n_S}$ of the target feature space $\mathbf{T}$. Notably, $\mathbf{Z}$ does not necessarily hold data for $n_S$ distinct queries as different source queries may have been mapped to the same target query. $\mathbf{Z}$ is subsequently normalised using min-max. To make both feature spaces ($\mathbf{S}$, $\mathbf{Z}$) numerically compatible we scale the latter based on their mean, column-wise (per search query) ratio $\mathbf{r} \in \mathbb{R}^{n_S}_{\geq 0}$, i.e. $\mathbf{Z}_S = \mathbf{Z} \odot \mathbf{r}$. Now, we can deploy the ensemble source models to the target space, making multiple inferences (for different $\lambda_1$ values) held in $\mathbf{Y} \in \mathbb{R}^{n_S \times q}$:

$$\mathbf{Y} = \mathbf{Z}_S \mathbf{W} + \mathbf{b}\,. \tag{13}$$

We then reverse the min-max normalisation for each one of the inferred time series (columns of $\mathbf{Y}$) using values from the source model's ground truth $\mathbf{y}$ (prior to its normalisation). Finally, we compute the mean of the ensemble (across the rows of $\mathbf{Y}$) as our target estimate, and also use two standard deviations to form a 95% confidence interval.

**Correlation and regression analysis.** The relationship of search frequency time series and confirmed cases can uncover symptoms or behaviours related to COVID-19. However, since confirmed cases data may not be representative of community level disease rates, looking at this relationship separately for each country might produce misleading outcomes. To mitigate this to the extent possible, we combine the data from $C$ countries and produce an aggregate set of query frequencies, $\mathbf{Z}_\alpha \in \mathbb{R}^{CM \times n}$, where $M$, $n$ denote the considered days and search queries, respectively. We denote the aggregated daily confirmed COVID-19 cases for these countries with $\mathbf{y}_\alpha \in \mathbb{R}^{CM}$. Prior to the aggregation, we apply min-max normalisation on the query frequency, and confirmed cases time series separately for each country.

Initially, we compute the Pearson correlation between the columns of $\mathbf{Z}_\alpha$ and $\mathbf{y}_\alpha$. Correlation is an informative metric, but considers each search query in isolation. Therefore, we also perform a multivariate regression analysis to more rigorously estimate the impact of each search query in estimating confirmed cases. To do this, we apply elastic net regularised regression (see Eq. 12), training and testing models for the past $K$ days. During each of the training phases, we use data up to and including the past $k-1$ days, and test only the $K$-th day (unseen); this results into daily test sets of size $C$ (one value for each country). We explore elastic net's regularisation path to consider $L$ models that maintain (by assigning a nonzero weight) up to a reasonable percentage of the features (e.g. 50%), so as a solution is not overfitting. We do this gradually, selecting first 1% of the features and moving towards the maximum considered percentage. In this experiment, we use the test set to identify the most accurate (in terms of mean squared error) model at each density level. For this model, we determine the impact of each one of the features (search queries) by considering both its frequency and allocated weight. The impact $\Theta(\cdot)$ of a query $q$ is equal to

$$\Theta(q) = \frac{\sum_{\ell=1}^{L}\sum_{t=1}^{K}\sum_{j=1}^{C} f_{t,j}\, w_{\ell,t}}{\sum_{\ell=1}^{L}\sum_{t=1}^{K}\sum_{j=1}^{C} \hat{y}_{\ell,t,j}}\,, \tag{14}$$

where $f_{t,j}$ denotes the query frequency at time point (day) $t$ and for country $j$, $w_{\ell,t}$ the corresponding weight at sparsity level $\ell$, and $\hat{y}_{\ell,t,j}$ the respective estimated confirmed cases. Impacts are summed across all the considered days, and model densities, and normalised at the end by the sum of all the corresponding COVID-19 case estimates.[8] These normalised impacts are used to inform our regression analysis.

## Results

The current online search based scores for COVID-19 in 8 nations are depicted in Figures 2 and 3 (data up to April 14, 2020). For a better visualisation, all time series are smoothed using a 7-point moving average, 3 days prior and after each point.

