Modelling infectious diseases using online search activity

Vasileios Lampos

Computer Science, UCL

lampos.net

Presentation structure

A. **Nowcasting** flu prevalence using web search activity

- Lampos, Miller, Crossan, Stefansen. *Advances in nowcasting influenza-like illness rates using search query logs*. Scientific Reports **5** (12760), 2015. doi:10.1038/srep12760
- Lampos, Zou, Cox. *Enhancing feature selection using word embeddings: The case of flu surveillance.* WWW '17, pp. 695-704, 2017. doi:10.1145/3038912.3052622

B. **Transferring** a disease model from one country to another using web search activity

- Zou, Lampos, Cox. *Transfer learning for unsupervised influenza-like illness models from online search data*. WWW '19, pp. 2505-2516, 2019. doi:10.1145/3308558.3313477

C. Modelling **COVID-19** prevalence using web search activity

- Lampos *et al*. *Tracking COVID-19 using online search*. npj Digital Medicine **4** (17), 2021. doi:10.1038/s41746-021-00384-w

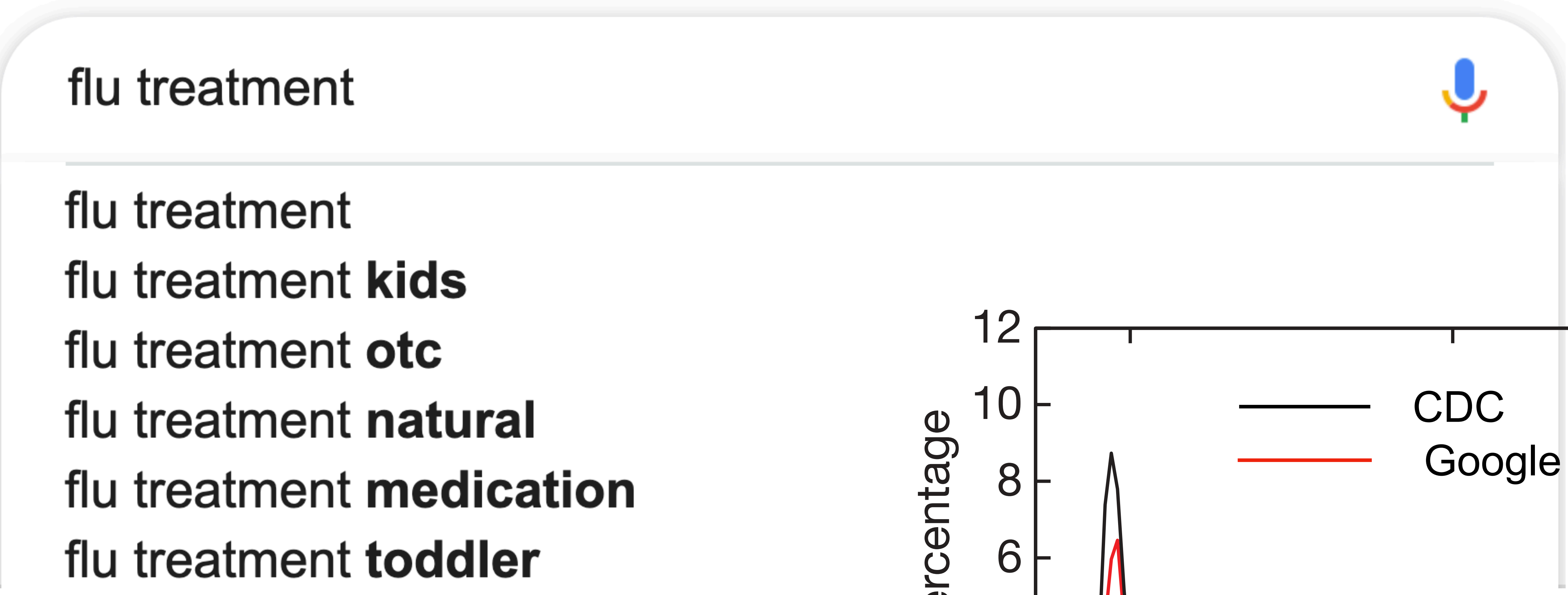
D. Advanced models (*neural network architectures*) for disease **forecasting**

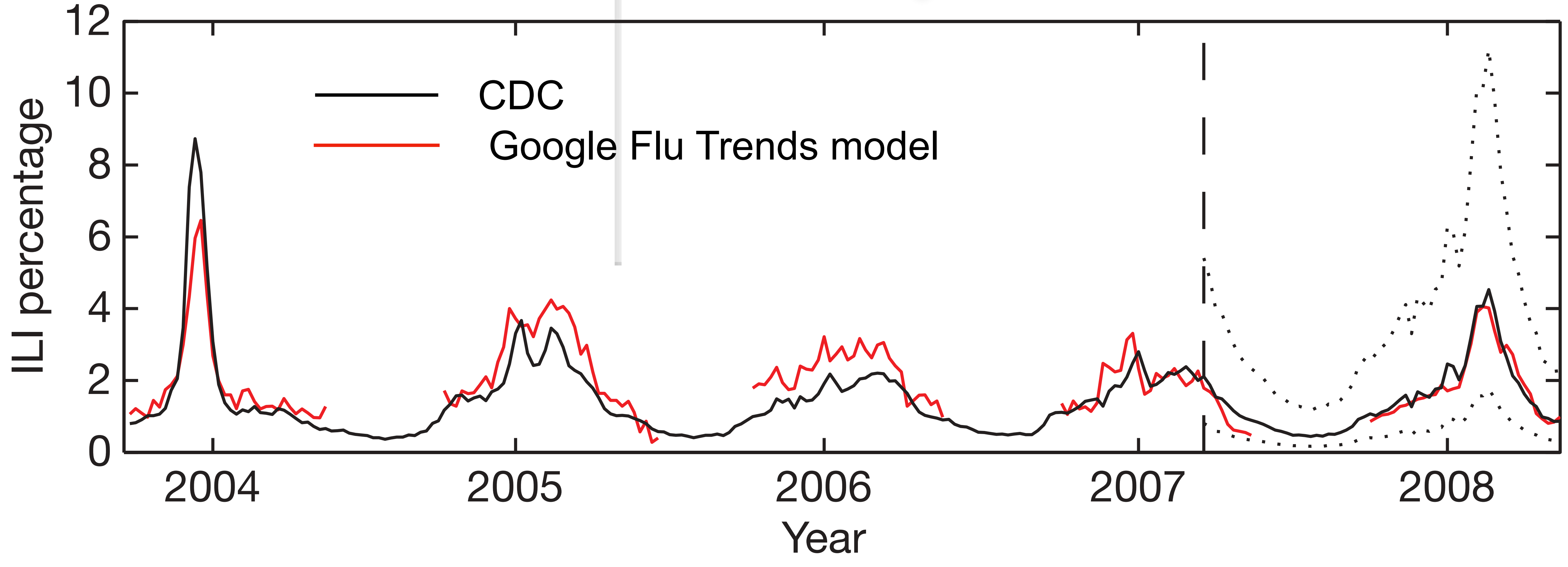
- Morris, Hayes, Cox, Lampos. *Neural network models for influenza forecasting with associated uncertainty using Web search activity trends.* PLOS Computational Biology **19** (8), 2023. doi.org/10.1371/journal.pcbi.1011392

Part A

Estimating flu prevalence using web search activity

From web searches to influenza (*flu*) rates





Eysenbach (2006), *AMIA*; Polgreen *et al.* (2008), *Clin. Infect. Dis.*; Ginsberg *et al.* (2009), *Nature*

Why estimate disease rates from web search?

- **Complements** conventional syndromic surveillance systems
  - larger *cohort*
  - broader *demographic coverage*
  - more granular *geographic coverage*
  - not affected by *closure days* (weekends, holidays)
  - *timeliness*
  - *lower cost*
- Applicable to locations that **lack** an established health surveillance infrastructure
- Track **novel** infectious diseases

Conventional (*traditional*) syndromic surveillance methods: disease prevalence, i.e. the % of infected people in a population, is determined via doctor (GP) visits and other related indicators, such as laboratory-confirmed infections, associated hospitalisations or deaths.

Wagner *et al.* (2018), *Sci. Rep.*; Budd *et al.* (2020), *Nat. Med.*

Google Flu Trends (GFT) — *discontinued*

google.org **Flu Trends**

Language: English (United States)

Google.org home

Dengue Trends

Flu Trends

Home

Select country/regior

How does this work?

