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]National Infection Service, 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**

ABSTRACT

Research outcomes over the years have showcased the capacity of online search behaviour to model various properties of infectious diseases. In this work we use online search query frequency time series to gain insights about the prevalence of COVID-19 in multiple countries. We first develop unsupervised modelling techniques based on identified symptom categories by United Kingdom's National Health Service. We then propose ways for minimising an expected bias in these signals partially generated by the early and continuous exposure to news media. We also look into transfer learning techniques for mapping supervised models from countries where the disease spread has progressed to countries that are in earlier phases of the epidemic curve. Furthermore, we analyse the time series of online search queries in relation to confirmed COVID-19 cases data jointly across multiple countries, uncovering interesting patterns. Finally, we show results from short-term forecasting models based on Gaussian Processes that combine confirmed cases and online search data time series.

Disclaimer. The first version of this manuscript was published on March 18, 2020. The current version considers data up to and including April 27, 2020. Findings must be read with caution prior to the completion of the first wave of the COVID-19 pandemic. More frequently updated versions 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]. Apart from timeliness and a non-stop operational capacity, online user trails can provide denser spatial coverage and expand to different demographic groups compared to traditional methodologies[7,8]. During emerging epidemics, they could additionally offer community level insights that current monitoring systems may not be able to obtain given the widespread social distancing and self-isolation measures.

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, for most locations, if not all, no sufficient data —in terms of validity, representativeness, and time span— currently 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 historical ground truth data[9]. 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

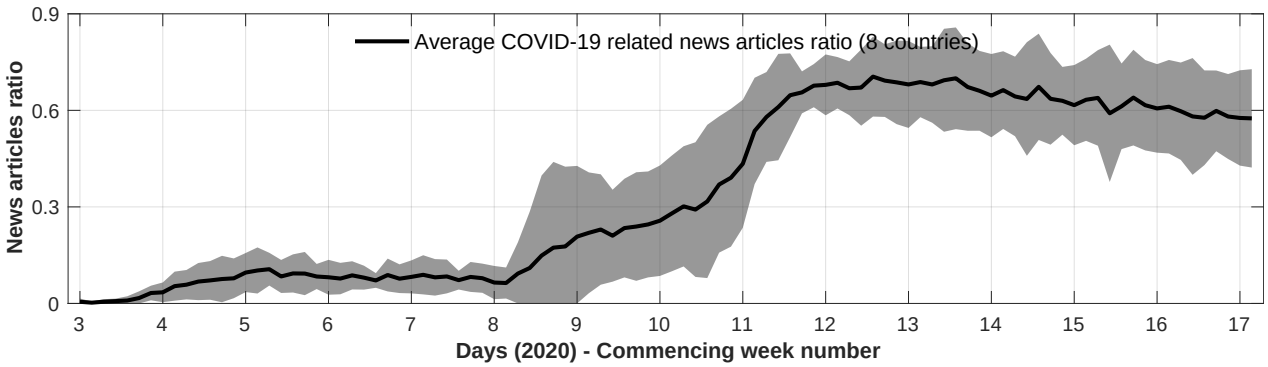

Figure 1. Average daily news articles ratio about COVID-19 across all countries in our analysis and corresponding confidence intervals (two standard deviations above and below the mean).

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 to areas that are at earlier stages of the epidemic. Taking noisy supervision a step further, 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. Finally, we develop short-term forecasting models for predicting confirmed COVID-19 cases up to 2 weeks in advance. This latter part of the work should be read with additional caution as we have limited temporal data to evaluate the quality of the forecasting models. 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 27, 2020 for the UK, US, 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]

COVID-19 related news media coverage. 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 the number of media sources considered per country. These counts were collected from September 30, 2019 through April 27, 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 variance 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) survey based on people who have contracted SARS-CoV-2. FF100 provides a probability for each identified symptom.[4]

[1] Google defines a search session as a grouping of consecutive searches by the same user within a short time interval.
[2] The search queries used in our analysis will be publicly shared at a later point in time.
[3] MediaCloud, mediacloud.org
[4] The outcomes of FF100 will be published (by the NHS/PHE) at a later point in time.

Confirmed COVID-19 cases time series. For all countries in our analysis, we obtain daily confirmed COVID-19 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 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. 9 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 27, 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

$$g_p = \gamma g \;\text{ and }\; g_c = (1-\gamma)g\,. \tag{2}$$

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

$$\underset{a_1,a_2}{\arg\min}\,\frac{1}{N}\sum_{i=1}^{N}(d_i - a_1 g_i - a_2 m_i)^2\,, \tag{3}$$

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. $a_1 g_c \approx -a_2 m$. Solving this for γ, we get

$$\gamma = 1 + \frac{a_2 m}{a_1 g}\,. \tag{4}$$

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

$$\gamma = a_1 + a_2\frac{m}{g}\,. \tag{5}$$

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 γ are thresholded so that γ is always in $[0,1]$.

