# Tracking COVID-19 using online search

**Vasileios Lampos[1*], Simon Moura[1], Elad Yom-Tov[2], Michael Edelstein[3], Maimuna Majumder[4], Rachel A. McKendry[1], and Ingemar J. Cox[1]**

[1]University College London
[2]Microsoft Research
[3]Public Health England
[4]Harvard Medical School
[*]Corresponding author, v.lampos@ucl.ac.uk

**Disclaimer:** The current version considers data up to and including March 29, 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. To this end, we envision that an updated report will be uploaded at least on a weekly basis. The most up-to-date versions of this report (due to the instant turnaround) can be found at github.com/vlampos/covid-19-online-search.

## Introduction

Online search data is routinely used to monitor the prevalence of infectious diseases, such as influenza[1–4]. Previous work has focused on supervised learning solutions, where *ground truth* data, in the form of historical syndromic surveillance reports, can be used to train machine learning models. However, no sufficient data —in terms of accuracy 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. Recent outcomes have shown that it is possible to transfer an online search based model for influenza-like illness (ILI) from a source to a target country without using ground truth data for the target location[5]. The transferred model's accuracy depends on choosing search queries and their corresponding weights wisely, via a transfer learning methodology, for the target location. 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 may be also driven by concern rather than infections, we devise a preliminary approach that attempts to minimise this part of the signal by incorporating a basic news media coverage metric in association with confirmed COVID-19 cases. Finally, we propose a transfer learning method for mapping supervised COVID-19 models from a country to another, in an effort to transfer knowledge from areas where the disease has a more extended progression. Results are presented for the UK, England, United States of America (US), Canada, Australia, France, Italy, and Greece.

## 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 search query frequencies for a day and a location. Query frequencies are defined as the sum of search sessions that include a target search term divided by the total search sessions for this day and location. We have obtained data from September 30, 2011 to March 29, 2020 for the UK, England, US, Canada, Australia, France, Italy, and Greece. The list of search terms is determined by COVID-related symptoms and keywords. For each country, we mainly used queries in its native language.[1]

**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[2] via national corpora for the UK (93), US (225), Canada

[1]Please note that we are avoiding to mention the exact search queries that we are using to discourage any kind of user search behaviour bias that may invalidate our current approach. These will be releases at a later time. At this stage, please contact us if you want to reproduce our technique.
[2]MediaCloud, mediacloud.org

(79), Australia (61), Italy (178), France (360), and Greece (75), where in the parentheses we state number of media sources considered per country. These counts were collected from September 30, 2019 through March 29, 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.

**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 number of confirmed COVID-19 cases on a daily basis for England is obtained by the corresponding PHE web page.[3] For all the remaining locations we obtain daily confirmed cases data from the European Centre for Disease Prevention and Control (ECDC).[4]

## Methods

**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 version, we also consider two additional groups one referring to an investigated symptom (anosmia or loss of smell), and another that includes 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. We 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 March 29, 2020) and a historical one (from September 30, 2011 to September 29, 2019). The corresponding data sets are denoted by $\mathbf{H} \in \mathbb{R}_{\geq 0}^{d_1 \times k}$ and $\mathbf{X} \in \mathbb{R}_{\geq 0}^{d_2 \times k}$, where $d_1$, $d_2$ represent the different numbers of days in the historical and current 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 ($\mathbf{h} = \mathbf{H}\mathbf{w}$, $\mathbf{x} = \mathbf{X}\mathbf{w}$). 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. Finally, we standardise $\mathbf{x}$ using the mean and standard deviation of the weighted time series of the current season augmented with points from the previous season (2018-19) to cover up for the missing (and potentially important) seasonal components. This is compared to a standardised version of the historical time series and their confidence intervals.

**Minimising the effect of news media.** 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

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

[3] Available at arcgis.com/home/item.html?id=e5fd11150d274bebaaf8fe2a7a2bda11

[4] Available at ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

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. For any other combination of $a_1$ and $a_2$ weights, 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.