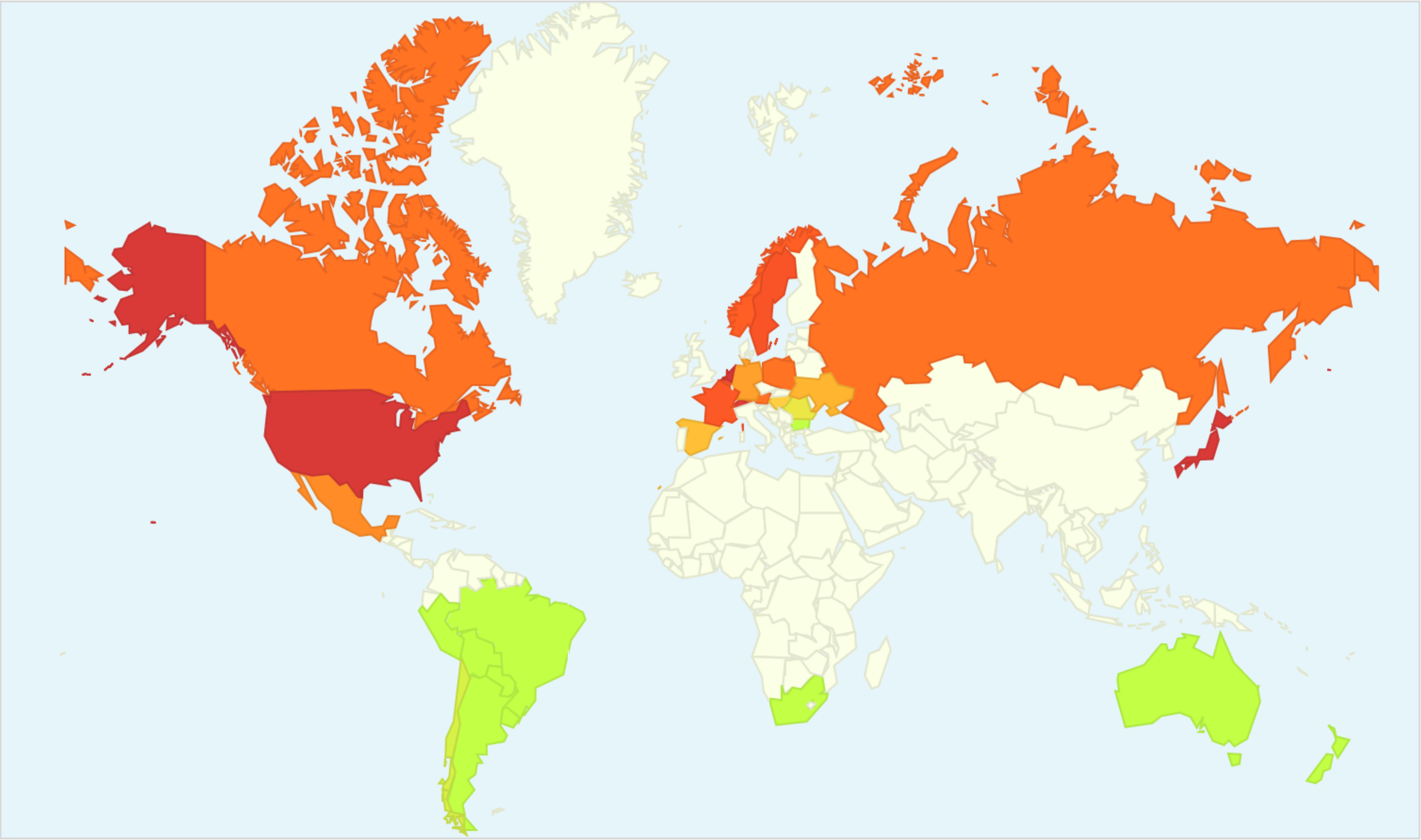
FAQ

Flu activity

- Intense
- High
- Moderate
- Low
- Minimal

Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »



Ginsberg *et al.* (2009), *Nature*

Google Flu Trends (GFT) — *regression function*

$$\text{logit}(P) = \beta_0 + \beta_1 \times \text{logit}(Q) + \epsilon$$

P : percentage of doctor visits due to influnza-like illness (ILI)

Q : aggregate frequency of a set of automatically selected search queries

β_0 : regression intercept (bias)

β_1 : regression weight (univariate regression)

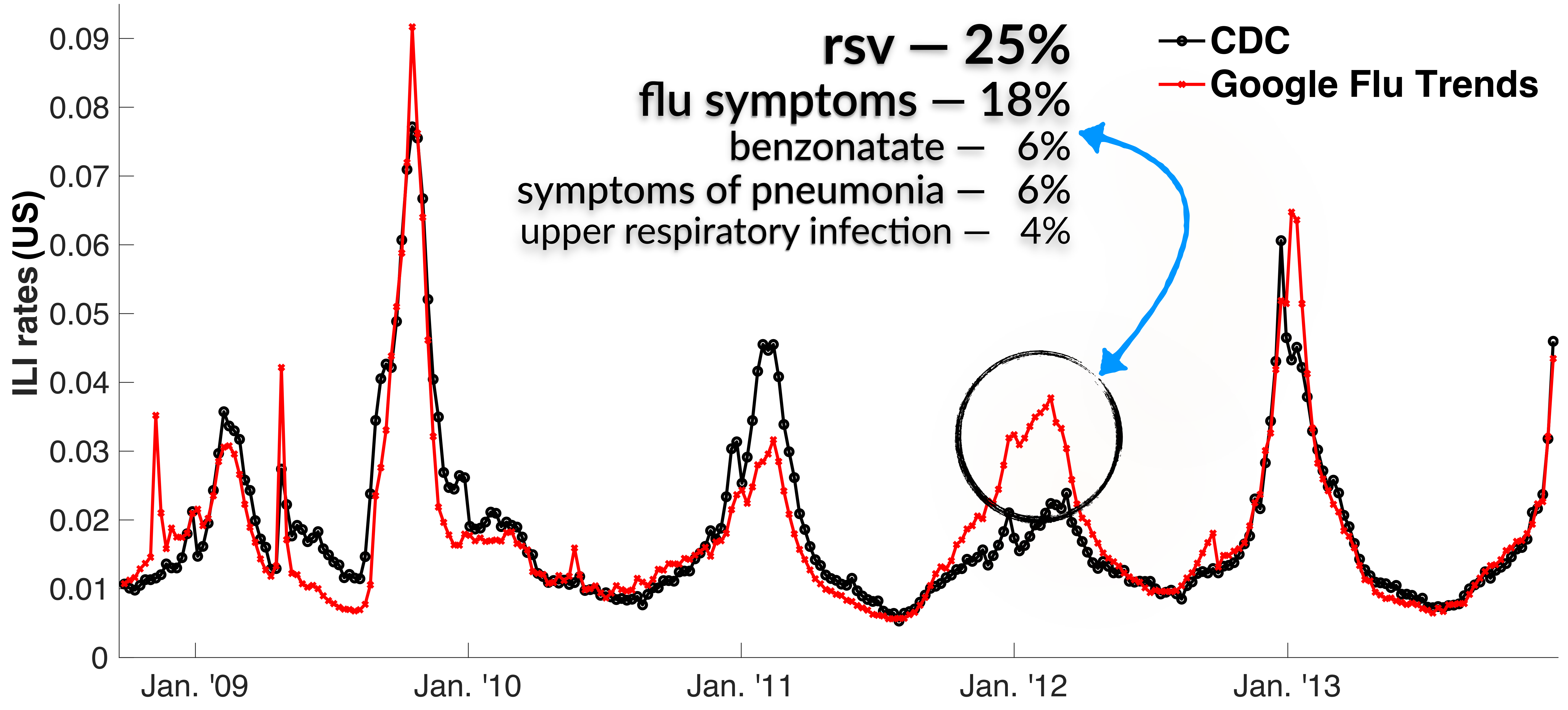
ϵ : independent, zero-centered noise

Ginsberg *et al.* (2009), *Nature*

Main issue

What if some of the selected queries are spurious or, in general, relate differently to flu rates compared to other selected search queries? This model makes a very naïve assumption.

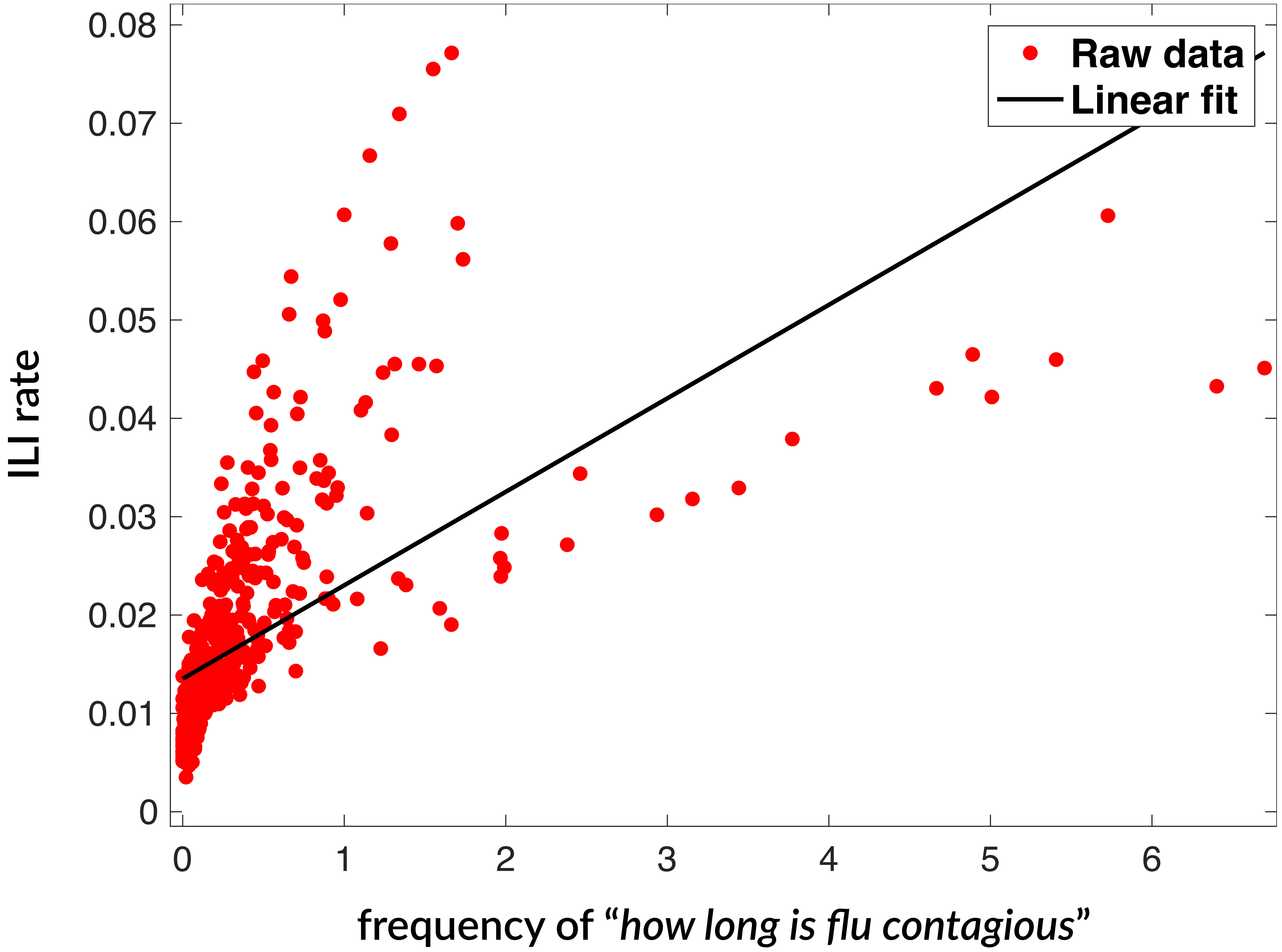
Google Flu Trends (GFT) — *shortcomings*