Using the above approach, we can learn a different γ 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.

[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.

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 γ using only the time series of g (online search score) and m (news media ratio about COVID-19).[7]

The rationale of this approach is similar to the logic behind a Granger causality test[10]. 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 γ for time point t. In particular, we first solve

$$\arg\min_{\mathbf{w},b_1} \frac{1}{N}\sum_{t=1}^{N}(g_t - w_1 g_{t-1} - w_2 g_{t-2} - b_1)^2\,, \tag{6}$$

to learn 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

$$\arg\min_{\mathbf{w},\mathbf{v},b_2} \frac{1}{N}\sum_{t=1}^{N}(g_t - w_1 g_{t-1} - w_2 g_{t-2} - v_1 m_t - v_2 m_{t-1} - v_3 m_{t-2} - b_2)^2\,, \tag{7}$$

to learn 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, ε_1 and ε_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

$$\gamma = \frac{\varepsilon_2}{\varepsilon_1}\,. \tag{8}$$

After obtaining a time series of γ'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[9]. 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:

$$s_i = \frac{1}{\sum_{p=1}^{D}\frac{1}{p}}\sum_{p=1}^{D}\frac{x_{i-p+1}}{p}\,, \tag{9}$$

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

We train an elastic net model on data from the source location[11], similarly to previous work on ILI[4,6,8] or other text regression tasks[12,13]. In particular, we solve the following optimisation task

$$\arg\min_{\mathbf{w},b}\left(\|\mathbf{y} - \mathbf{S}\mathbf{w} - b\|_2^2 + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2^2\right)\,, \tag{10}$$

where $\mathbf{y} \in \mathbb{R}^M$ denotes the daily number of confirmed COVID-19 cases in the source location, λ_1, $\lambda_2 \in \mathbb{R}_{>0}$ are the ℓ_1- and ℓ_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 λ's, and then train q models for different values of λ_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. 10). 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 λ_1 values) held in $\mathbf{Y} \in \mathbb{R}^{n_S \times q}$:

$$\mathbf{Y} = \mathbf{Z}_S \mathbf{W} + \mathbf{b}\,. \tag{11}$$

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 to balance out local properties.

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. 10), training and testing K models for the past K days. During each of K the training phases, assuming it contains η days, we use data up to and including the past $\eta - 1$ days to train, and test only the last day (η_{th}) which is 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

$$\Theta(q) = \frac{\sum_{\ell=1}^{L}\sum_{t=1}^{K}\sum_{j=1}^{C} f_{t,j}\, w_{\ell,t}}{\sum_{\ell=1}^{L}\sum_{t=1}^{K}\sum_{j=1}^{C} \hat{y}_{\ell,t,j}}\,, \tag{12}$$

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 ℓ, 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.

Short-term forecasting of confirmed cases. Short-term forecasts of confirmed COVID-19 cases can be conducted using the time series of past records. Augmenting this autoregressive (AR) signal with online user trails could help to improve accuracy, if we draw a parallel with influenza-like illness rate modelling[14].

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

Let $\mathbf{Z} \in \mathbb{R}_{\geq 0}^{M \times N}$ denote the frequency of N search queries for M days, and let $\mathbf{y} \in \mathbb{N}^M$ be the corresponding confirmed COVID-19 cases. We first solve a strictly AR task using L past values, meaning that at a time point t we use $\mathbf{y}_{\text{AR}}(t,L) = [y_t, y_{t-1}, \ldots, y_{t-L}] \in \mathbf{y}$ to forecast y_{t+D}, performing D days ahead forecasting. We denote this forecasting model as **AR-F**. We also augment our observations by incorporating search query frequency data for the current time instance, $\mathbf{z}_t \in \mathbf{Z}$, resulting to an input $\mathbf{x}_t = [\mathbf{z}_t; \mathbf{y}_{\text{AR}}(t,L)]$ that is held in the concatenated matrix $\mathbf{X} = [\mathbf{Z}; \mathbf{Y}_{\text{AR}}(L)] \in \mathbb{R}^{M \times (N+L+1)}$.[9] We denote this search AR forecasting model as **SAR-F**. For simplicity, we drop the notation t in subsequent references to these variables.