**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 search query frequency time series, from one country that has access to historical syndromic surveillance data to another that has not[5]. Here, we adapt this method to transfer a model for COVID-19 from a source country where the 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:

$$s_i = \frac{1}{\sum_{p=1}^{D} \frac{1}{p}} \sum_{p=1}^{D} \frac{x_{i-p+1}}{p}\,, \tag{8}$$

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

We train an elastic net model on data from the source location[6], similarly to previous work on ILI[1,3,7]. In particular, we solve the following optimisation task

$$\arg\min_{\mathbf{w},b} \left( \|\mathbf{y} - \mathbf{S}\mathbf{w} - b\|_2^2 + \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w}\|_2^2 \right)\,, \tag{9}$$

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 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 $\lambda_1$, under the constraint that only sparse solutions compared to the number of training instances are considered (for us to consider models with $\leq \xi$ nonzero weights approx. $2\xi \log(\xi)$ samples are required). All different regression models represented by the columns of $\mathbf{W} \in \mathbb{R}^{n_S \times q}$, and different 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 30 days backwards or forwards) so that the average correlation between search query frequencies in $\mathbf{S}$ and $\mathbf{T}$ are maximised. If no target search query exists for a certain symptom category, we simply use the best correlated query from all target queries available (irrespectively of the symptom category) as its mapping. After this