Lampos *et al.* (2015), *Sci. Rep.*

In the original paper (Ginsberg *et al.*, 2009), the GFT model was "evaluated" on just ~1 flu season! ***That is not a proper evaluation.***

Web search frequencies & flu rates: a *nonlinear* relationship



- Not all search queries have a linear (*or the same*) relationship with flu rates
- Example of a bi-modal relationship

Lampos *et al.* (2015), *Sci. Rep.*

Multivariate Gaussian Process (GP) kernels on search query clusters

Composite Gaussian Process (GP) kernel

$$k(\mathbf{x},\mathbf{x}') = \left(\sum_{i=1}^{C} k_{\mathrm{SE}}\left(\mathbf{c}_i,\mathbf{c}_i'\right)\right) + \sigma_{\mathrm{n}}^2 \cdot \delta(\mathbf{x},\mathbf{x}')$$

NB: *Queries are selected based on their **correlation** with ILI rates in the training data and an **elastic net** regression function*

$\mathbf{x},\mathbf{x}' \in \mathbb{R}^m_{\geq 0}$, where m is the number of search queries we consider

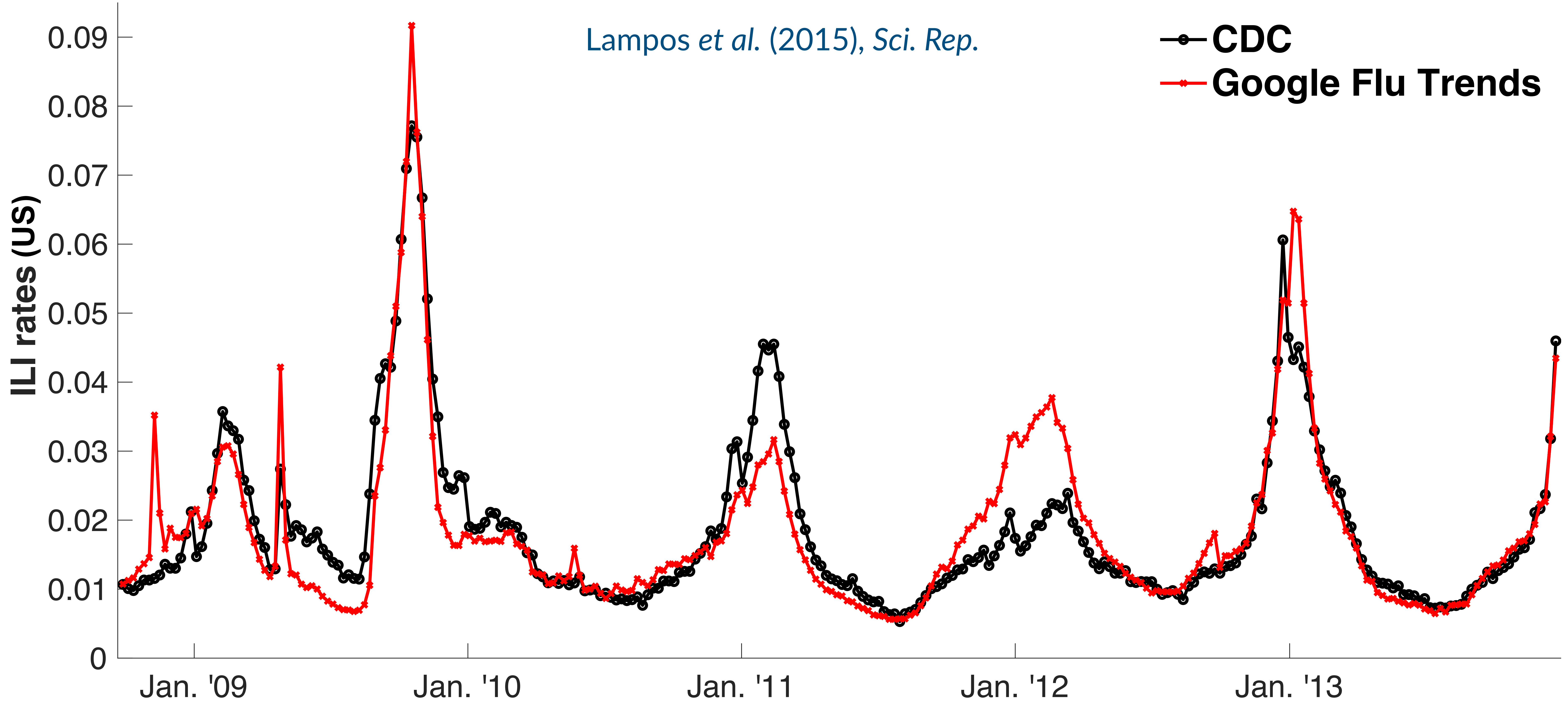
$\mathbf{c}_i,\mathbf{c}_i' \in \mathbb{R}^z_{\geq 0}$, $z < m$, C query clusters based on frequency time series

Squared Exponential (SE) kernel

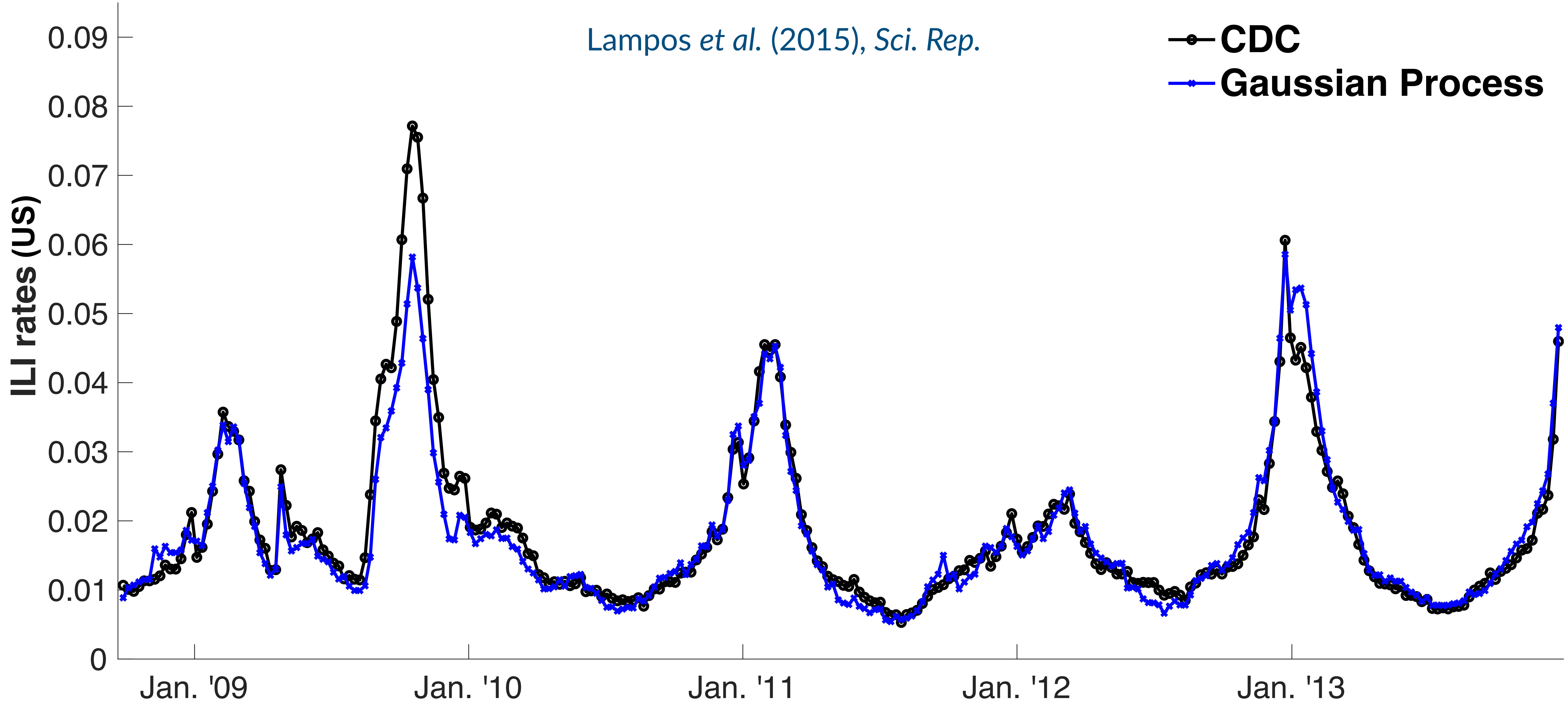
$$k_{\mathrm{SE}}(\mathbf{c}_i,\mathbf{c}_i') = \sigma^2 \exp\left(-\frac{\|\mathbf{c}_i - \mathbf{c}_i'\|_2^2}{2\ell^2}\right)$$