We deploy a series of Gaussian Process (GP) models, training a different model for each D days ahead forecasting task. The choice of GPs is justified by previous work on modelling infectious diseases[4,6,15]. GPs are defined as random variables any finite number of which have a multivariate Gaussian distribution[16]. GP methods aim to learn a function $f : \mathbb{R}^m \to \mathbb{R}$ drawn from a GP prior. They are specified through a mean and a covariance (or *kernel*) function, i.e. $f(\mathbf{x}) \sim \text{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$, where $\mathbf{x}$ and $\mathbf{x}'$ (both $\in \mathbb{R}^m$) denote rows of the input matrix $\mathbf{X}$. By setting $\mu(\mathbf{x}) = 0$, a common practice in GP modelling, we focus only on the kernel function. The specific composite kernel function used in our forecasting models is given by

$$k(\mathbf{x}, \mathbf{x}') = k_{\text{SE}}(\mathbf{x}, \mathbf{x}'; \sigma_1, \ell_1) + k_{\text{SE}}(\mathbf{x}, \mathbf{x}'; \sigma_2, \ell_2) + \sigma_3^2 \delta(\mathbf{x}, \mathbf{x}'), \tag{13}$$

where $k_{\text{SE}}(\cdot,\cdot)$ denotes a squared exponential (SE) covariance function, $\delta(\cdot,\cdot)$ denotes a Kronecker delta function used for an independent noise component, ℓ's denote lengthscale parameters, and σ^2's are scaling factors (variance). The SE kernel is defined by

$$k_{\text{SE}}(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{r^2}{2\ell^2}\right), \tag{14}$$

where $r = \mathbf{x} - \mathbf{x}'$. By replacing $\mathbf{x}$, $\mathbf{x}'$ with $\mathbf{y}_{\text{AR}}$, $\mathbf{y}'_{\text{AR}}$, we obtain the covariance functions for a strictly autoregressive model. The hyperparameters of the covariance function $(\sigma_1, \sigma_2, \sigma_3, \ell_1, \ell_2)$ are optimised using Gaussian likelihood and variational inference[16].

Results

The unsupervised online search based scores for COVID-19 in 8 countries are depicted in Figures 2 and 3 (data up to April 27, 2020). For a better visualisation, all time series are smoothed using a 7-point moving average, 3 days prior and after each point. 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 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 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. 6, 7, and 8), we observe more conservative estimates in all locations, including a recent drop or an altered trend in some of them (e.g. the US).[10]

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 ℓ_1-norm regularisation path. For all target countries, we transfer source models that activate (non zero weight) from 5% to 95% of the search queries (71,759 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 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[9].

[9] Under this notation, we assume that values of $\mathbf{y}$ before the first sample of $\mathbf{Z}$ are at our disposal.

[10] A detailed analysis of the results will be conducted as soon as we reach to the completion of the first wave of the pandemic.