[8] Each query frequency-weight product is a component of a sum used to derive a COVID-19 cases estimate.

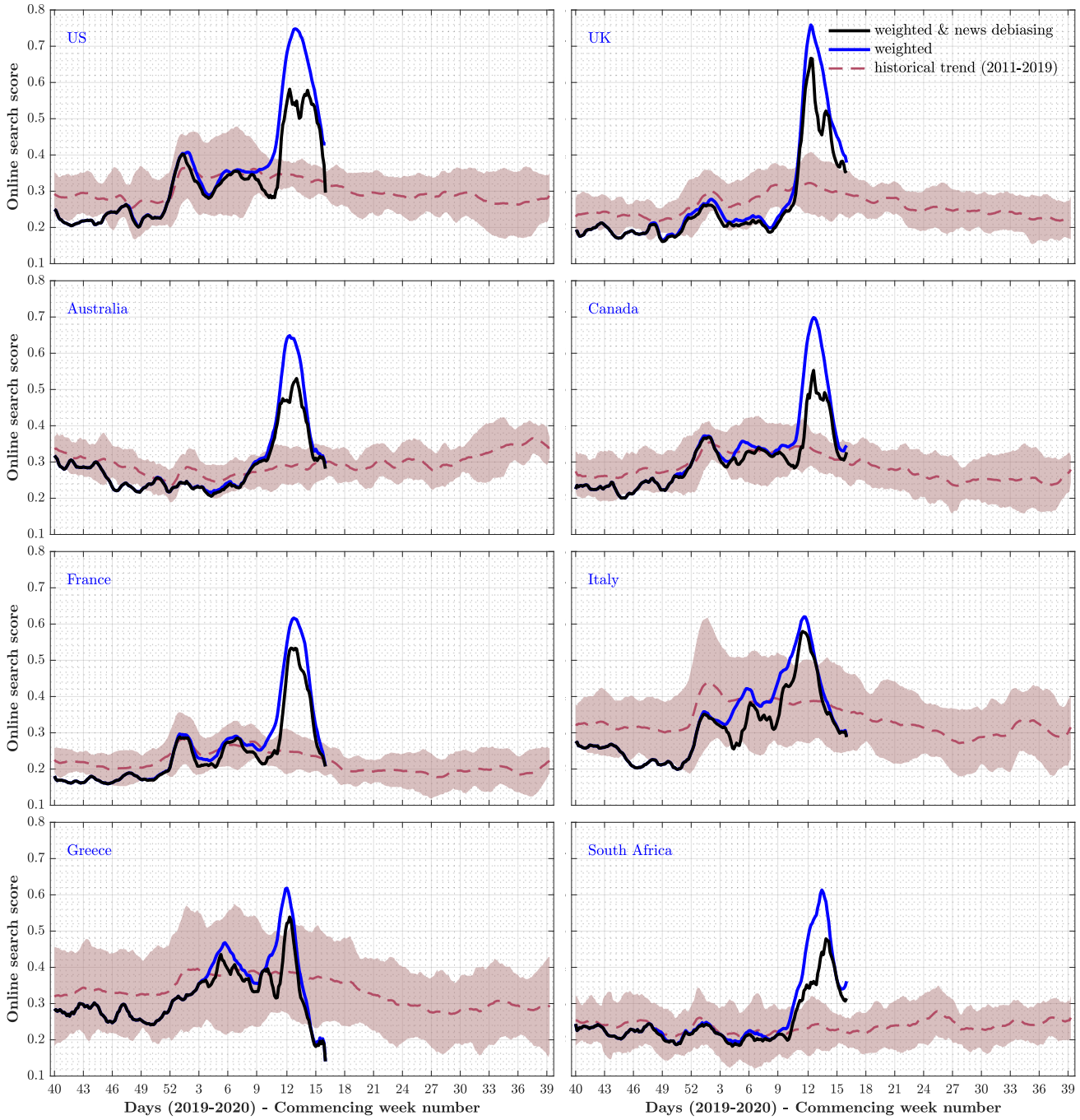


**Figure 2.** Online search based scores for COVID-19 related symptoms as identified by the NHS FF100 survey for 8 nations up to and including April 14, 2020. Query frequencies are weighted by symptom frequency as described in Methods (blue line). We have also included estimates after minimising news media effects using data from a global news media corpus (black line). These scores are compared with an average 8-year trend of the weighted model (dashed line) and its corresponding confidence intervals (shaded area). For a better visualisation all time series are smoothed using a 7-point moving average.