Lampos *et al.* (2015), *Sci. Rep.*;
Rasmussen, Williams (2006), *MIT Press*

Modelling ILI rates with Gaussian Process (GP) kernels

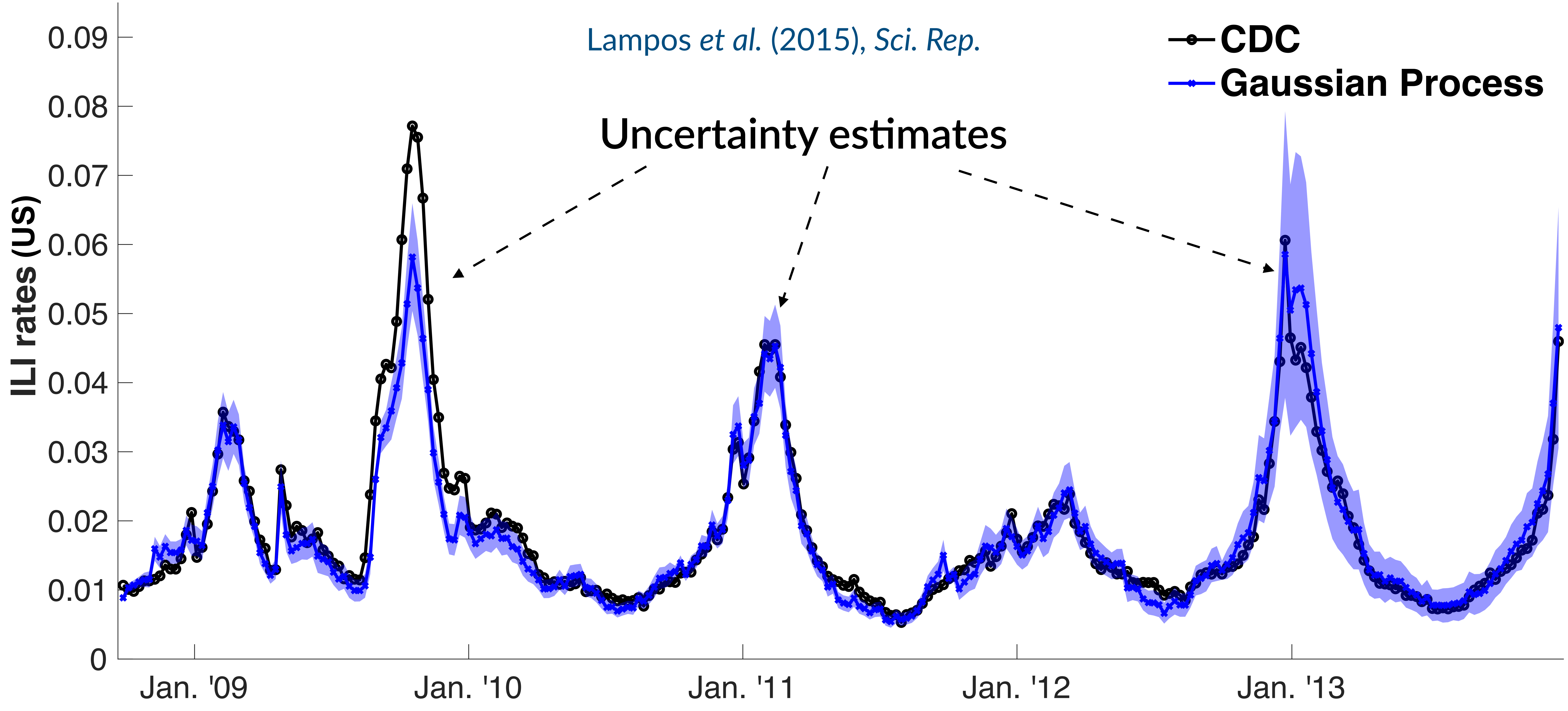


Modelling ILI rates with Gaussian Process (GP) kernels



- **42%** mean absolute error reduction compared to Google Flu Trends
- **.95** bivariate correlation (*previously* .89) with CDC rates

Modelling ILI rates with Gaussian Process (GP) kernels



- **42%** mean absolute error reduction compared to Google Flu Trends
- **.95** bivariate correlation (*previously* .89) with CDC rates

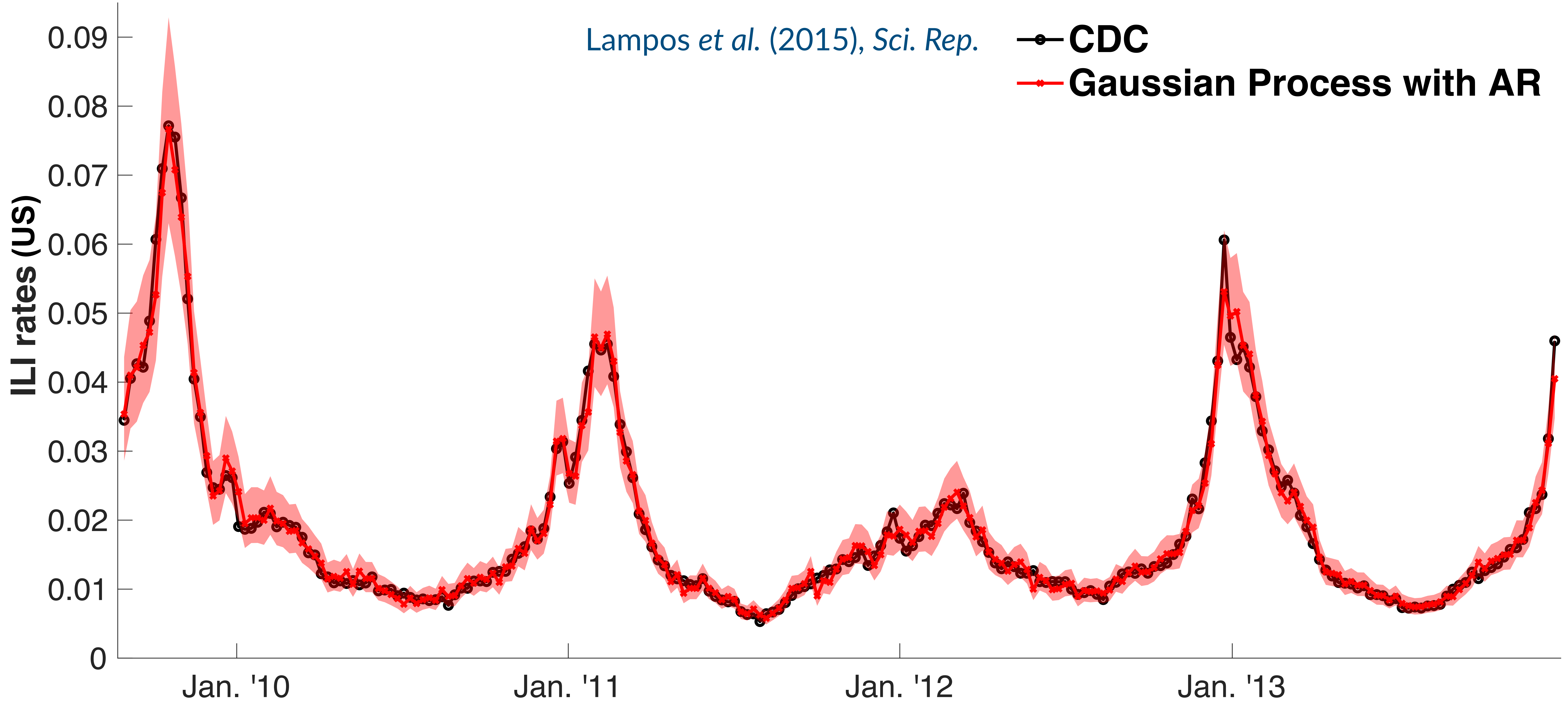
Autoregression (AR) with SARIMAX

$$y_t = \underbrace{\sum_{i=1}^{p} \phi_i y_{t-d} + \sum_{i=1}^{J} \omega_i y_{t-52-i}}_{\text{AR and seasonal AR}} + \underbrace{\sum_{i=1}^{q} \theta_i \epsilon_{t-d} + \sum_{i=1}^{K} \nu_i \epsilon_{t-52-i}}_{\text{MA and seasonal MA}} + \underbrace{\sum_{i=1}^{D} w_i h_{t,i}}_{\text{GP estimates}} + \epsilon_t$$

- SARIMAX: Seasonal AutoRegressive Integrated Moving Average with eXogenous variables
- d weeks delay in including past ILI rates as reported by CDC
- Choose model parameters based on the Akaike Information Criterion (AIC)
  - sometimes past seasons are helpful, but not always
  - the most important piece of information is the GP estimate for the ILI rate ***based on web search query frequencies***

Lampos *et al.* (2015), *Sci. Rep.*

Modelling ILI rates with Gaussian Process (GP) kernels & SARIMAX



- Incorporating historical CDC estimates into an autoregression (AR) using SARIMAX
- 27% MAE reduction compared to GFT with AR, 52% over the GP model without AR
- .99 bivariate correlation with CDC

Feature selection — *which search queries to use?*

- **Feature selection** was based on a **temporal relationship**
  - Is this sufficient? No / not always
- **Spurious search queries** such as “*NBA injury report*” or “*muscle building supplements*” were still included in the selection
  - query clustering: some guarantees for different treatment, but needs a more complex regression model
- Introduce a query *filter* based on **distributional semantics** using word embeddings
- **Hybrid combination** of this with previous feature selection regimes

Lampos *et al.* (2015), *Sci. Rep.*; Lampos, Zou, Cox (2017), *WWW ‘17*

Query selection based on distributional semantics

$$\operatorname{sim}\left(q, \mathbb{C}\right) = \frac{\sum_{i=1}^{P} \cos\left(\mathbf{e}_q, \mathbf{e}_{p_i}\right)}{\sum_{j=1}^{N} \cos\left(\mathbf{e}_q, \mathbf{e}_{n_j}\right) + \gamma}$$

$\mathbf{e}_{(\cdot)}$: embedding vector ***trained on Twitter data***

$\mathbb{C} = \left\{\mathbb{C}_P, \mathbb{C}_N\right\}$ – a **concept** about influenza