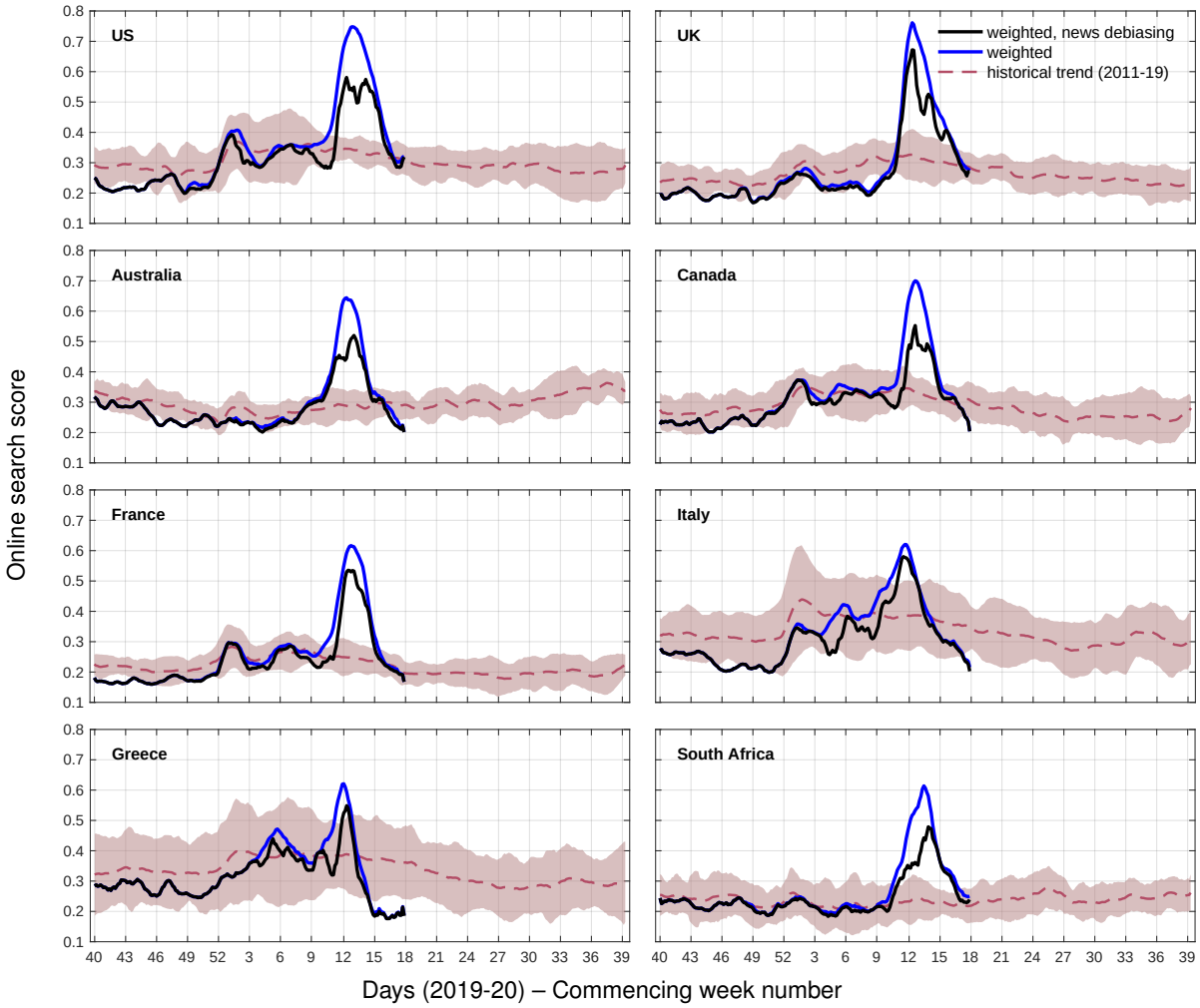

Figure 2. Online search based scores for COVID-19 related symptoms as identified by the NHS FF100 survey for 8 countries up to and including April 27, 2020. Query frequencies are weighted by symptom frequency (blue line), and have news media effects minimised (black line). These scores are compared with an average 8-year trend of the weighted model (dashed line) and its corresponding confidence intervals (shaded area). All time series are smoothed using a 7-point moving average.

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 US, UK, 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 27, 2020. Results showcasing the top correlated and anti-correlated search queries are depicted in Figure 5(A). It is quite striking that the search query "stay home" comes up second ($r = .73$), tied with "sars cov 2", and right after the keyword "covid" ($r = .78$), as it is not a symptom but a behaviour associated with reducing the spread of the virus. Focusing on symptoms, we can also see that loss of taste (ageusia), anosmia, sneezing, chest pain, and 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. Blue face, headache, anosmia, high temperature, breathing difficulty, chest tightness, nasal congestion, nausea, and appetite loss are some of the symptoms (in that order) that are also impactful from the perspective of online search predictiveness. In addition, recommended behaviours such as isolation, and staying at home are also very relevant.

Finally, we evaluated the performance of the strictly autoregressive (AR-F) model using the $L = 6$ past values of confirmed