Figure 2 shows scores based on symptom-related query frequencies that are weighted by the actual symptom probability as reported in the NHS FF100 survey for COVID-19. Expanding on this, Figure 3 shows scores when search queries that are about the symptom of anosmia as well as strictly about COVID-19 are added as additional query groups. We set the weight of the anosmia symptom category to 0.4 (2 in 5 cases), as we wait for confirmation from an expert analysis. The weight of the strictly COVID-19 related queries is set equal to 1. The rationale behind including the latter category is that by now (and

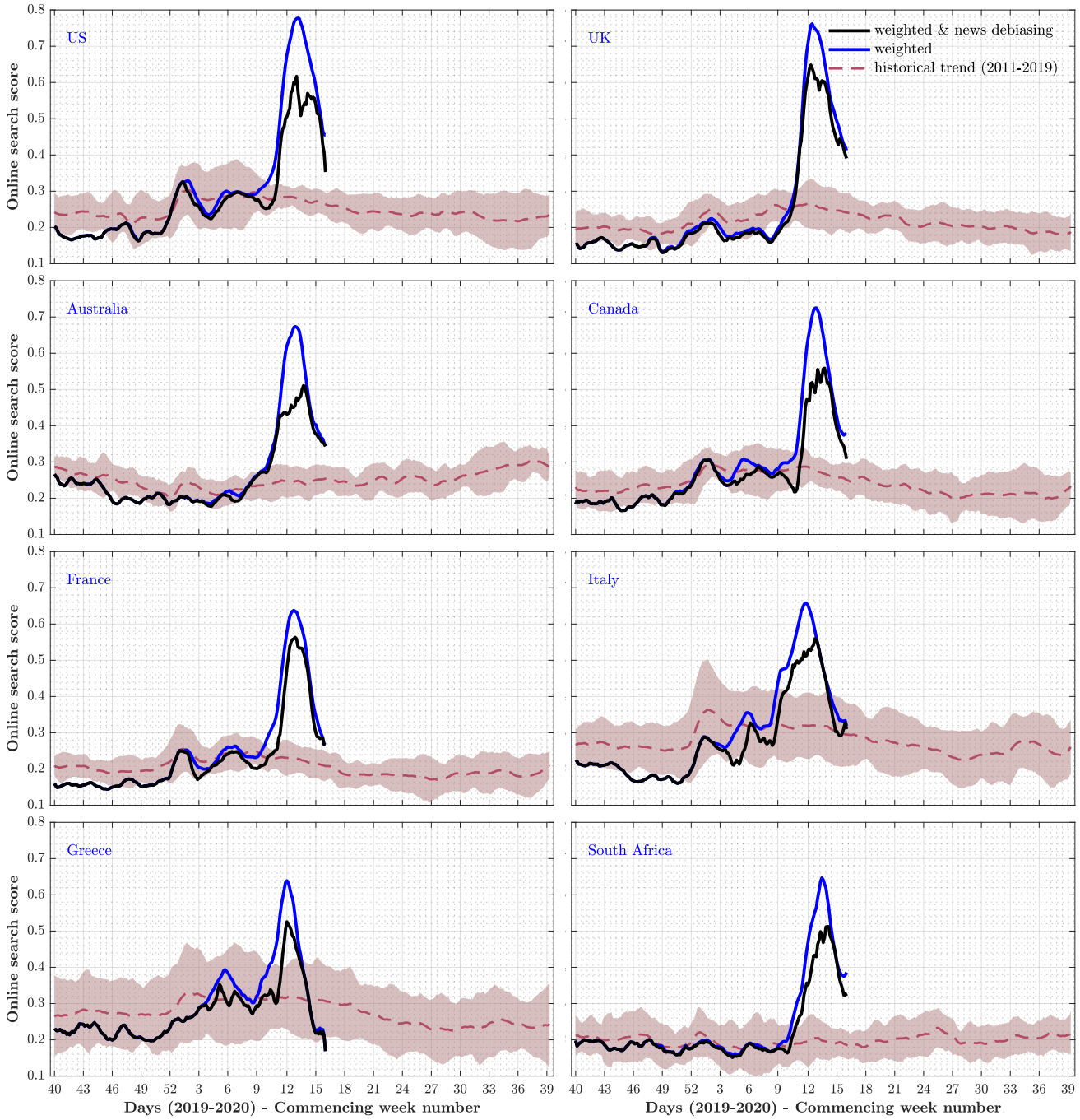


**Figure 3.** Online search based scores for COVID-19 related symptoms as identified by the NHS FF100 survey, in addition to queries about the symptom of anosmia, and a group of coronavirus-related terms, for 8 nations up to and including April 14, 2020. Query frequencies are weighted by symptom frequency as described in Methods (blue line). We have also included estimates after minimising news media effects using data from a global news media corpus (black line). These scores are compared with an average 8-year trend of the weighted model (dashed line) and its corresponding confidence intervals (shaded area). For a better visualisation all time series are smoothed using a 7-point moving average.

perhaps at an earlier time point) people who experience COVID-19 related symptoms might search about the disease directly as its name(s) and associated symptoms are broadly known.

Focusing on the weighted signal (blue lines), we observe exponentially increasing rates that exceed the estimated confidence intervals in most investigated countries. At the same time, we are also observing a recent drop of the score in all countries. The