$\mathbb{C}_P$: *n*-grams of a **positive** context for concept $\mathbb{C}$

$\mathbb{C}_N$: *n*-grams of a **negative** context for concept $\mathbb{C}$

$\theta = \cos\left(\,\cdot\,\right) \rightarrow \in [0,1]$ via $(\theta + 1)/2$ *to avoid negative components*

$\gamma \in \mathbb{R}_{>0}$ *to avoid, in theory, division by 0*

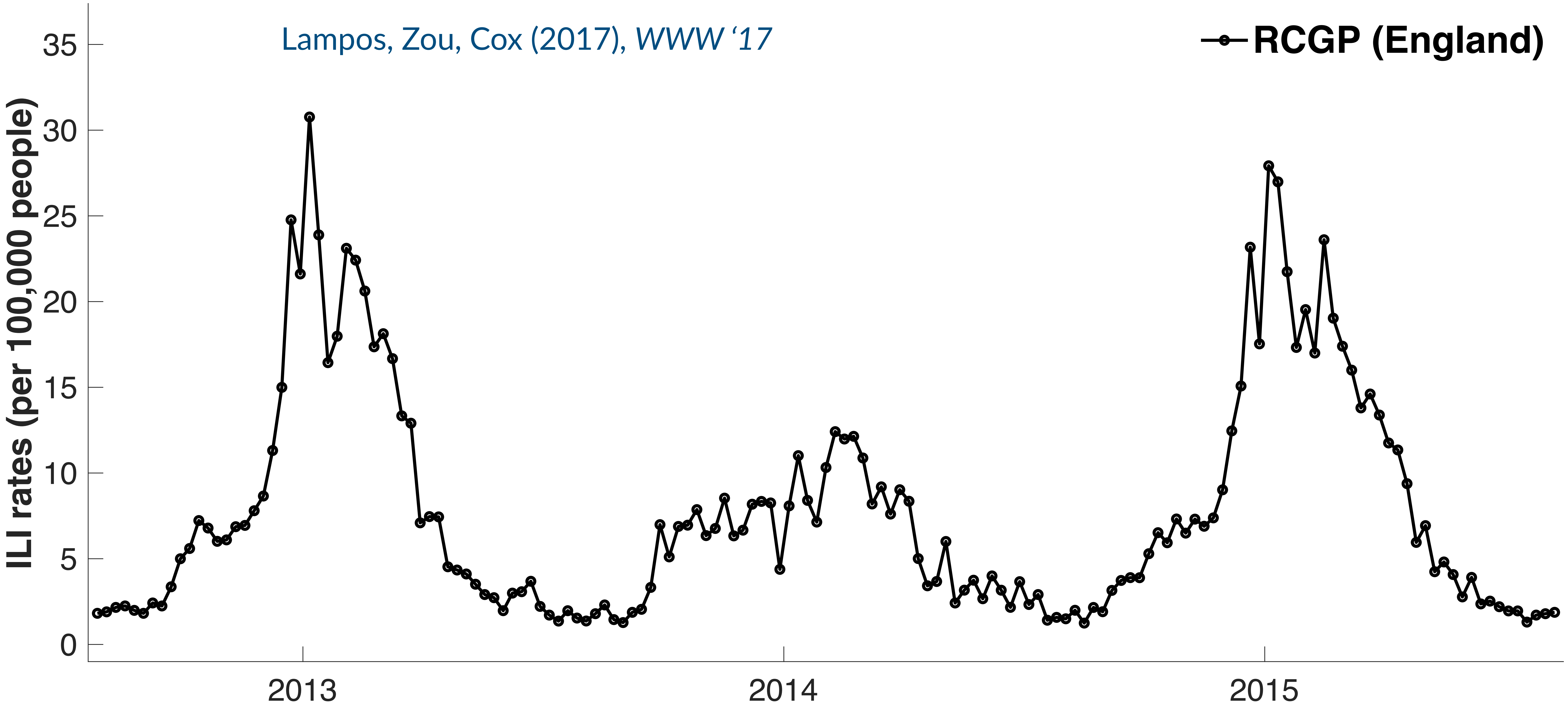
Lampos, Zou, Cox (2017), *WWW '17*;
Levy, Goldberg (2014), *CoNLL '14*

Query selection based on distributional semantics

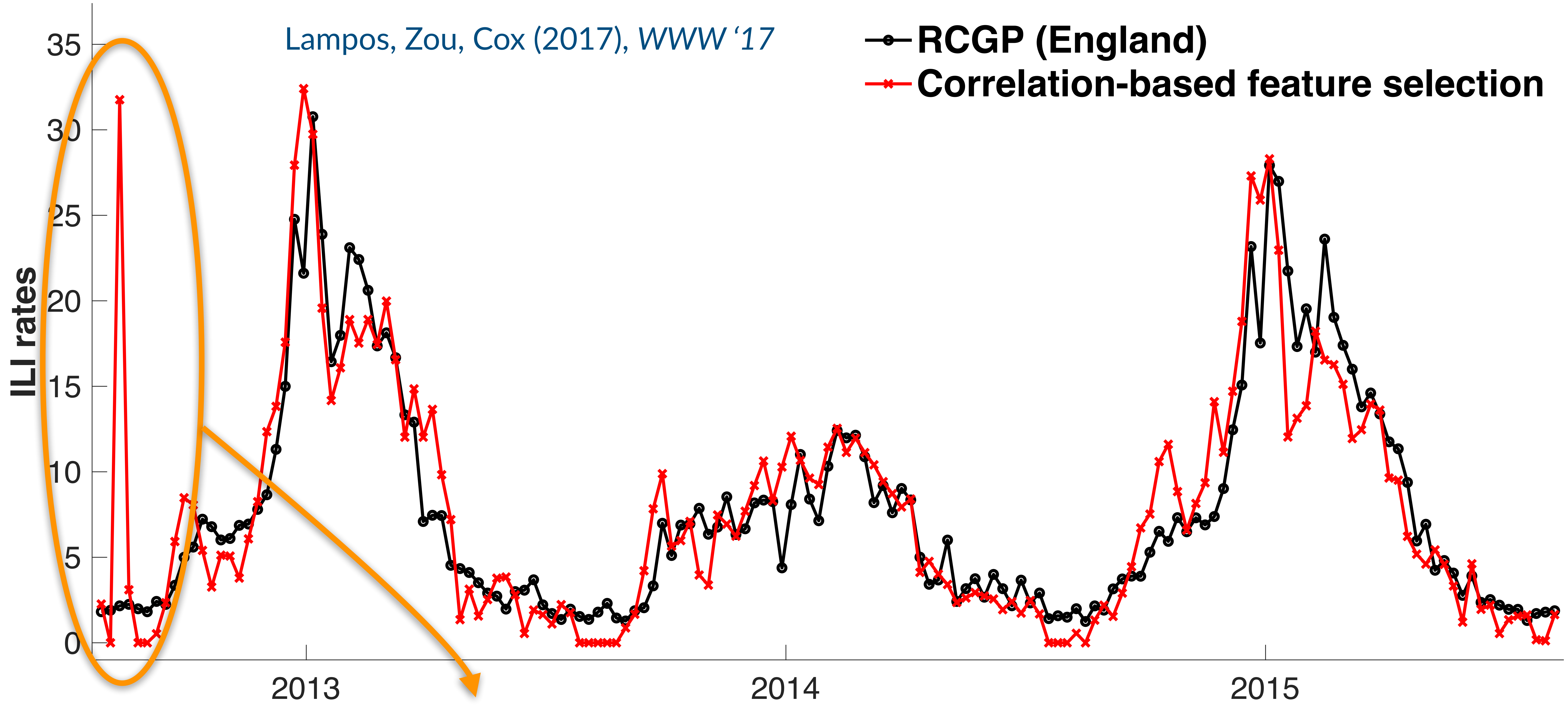
Positive context	Negative context	Most similar queries
#flu fever flu flu medicine GP hospital	Bieber ebola Wikipedia	*"cold flu medicine"* *"flu aches"* *"cold and flu"* *"cold flu symptoms"* *"colds and flu"*
flu flu GP flu hospital flu medicine	ebola Wikipedia	*"flu aches"* *"flu"* *"colds and flu"* *"cold and flu"* *"cold flu medicine"*