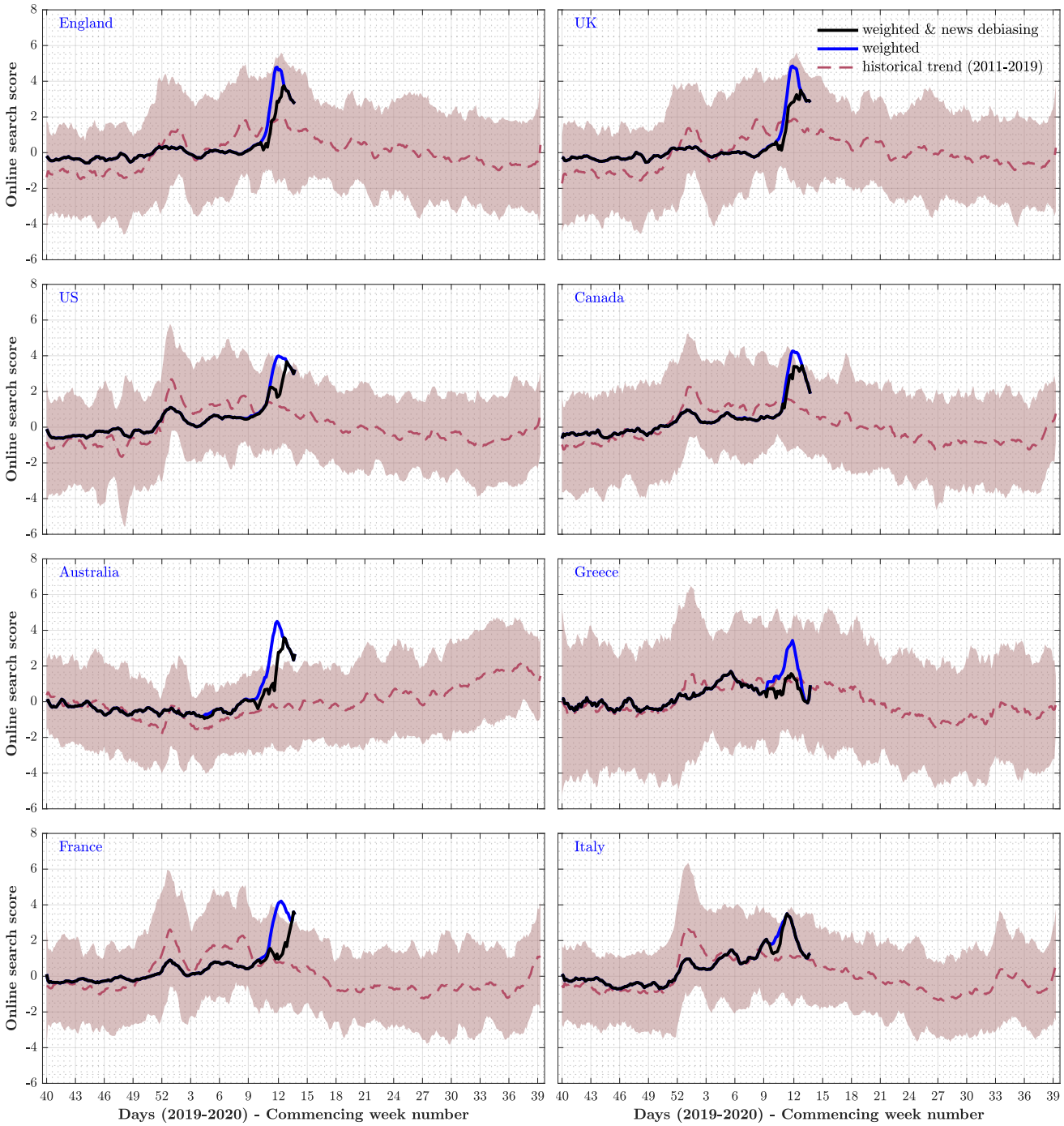


**Figure 1.** Standardised online search based scores for COVID-19 related symptoms as identified by the NHS FF100 survey for 8 nations up to and including March 29, 2020. Query frequencies are weighted by symptom frequency as described in Methods (blue line). We have also included estimates after minimising the news media effect using data from PHE, ECDC, and 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.

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 $\mathbf{Y} \in \mathbb{R}^{n_S \times q}$:

$$\mathbf{Y} = \mathbf{Z}_S \mathbf{W} + \mathbf{b}\,. \tag{10}$$

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.

## Results

The current online search based scores for COVID-19 in 8 nations are depicted in Figures 1 and 2 (data up to March 29, 2020). For a better visualisation, all time series are smoothed using a 7-point moving average, 3 days prior and after each point. Figure 1 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 2 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 perhaps at an earlier time point) people who experience COVID-19 related symptoms might search about this 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 at some points or periods in time 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. 2) slightly increase the maximum scores per country.

Looking at the scores where we have attempted to minimise the effect of news (black lines), and at the same time introduce some form of supervision in conjunction with the daily number of confirmed cases per country (note that we look back at the previous $N = 10$ days to determine this), we observe more conservative estimates in most locations, including a recent drop or an altered trend (e.g. increasing vs. decreasing) in some of them. Notably, the caveat of this approach is that it relies on the existence of a representative population sample based on the number and distribution of COVID-19 tests conducted in each country. If the reported daily number of COVID-19 cases is not a representative proportion, then variable $d$ (confirmed number of infections) in Eq. 4 is not reliable, and the regression task becomes ill-posed, together with any interpretation of the derived weights.

Finally, Figure 3 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. An interesting observation while implementing the transfer learning model was that the search query frequency data for Italy were best correlated with other countries after shifting them by a number of days; 1 for Canada, 7 for the UK, US, Australia, and France, 14 for England, and $-3$ days for Greece. This indicates that in most occasions Italy is, indeed, in front by a few days at least in terms of user search behaviour. In total, we learn and transfer 35,520 elastic net models that select, by assigning a nonzero weight, from 2 to 18 search queries. Interestingly, the mapped trends correspond sufficiently well to confirmed cases data in most countries considering the fact that they lack supervision at the target locations. The same caveat, explained in the previous paragraph, applies for this analysis as well.

## References


1. Lampos, V., Miller, A. C., Crossan, S. & Stefansen, C. Advances in nowcasting influenza-like illness rates using search query logs. *Sci. Rep.* **5** (2015).
2. Yang, S., Santillana, M. & Kou, S. C. Accurate Estimation of Influenza Epidemics using Google Search Data via ARGO. *PNAS* **112**, 14473–14478 (2015).
3. 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).
4. 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).
5. Zou, B., Lampos, V. & Cox, I. J. Transfer Learning for Unsupervised Influenza-like IllnessModels from Online Search Data. In *Proc. of the 28th International Conference on World Wide Web*, 2505–2516 (2019).
6. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. Royal Stat. Soc.: Ser. B* **67**, 301–320 (2005).