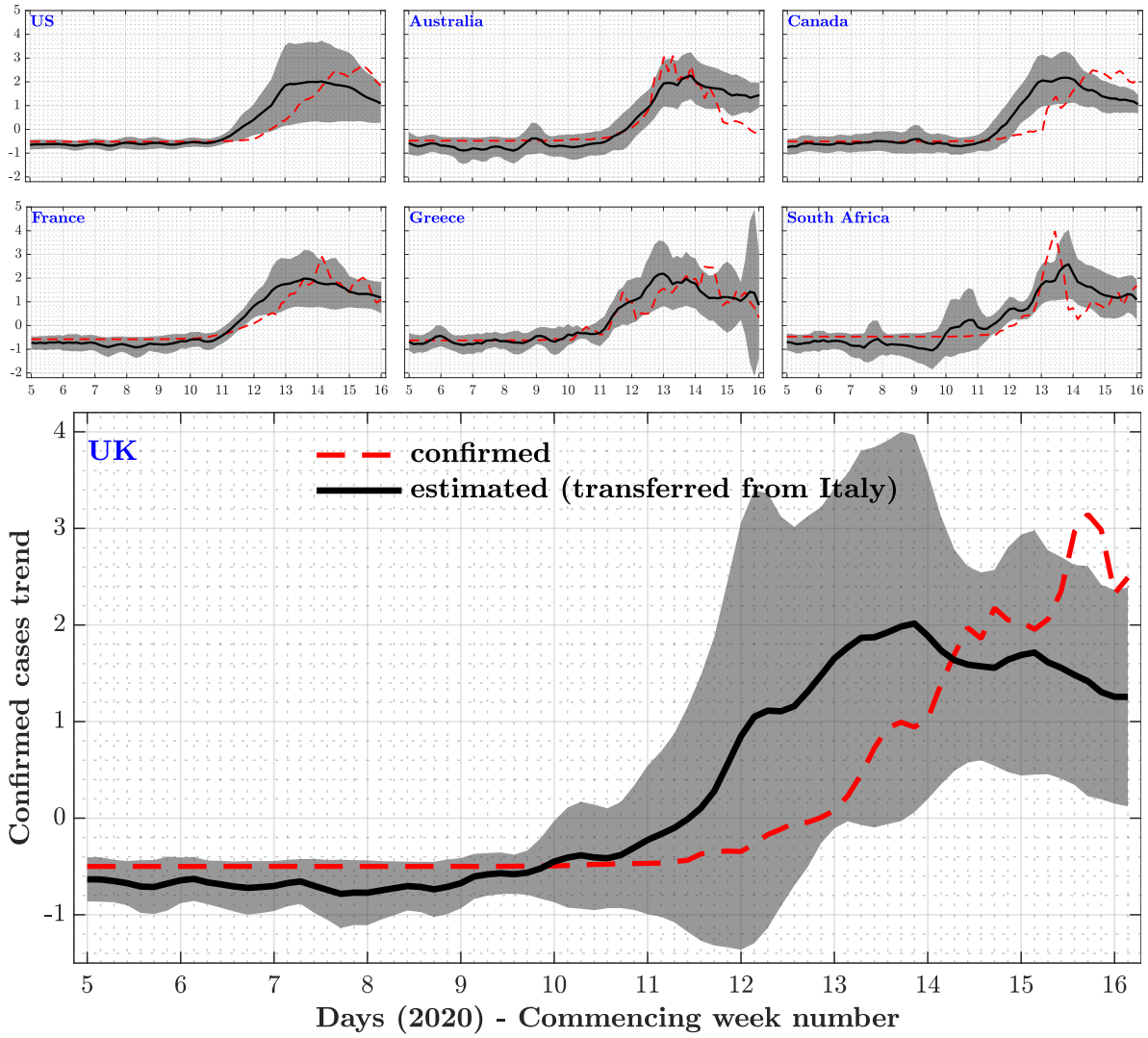


**Figure 4.** Transferring a supervised model for Italy to other countries in our analysis. The figures show an estimated confirmed cases trend (with confidence intervals) for all locations in our analysis (minus Italy) compared to the recorded confirmed cases as reported by PHE and the ECDC. Plot lines have been standardised, and then smoothed using a 3-point moving average.

added query categories (Fig. 3) increase the maximum scores per country, but do not affect the overall trend.

Looking at the scores where we have attempted to minimise the effect of news (black lines) using the autoregressive approach (note that we use the previous $N = 56$ days to determine this; see Eqs. 8, 9, and 10), we observe more conservative estimates in all locations, including a recent drop or an altered trend (e.g. increasing vs. decreasing) in some of them (e.g. the US). A detailed analysis of the observed trends is left for a later version of this manuscript as conclusions cannot be drawn before the (approximate) end of the first wave of the pandemic.

Figure 4 showcases the outcome of an experiment where we trained a model for Italy and then transferred it to the rest of countries in our analysis. Italy was chosen as the source country because it is considered to be in front of the rest in terms of epidemic progression. During this experiment search query frequency time series were smoothed (as explained in Methods) using a harmonic mean of the past 14 days. We train 100,000 elastic net models for the source location, exploring the entire $\ell_1$-norm regularisation path. For all target countries, we transfer source models that activate (non zero weight) from 1% to 99% of the search queries (88,138 models in total). Interestingly, the mapped trends correspond sufficiently well to confirmed cases data in most countries taking into account the fact that they lack supervision at the target locations. Notably though, here the goal is not necessarily for the two trends (estimated and confirmed cases) to match, as the transferred models may be capable of capturing signals that are missed out by the current surveillance systems. The caveat of this approach is that it relies on