Lampos, Zou, Cox (2017), *WWW '17*

Feature selection based on *correlation* and *regularised regression*



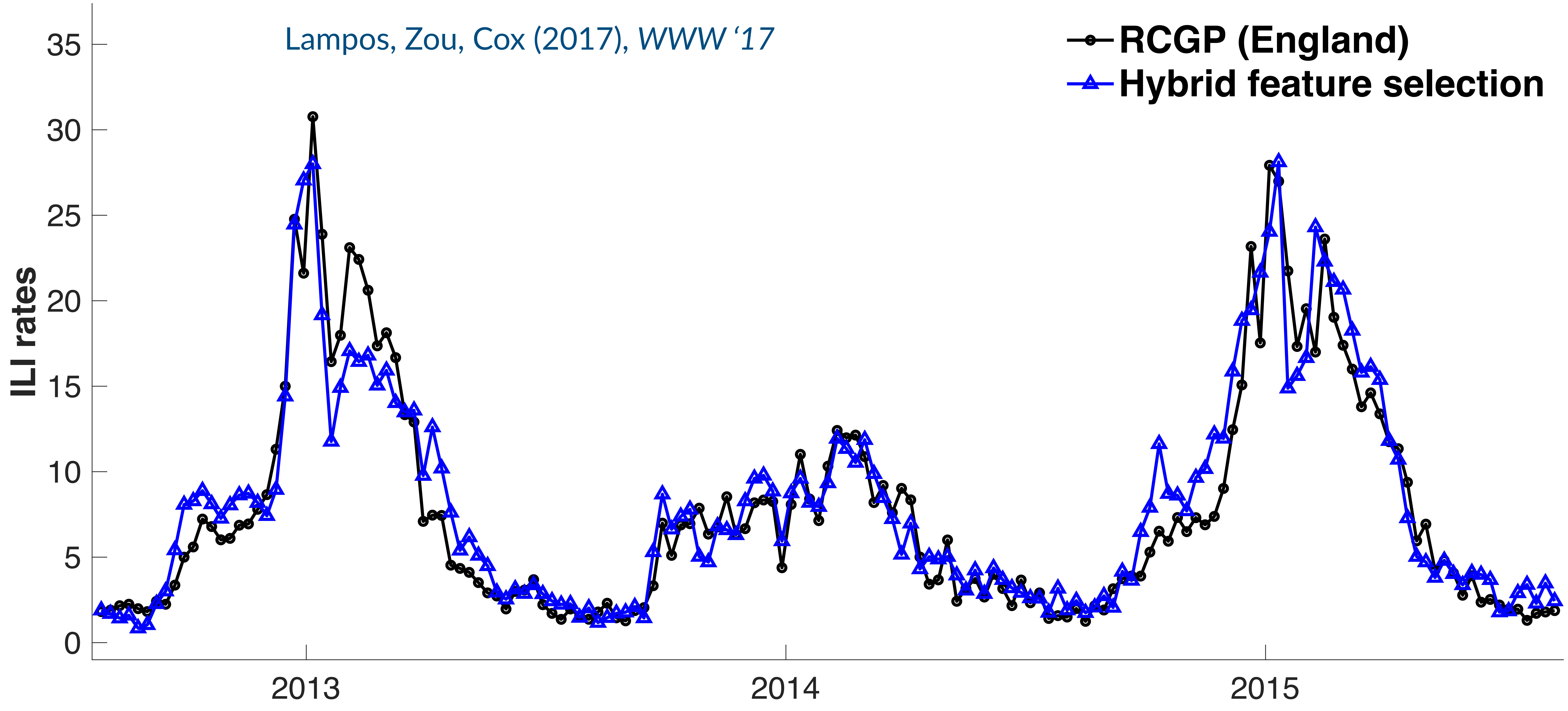
Feature selection based on *correlation* and *regularised regression*



Examples of **problematic** query selections

prof. *surname*: 70%
name surname: 27%
heating oil: 21%
name surname recipes: 21%
blood game: 12.3%
swine flu vaccine side effects: 7.2%

Hybrid feature selection: *distributional semantics* and *correlation*

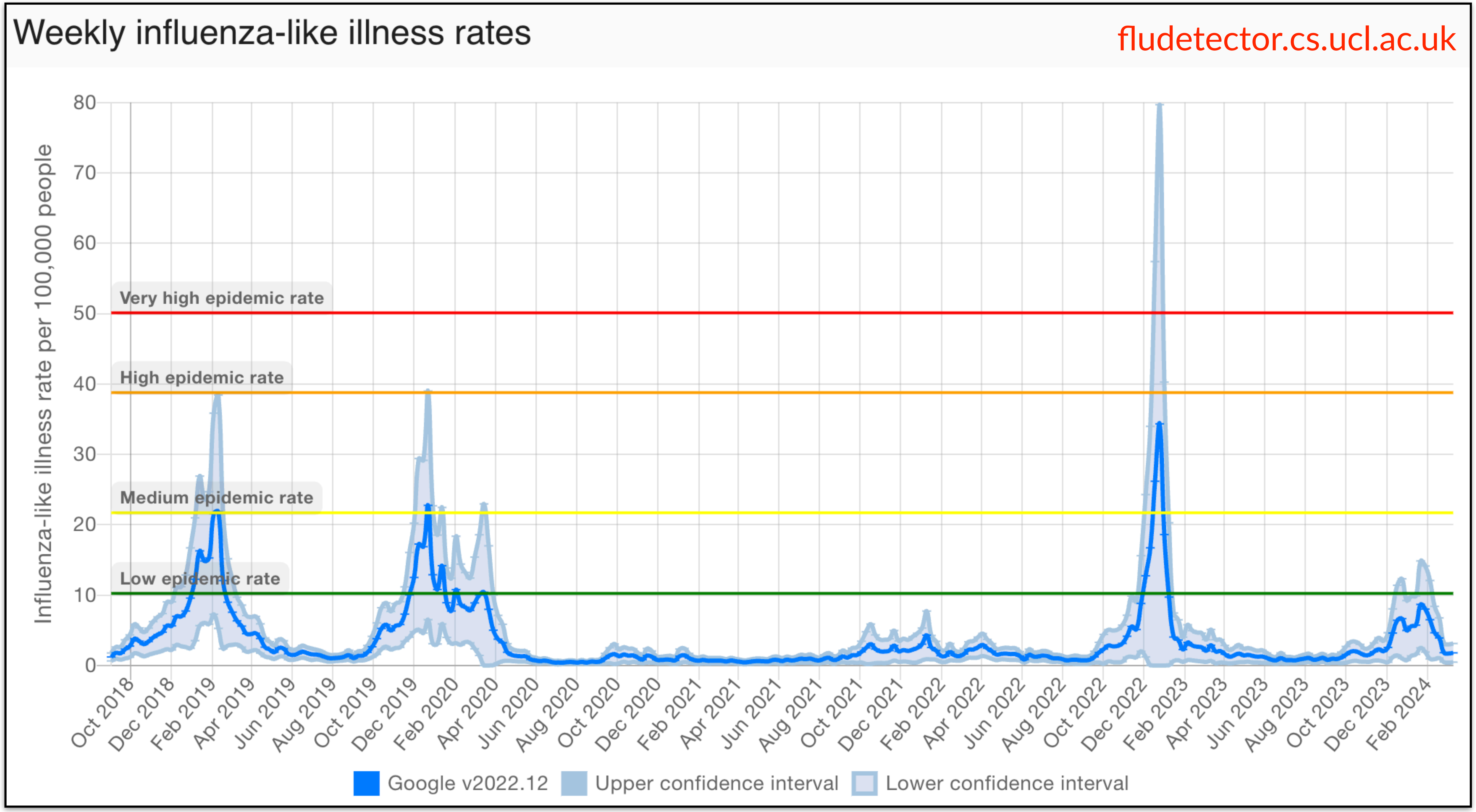


- **12.3%** accuracy improvement in terms of mean absolute error
- **.913** bivariate correlation with the ground truth (*RCGP ILI rates*)

Flu detector, part of UK's influenza surveillance

UK Health Security Agency

gov.uk/government/statistics/national-flu-and-covid-19-surveillance-reports-2023-to-2024-season



Why estimate disease rates from web search?

- Complements conventional syndromic surveillance systems
  - larger *cohort*
  - broader *demographic coverage*
  - broader, more granular *geographic coverage*
  - not affected by *closure days* and other *temporal biases*
  - *timeliness*
  - *lower cost*

oxymoron: public health data is needed to train machine learning models!

- Track novel infectious diseases

Conventional (*traditional*) syndromic surveillance methods: disease prevalence, i.e. the % of infected people in a population, is determined via doctor (GP) visits and other related indicators, such as laboratory-confirmed infections, associated hospitalisations or deaths.

Wagner *et al.* (2018), *Sci. Rep.*; Budd *et al.* (2020), *Nat. Med.*

Part B

Transfer learning for disease modelling from web search activity from one location to another

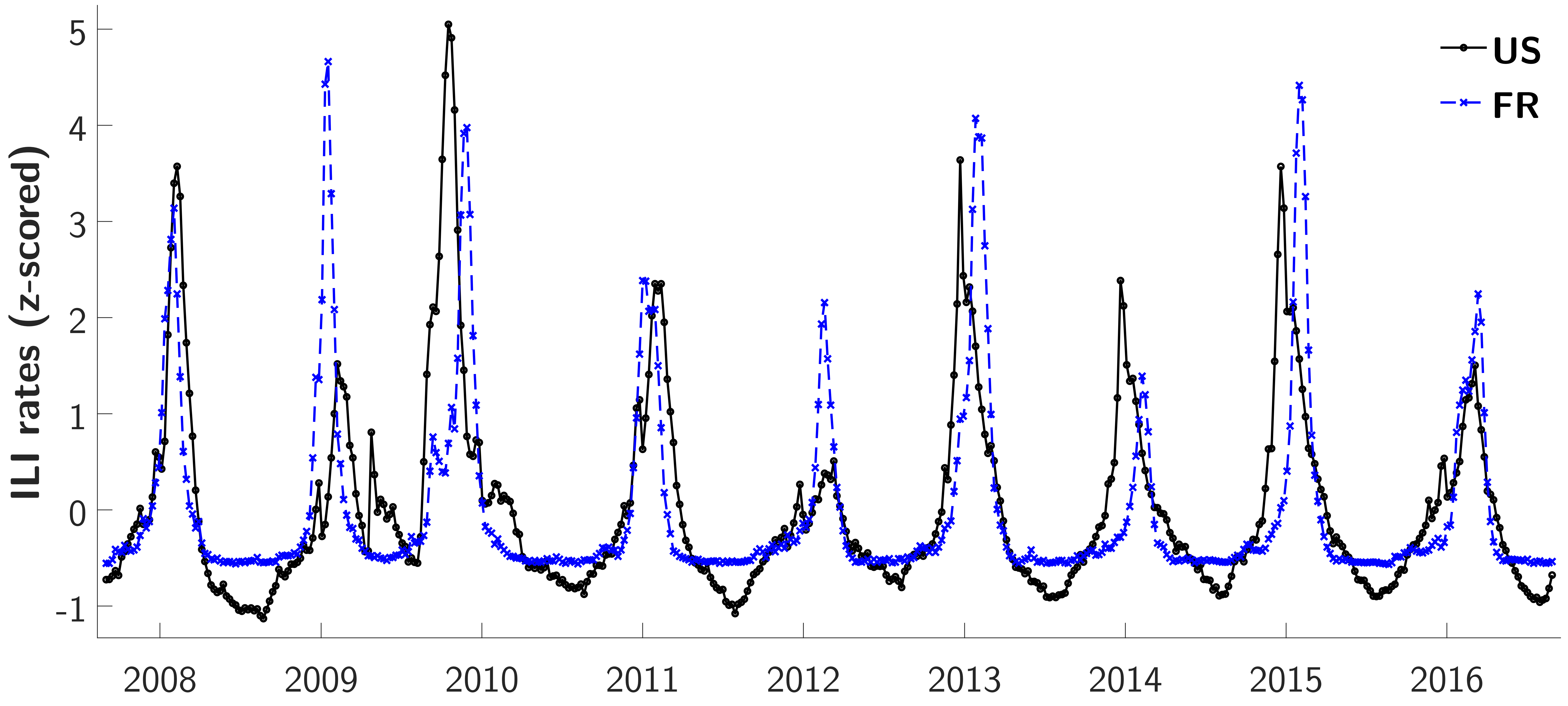
Zou, Lampos, Cox (2019), *WWW '19*

Transfer learning across countries for flu models from web search

- **Transfer learning** *in general*
  - Gain knowledge from one domain/task, apply it to another one