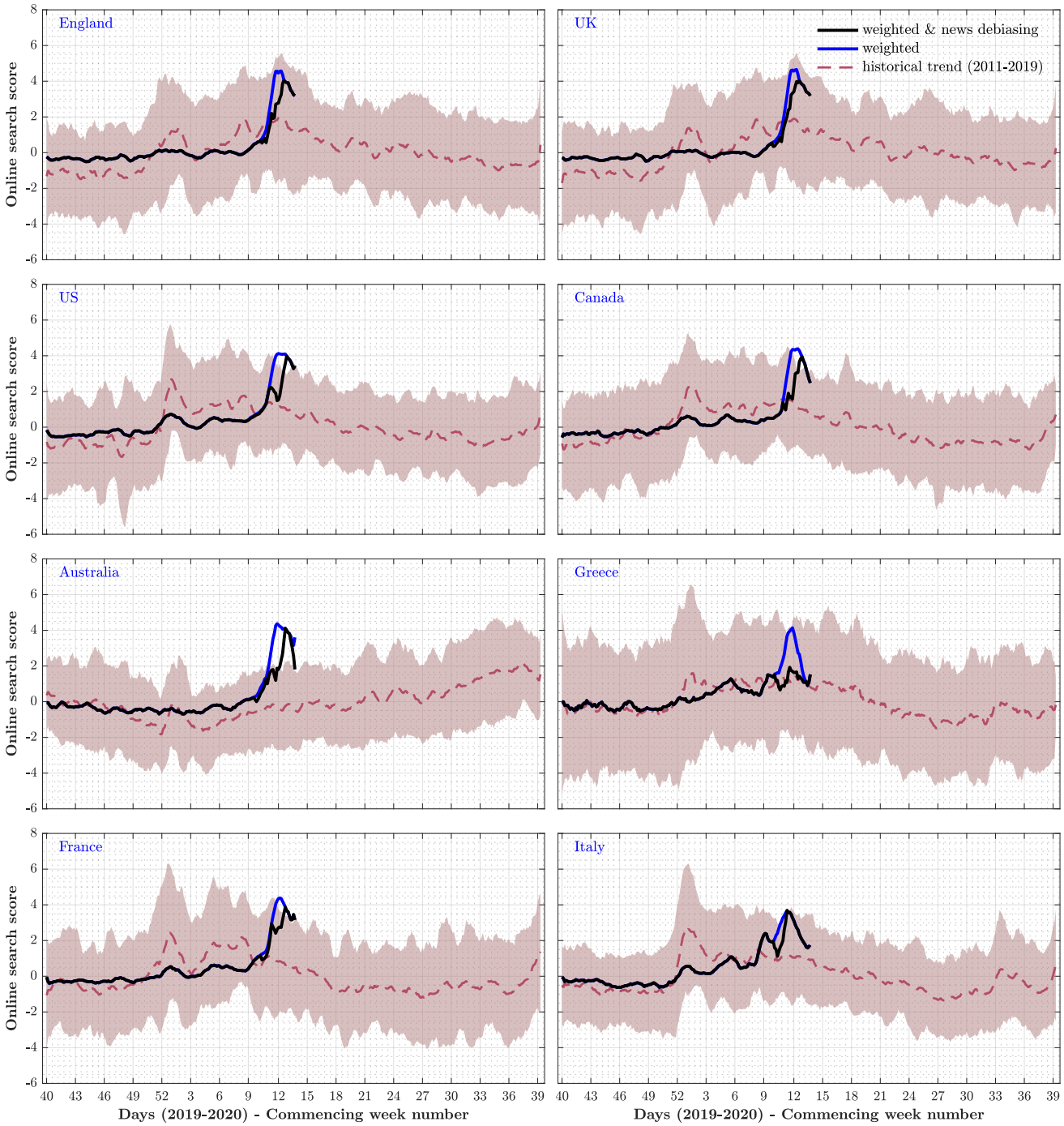


**Figure 2.** Standardised online search based scores for COVID-19 related symptoms as identified by the NHS FF100 survey, in addition to queries about the symptom of "anosmia", and a group of coronavirus-related terms, for 8 nations up to and including March 29, 2020. Query frequencies are weighted by symptom frequency as described in Methods (blue line). We have also included estimates after minimising the news media effect using data from PHE, ECDC, and 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.