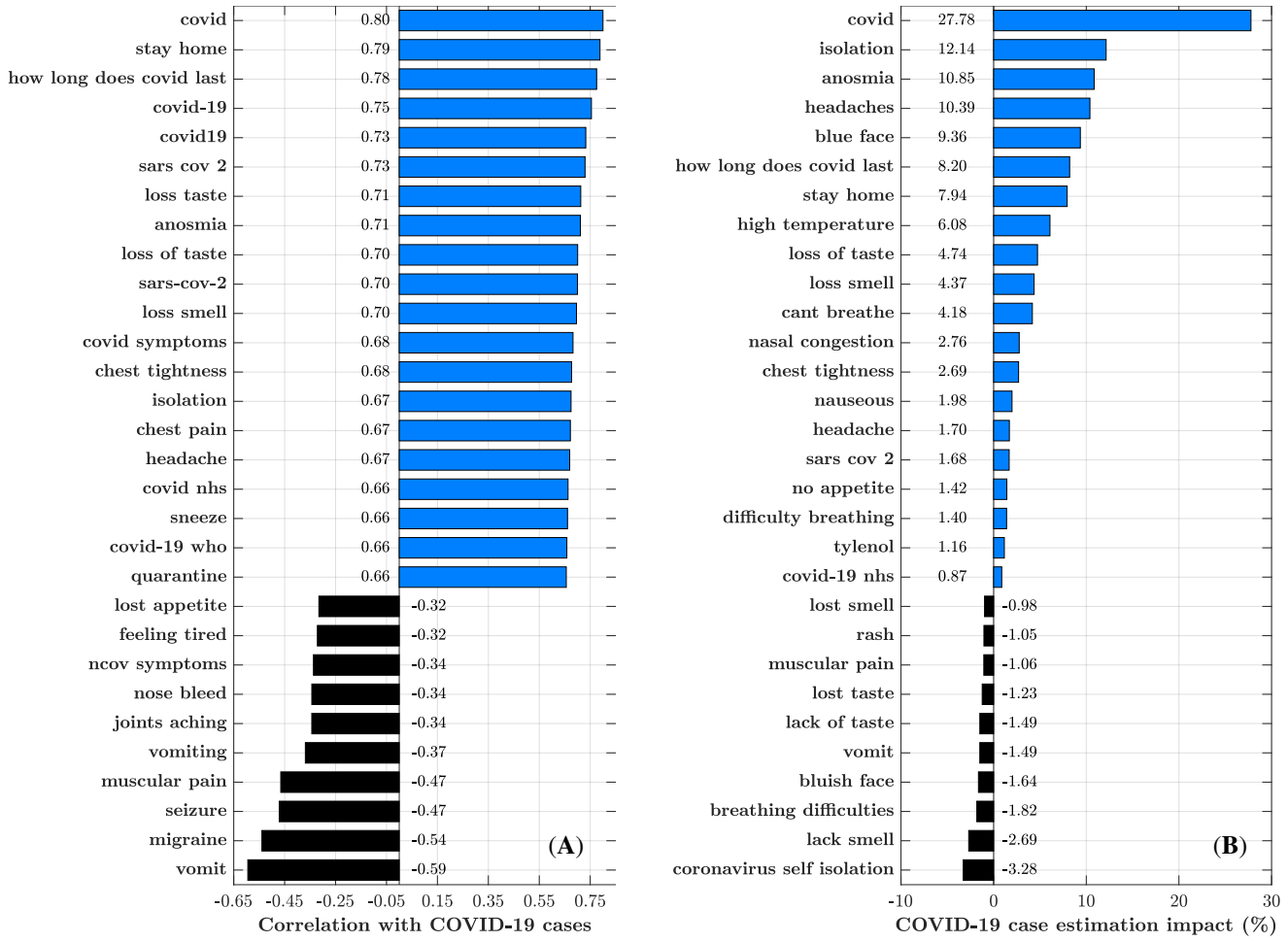


**Figure 5.** Correlation and regression analysis of search query frequencies against confirmed COVID-19 cases in four countries (US, UK, Australia, and Canada). (**A**) Top-20 positively and top-10 negatively correlated search queries with COVID-19 confirmed cases; (**B**) Top-20 positively and top-10 negatively impactful queries in nowcasting COVID-19 confirmed cases.

the existence of a representative population sample of confirmed COVID-19 cases at the source country (in this case Italy). Otherwise, the transferred model will inherit the biases of the source model. Furthermore, the transfer learning approach itself requires a significant level of similarity in user search behaviour about COVID-19 between different countries, as explained in a similar attempt to transfer models for influenza-like illness[8].