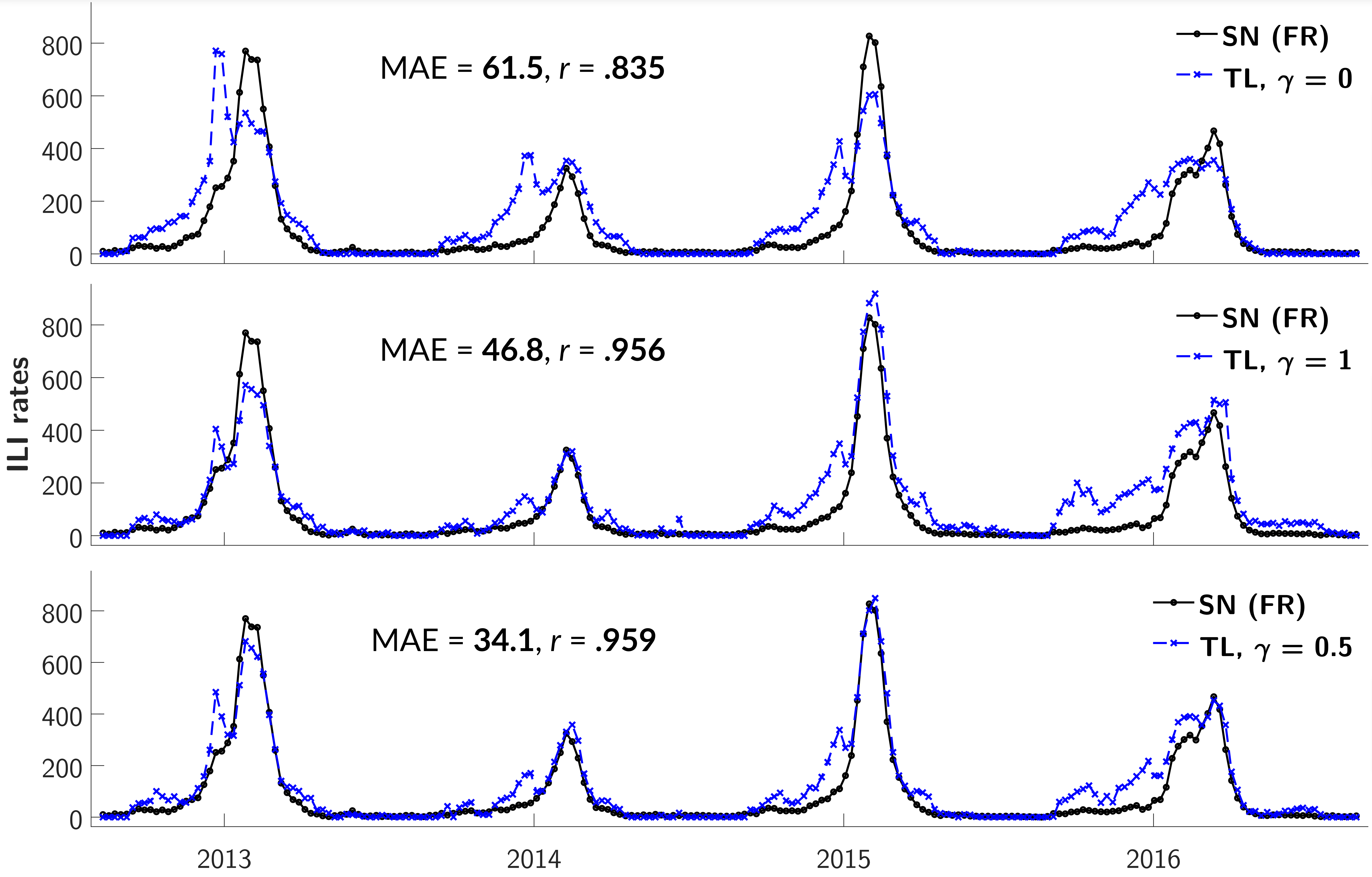
- **Transfer learning** *for estimating flu rates across different countries*
  - Locations: **source** (*no missing data*), **target** (*no disease rates*)
  - regularised **regression model** for a source location based on web search activity and historical disease rates
  - **map search queries** from the source to the target location
    — *semantic similarity* (bilingual if necessary)
    — *temporal similarity*
    — *hybrid similarity* (their linear combination controlled by γ)
  - transfer regression model (*equivalent to zero-shot learning*)

Transferring a flu model based on web searches: *from* US *to* France



How similar are the flu rates between the **US** and **France (FR)**?
— temporal differences (e.g. different onset/peak moments), intensity differences