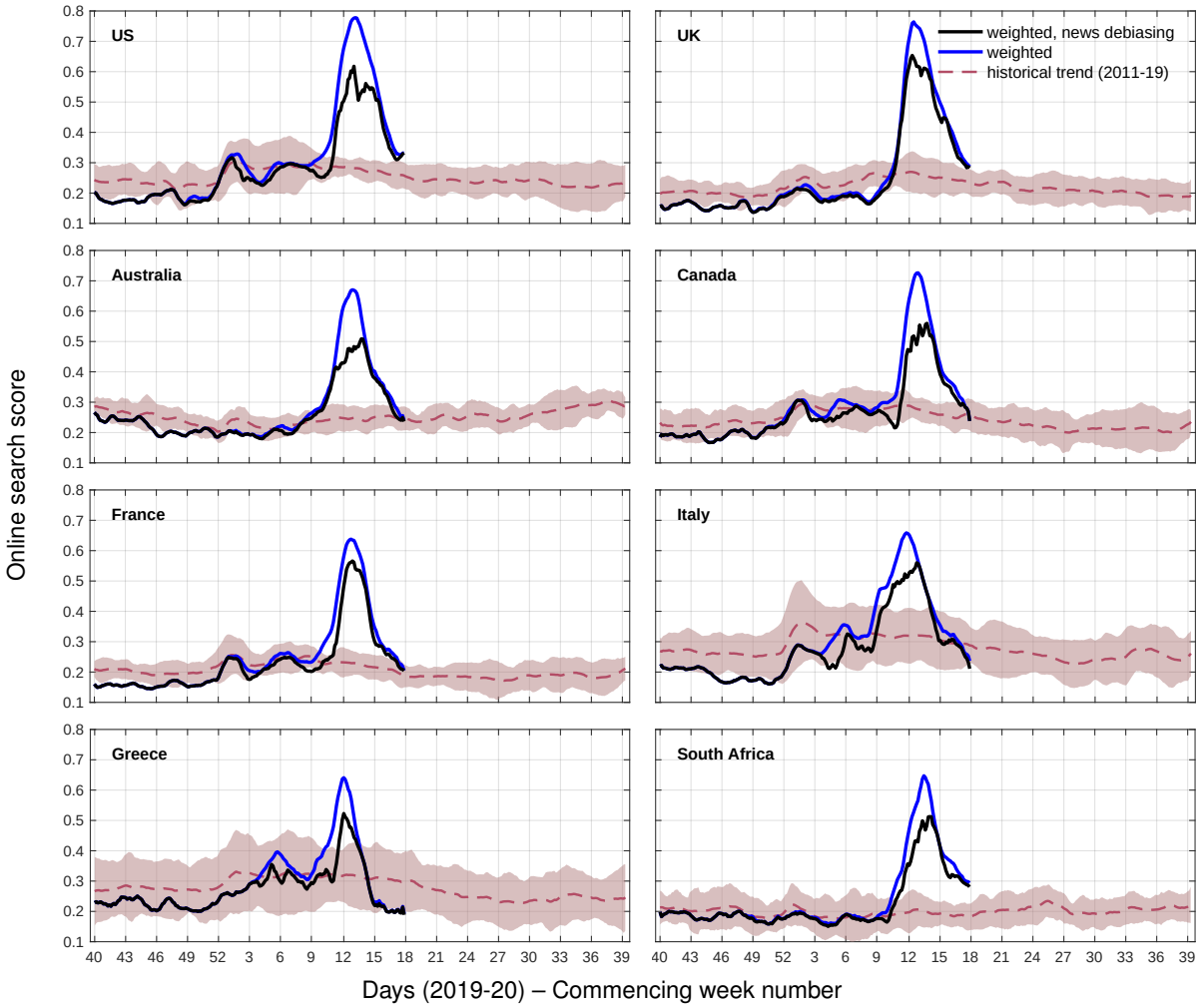

Figure 3. Online search based scores for COVID-19 related symptoms as identified by the NHS FF100 survey, in addition to queries about the symptom of anosmia, and a group of coronavirus-related terms, for 8 countries up to and including April 27, 2020. Query frequencies are weighted by symptom frequency (blue line), and have news media effects minimised (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). All time series are smoothed using a 7-point moving average.

COVID-19 cases (see Methods), and its expanded version that incorporates search data (SAR-F), on 3 forecasting tasks, predicting COVID-19 cases 1, 7, and 14 days ahead. We used a basic persistence model (PER-F) as our baseline. For a D days ahead forecasting task, PER-F uses y_t as the forecasting estimate for the time instance $t+D$ ($\hat{y}_{t+D} = y_t$). We test on the most recent 42 days, retraining models at every time step, and assessing accuracy using the mean absolute error (MAE) between actual and predicted values. Table 1 enumerates these results, including a normalised average (using min-max) across locations, models, and forecasting tasks. We see that in the more challenging forecasting tasks the AR-F and SAR-F models are better than the PER-F baseline, and that the overall best model by a significant margin is SAR-F. This showcases how online search information can contribute in obtaining more accurate forecasts. We also applied these models to conduct forecasting estimates on a daily basis for the coming 2 weeks that at this point in time cannot be evaluated. Figure 6 depicts these results. For most countries in our analysis a decreasing trend is predicted.

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).