Finally, we performed a correlation and regression analysis on online search frequency and COVID-19 confirmed cases time series as described in Methods. We aggregated data from 4 English speaking countries, namely the UK, US, Australia, and Canada, as their share the exact same search queries. We first estimated the Pearson correlation between queries and confirmed cases at all locations (in an aggregate fashion), from December 31, 2019 up to and including April 14, 2020. Results are depicted in Figure 5(A). It is quite striking that the search query "stay home" comes up on top, which is not a symptom, but a behaviour associated with reducing the spread of the virus. Focusing on symptoms, we can also see that anosmia, loss of taste (ageusia), chest pain, headache show quite strong positive correlations. On the other hand symptoms such as migraine and vomiting show strong anticorrelation. For the regression analysis, we focused on the past 6 weeks, meaning that we trained models for and tested them on each day of that period. We considered models up to a 50% feature density (nonzero weights for half of the considered search queries), at which point we saw signs of overfitting. The outcomes of this analysis are depicted in Figure 5(B). We see that the most impactful feature is search queries that include the term "covid", which is something expected given the magnitude of this pandemic. Anosmia, loss of taste, high temperature, headache, nasal congestion, breathing difficulty, chest tightness, and nausea are symptoms (in that order) that are also impactful from the perspective of search predictiveness. In addition, behaviours such as isolation, staying at home are also very relevant.

## References


1. Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D. & Weinstein, R. A. Using Internet Searches for Influenza Surveillance. *Clin. Infect. Dis.* **47**, 1443–1448 (2008).
2. Lampos, V. & Cristianini, N. Tracking the flu pandemic by monitoring the social web. In *Proc. of the 2nd International Workshop on Cognitive Information Processing*, 411–416 (2010).
3. Culotta, A. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proc. of the 1st Workshop on Social Media Analytics*, 115–122 (2010).
4. Lampos, V., Miller, A. C., Crossan, S. & Stefansen, C. Advances in nowcasting influenza-like illness rates using search query logs. *Sci. Rep.* **5** (2015).
5. Yang, S., Santillana, M. & Kou, S. C. Accurate Estimation of Influenza Epidemics using Google Search Data via ARGO. *PNAS* **112**, 14473–14478 (2015).
6. Lampos, V., Zou, B. & Cox, I. J. Enhancing Feature Selection Using Word Embeddings: The Case of Flu Surveillance. In *Proc. of the 26th International Conference on World Wide Web*, 695–704 (2017).
7. Wagner, M., Lampos, V., Cox, I. J. & Pebody, R. The added value of online user-generated content in traditional methods for influenza surveillance. *Sci. Rep.* **8** (2018).
8. Zou, B., Lampos, V. & Cox, I. J. Transfer Learning for Unsupervised Influenza-like Illness Models from Online Search Data. In *Proc. of the 28th International Conference on World Wide Web*, 2505–2516 (2019).
9. Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438 (1969).
10. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. Royal Stat. Soc.: Ser. B* **67**, 301–320 (2005).
11. Lampos, V., Yom-Tov, E., Pebody, R. & Cox, I. J. Assessing the impact of a health intervention via user-generated internet content. *Data Min. Knowl. Discov.* **29**, 1434–1457 (2015).


## Acknowledgements


V.L, S.M., I.J.C, and R.M. would like to acknowledge all levels of support from the EPSRC projects EP/K031953/1 ("EPSRC IRC in Early-Warning Sensing Systems for Infectious Diseases") and EP/R00529X/1 ("i-sense: EPSRC IRC in Agile Early Warning Sensing Systems for Infectious Diseases and Antimicrobial Resistance"). The authors would like to thank the NHS FF100 team for their effort in collecting symptom information for COVID-19. We would also like to thank Ettore Severi, Anna Odone, and Daniela Paolotti for assisting in the translation of search queries from English to Italian. Finally, V.L. would like to thank Sam J. Gilbert for interesting discussions and pointers during the development of this work.


## Author contribution statement

V.L. conceived this research, formed the majority of the data sets, developed the methods, ran the experiments, and wrote the manuscript. M.M. provided news coverage data that were used to minimise the effect of news media in our models. S.M. provided a translation of search query groups from English to French, and M.X.R. and Y.H. from English to various languages spoken in South Africa. S.M., E.Y.T., M.E., M.M., R.M., and I.J.C. provided feedback in various levels of this work, including the methodological approach, and contributed in writing the manuscript.