**7.** 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).

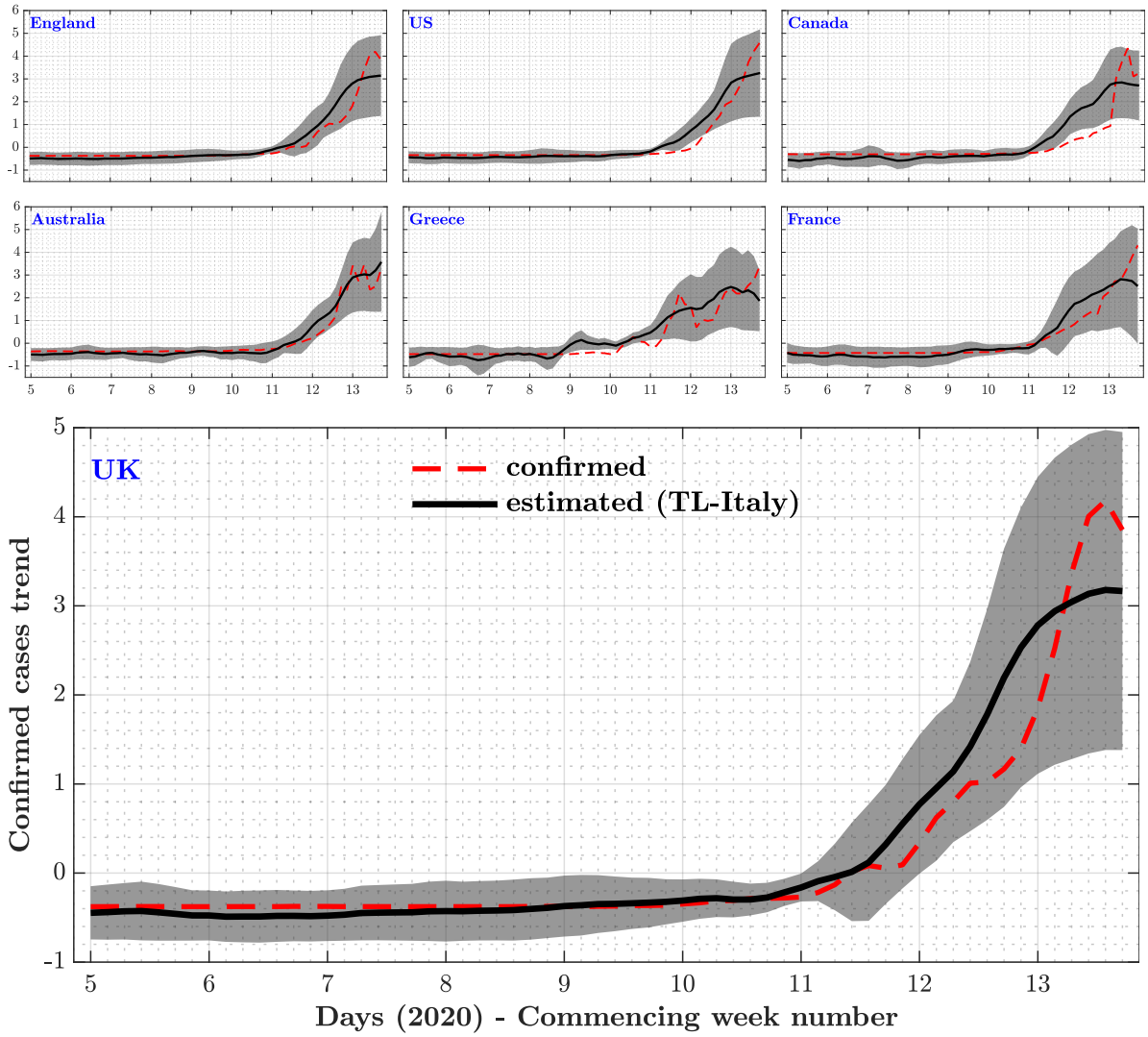


**Figure 3.** Transferring a supervised model for Italy to other countries in our analysis. The figures show an estimated confirmed cases trend (with confidence intervals) for all locations in our analysis (minus Italy) compared to the recorded confirmed cases as reported by PHE and the ECDC. Plot lines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

## 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

**Note:** The author contribution statement will be updated as this research project unfolds.

V.L. conceived this research, formed the majority of the data sets, developed the methods, ran the experiments, and wrote the manuscript. S.M. provided a translation of search query groups to French. M.M. provided news coverage data that were used to minimise the effect of news media in our models. 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.