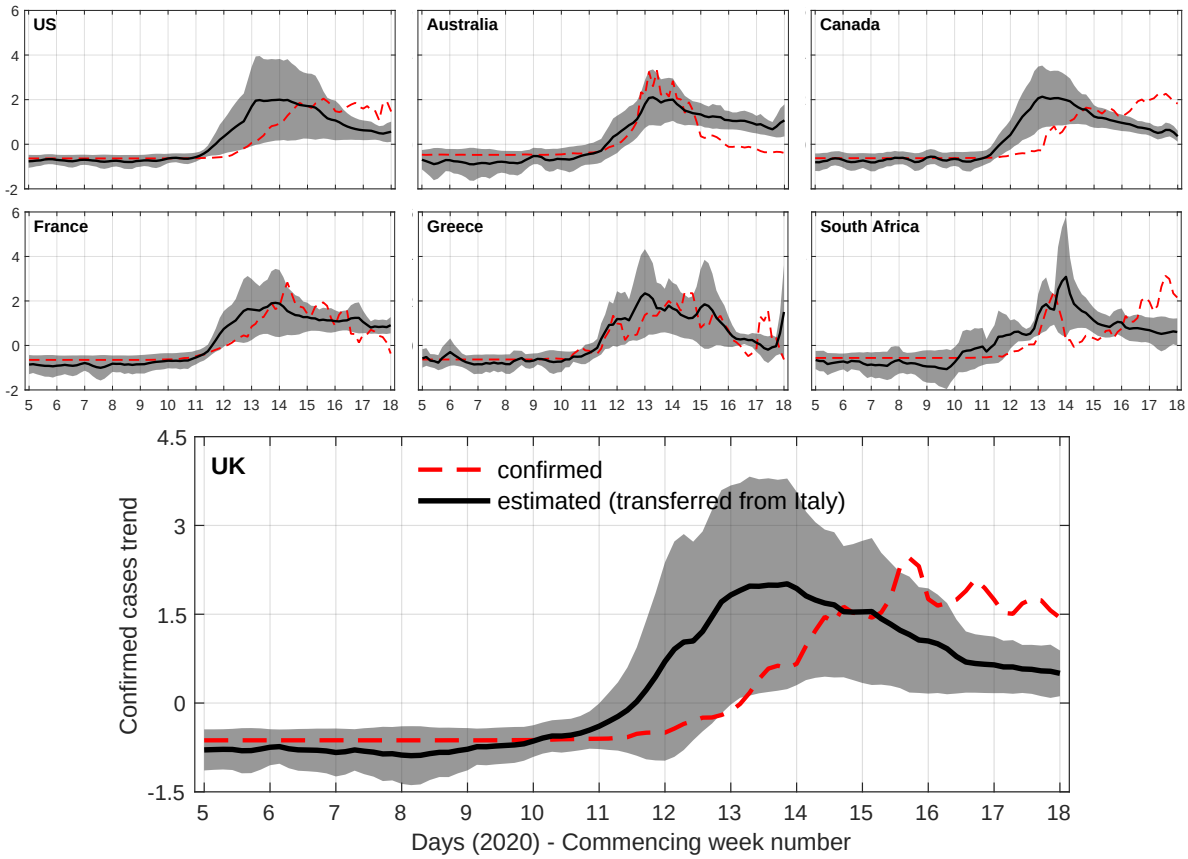

Figure 4. Transfer learning models using data from Italy. The figures show an estimated confirmed COVID-19 cases trend (with confidence intervals) compared to the actual one. Time series are standardised, and smoothed (3-point moving average).

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. 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).

9. 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).

10. Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438 (1969).

11. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. Royal Stat. Soc.: Ser. B* **67**, 301–320 (2005).

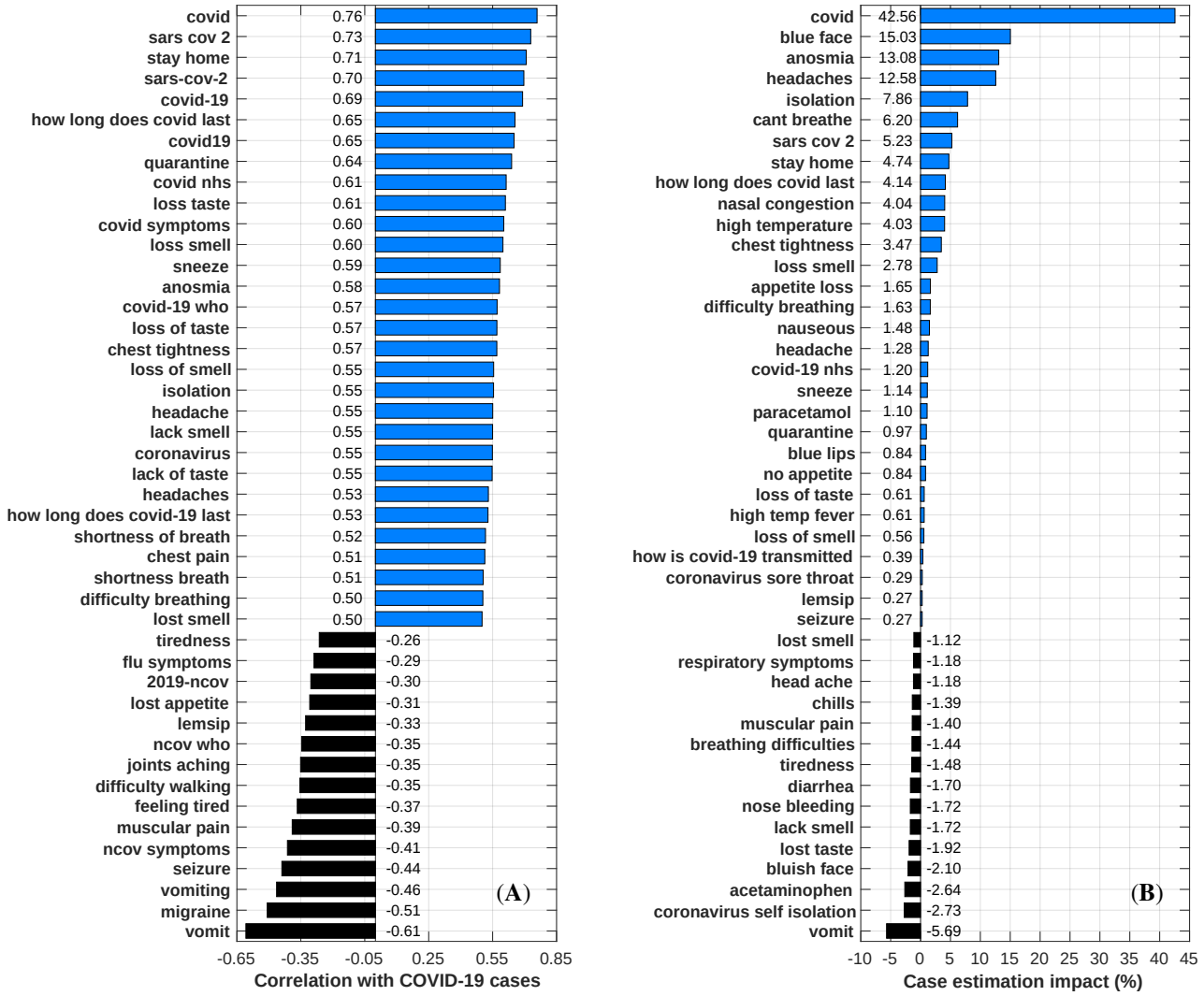

Figure 5. Correlation and regression analysis of search query frequencies against confirmed COVID-19 cases in four countries (US, UK, Australia, and Canada). (**A**) Top-30 positively and top-15 negatively correlated search queries with COVID-19 confirmed cases; (**B**) Top-30 positively and top-15 negatively impactful queries in nowcasting COVID-19 confirmed cases.

12. Lampos, V., Preoţiuc-Pietro, D. & Cohn, T. A user-centric model of voting intention from social media. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, 993–1003 (2013).

13. Lampos, V., Aletras, N., Preoţiuc-Pietro, D. & Cohn, T. Predicting and characterising user impact on twitter. In *Proc. of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 405–413 (2014).

14. Paul, M. J., Dredze, M. & Broniatowski, D. Twitter Improves Influenza Forecasting. *PLoS Curr.* **6** (2014).

15. Zou, B., Lampos, V. & Cox, I. Multi-Task Learning Improves Disease Models from Web Search. In *Proc. of the 27th International Conference on World Wide Web*, 87–96 (2018).

16. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (MIT Press, 2006).

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

Country	1 day ahead			7 days ahead			14 days ahead		
	AR-F	**SAR-F**	**PER-F**	**AR-F**	**SAR-F**	**PER-F**	**AR-F**	**SAR-F**	**PER-F**
UK	1085 (1171)	805 (681)	708 (826)	1293 (1156)	1064 (964)	1143 (1014)	1284 (1094)	906 (756)	1870 (1441)
US	5710 (5052)	4176 (5510)	4390 (5778)	9181 (7542)	4687 (5410)	7471 (4345)	11134 (9484)	6452 (6055)	11743 (8388)
Australia	104 (133)	94 (139)	85 (129)	137 (127)	98 (112)	144 (143)	115 (105)	118 (138)	207 (161)
Canada	337 (374)	324 (331)	215 (231)	345 (378)	299 (284)	350 (309)	513 (425)	308 (310)	540 (416)
Greece	32 (24)	31 (24)	30 (33)	35 (24)	33 (30)	33 (25)	36 (28)	37 (28)	43 (27)
Italy	740 (550)	589 (476)	540 (343)	664 (531)	660 (465)	1074 (833)	956 (702)	865 (722)	2050 (1357)
France	1342 (1105)	1258 (1124)	1182 (1011)	1325 (1103)	1424 (1200)	1443 (1041)	1527 (998)	1483 (1452)	1953 (1298)
South Africa	72 (65)	57 (61)	46 (47)	68 (52)	63 (60)	76 (62)	66 (63)	59 (46)	82 (65)
Norm. mean	.187 (.182)	.161 (.172)	**.143** (.163)	.211 (.188)	**.177** (.171)	.216 (.174)	.239 (.210)	**.192** (.181)	.318 (.234)

Table 1. Average mean absolute error and standard deviation (in parentheses) of forecasting models (1, 7, and 14 days ahead) for daily confirmed COVID-19 cases. The last row contains min-max normalised averages across countries, methods, and forecasting tasks. **AR-F**: autoregressive forecasting using past confirmed cases; **SAR-F**: combined online search and autoregressive forecasting; **PER-F**: persistence model.

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.

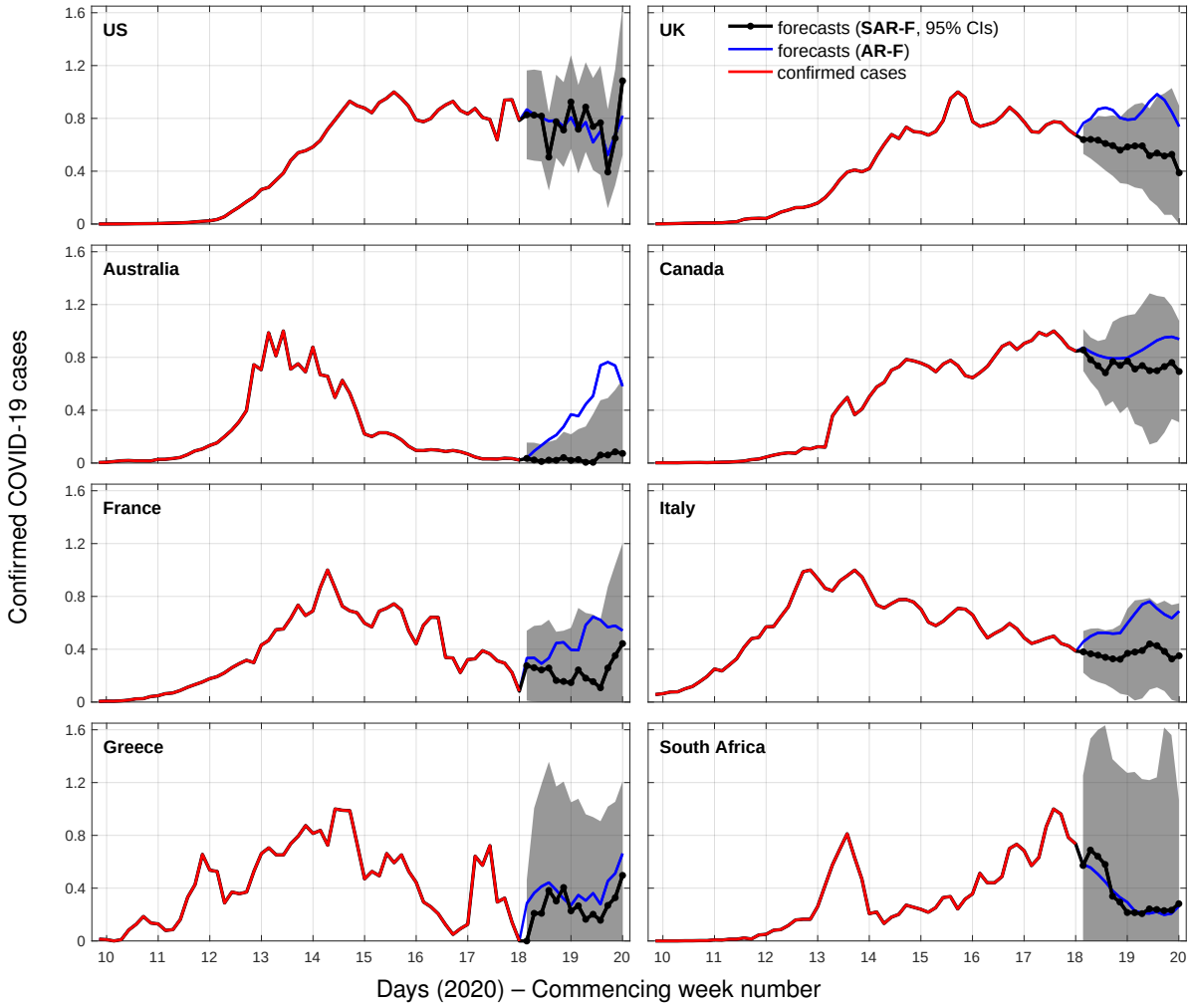

Figure 6. 1 to 14-days ahead forecasting estimates of confirmed COVID-19 cases for 8 countries starting from April 28, 2020. The blue line shows estimates from a strictly autoregressive model (**AR-F**). The black line shows estimates from a model that incorporates online search information (**SAR-F**). The shaded area denotes the corresponding 95% confidence intervals for the latter estimates. For a better visualisation, all values are normalised using min-max solely based on the confirmed cases time series of each country (red line), and are smoothed using a 3-point moving average.