Transferring a flu model based on web searches: *from* US *to* France

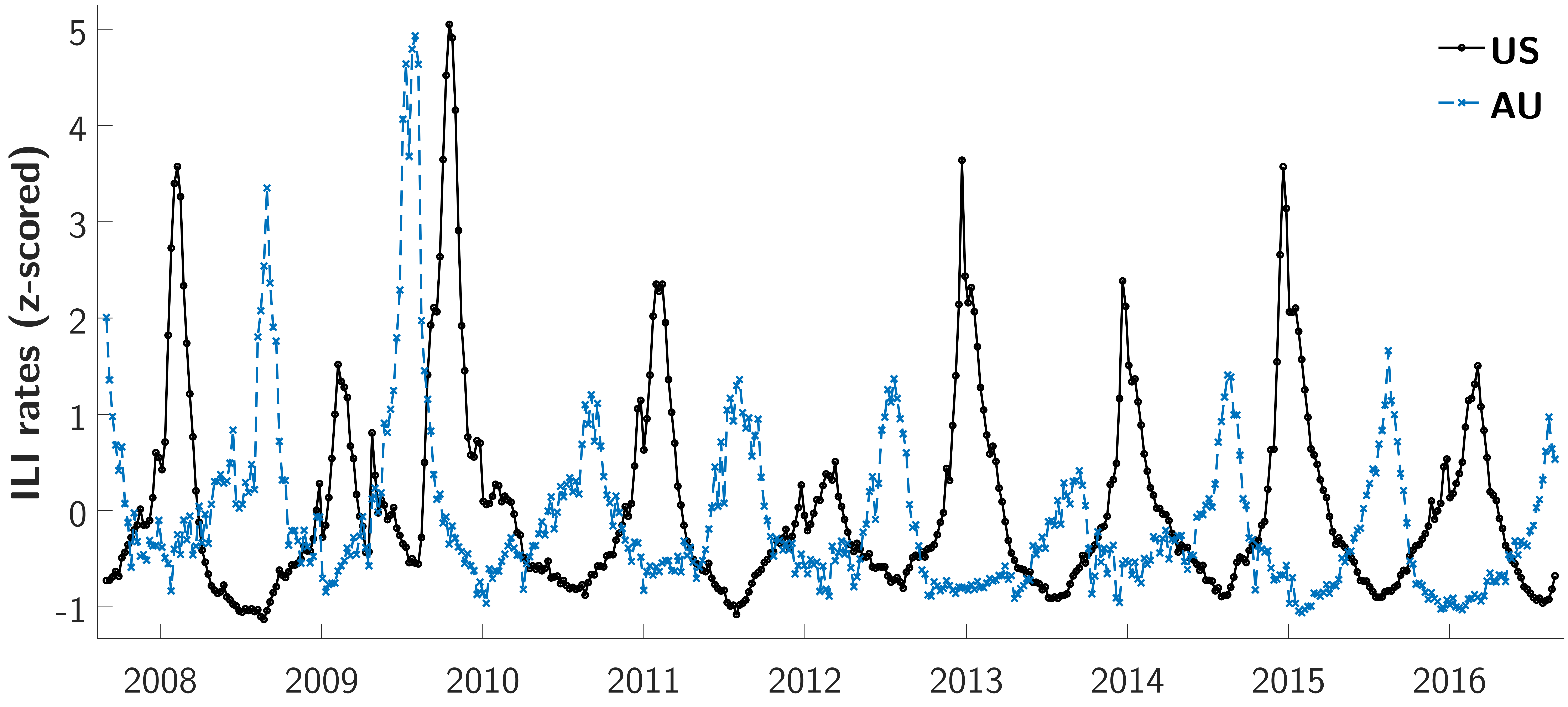


temporal correlation

semantic similarity

hybrid similarity

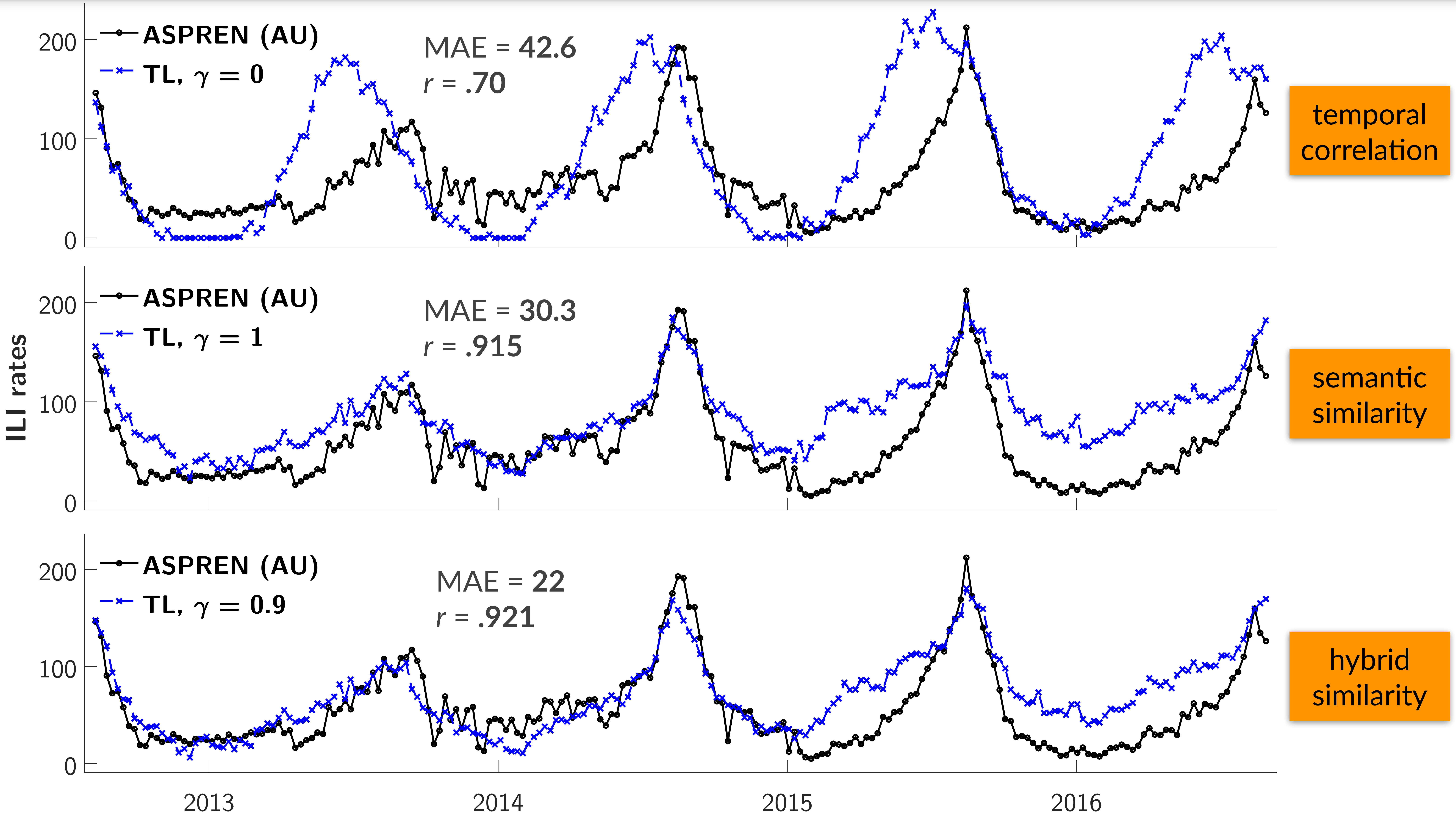
Transferring a flu model based on web searches: *from* US *to* Australia



How similar are the flu rates between the **US** and **Australia (AU)**?

— different (≈opposite) seasons, significant intensity differences in more recent years

Transferring a flu model based on web searches: *from* US *to* Australia



Part C

Tracking COVID-19 using online search

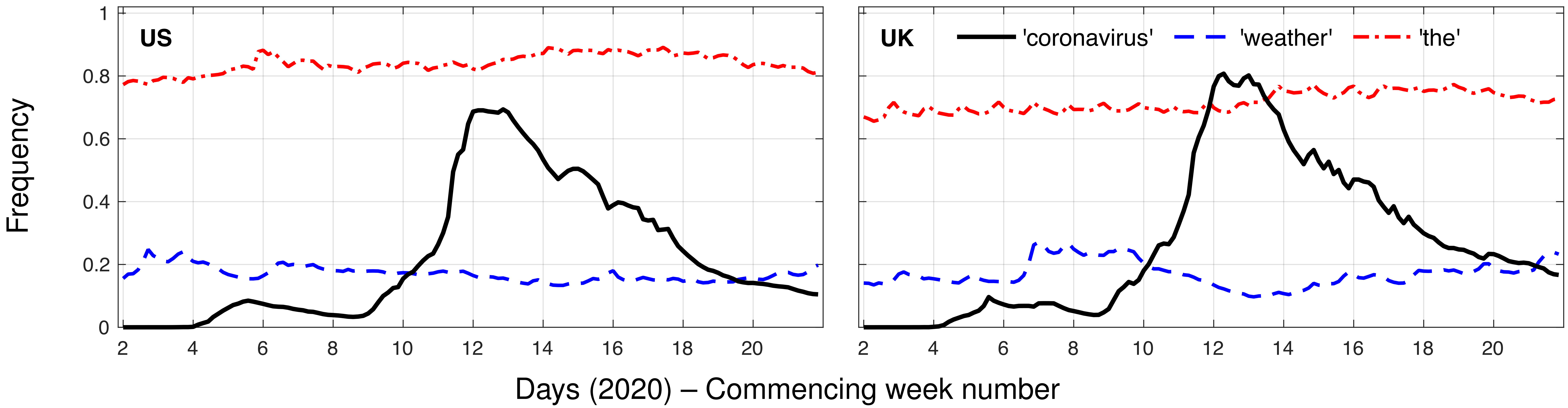
Lampos *et al*. (2021), *npj Digit. Med.*

Google search activity

Google Health Trends: frequency $y_{L,d}$ of web search query q for a location L during a day d

$$y_{L,d} = \frac{\text{number of times } q \text{ was issued by users in location } L \text{ during day } d}{\text{total number of searches by users in location } L \text{ during day } d}$$

Unprecedented search frequency trends during the first COVID-19 pandemic waves

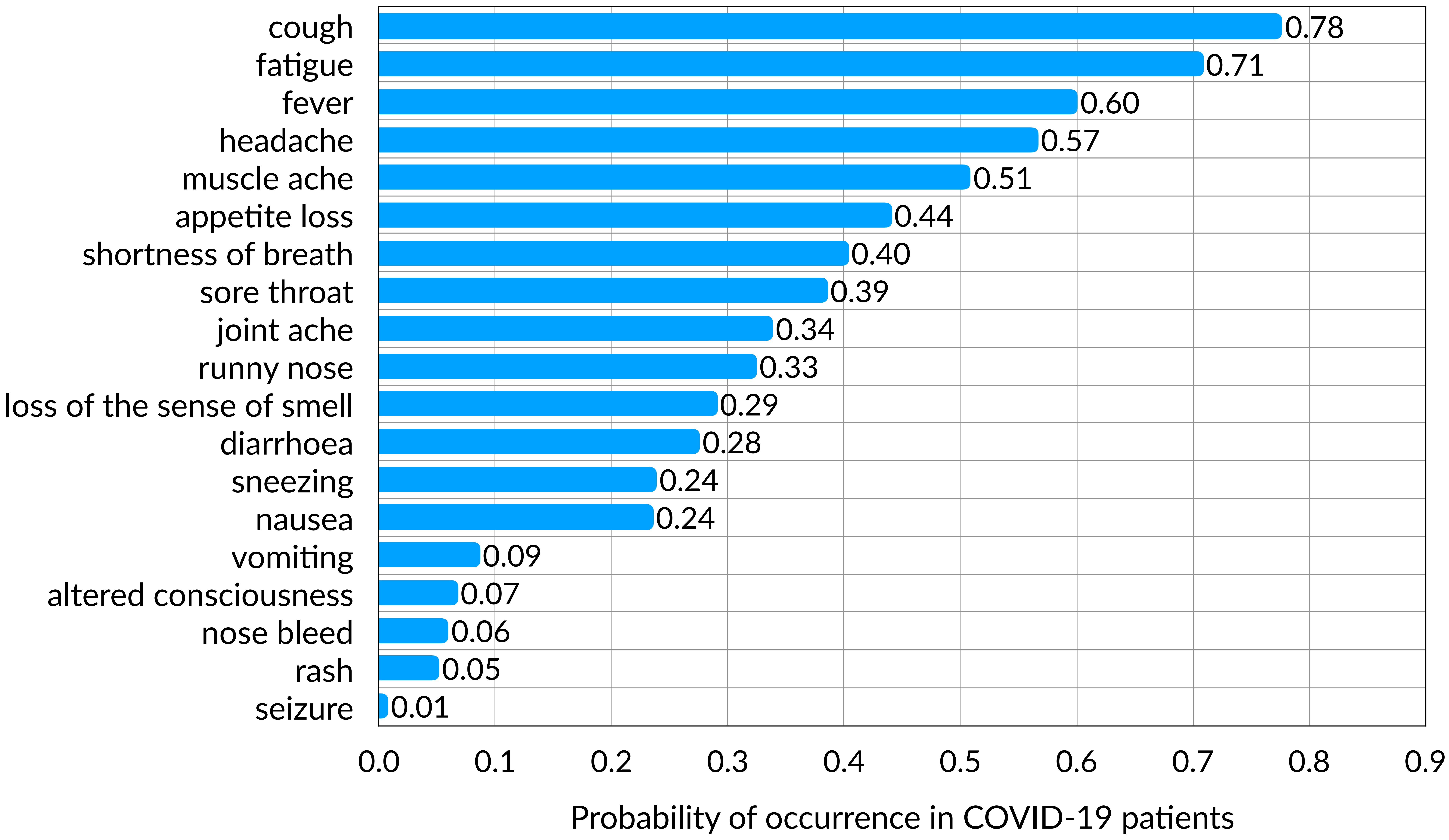


Challenges in modelling COVID-19 using web search activity

- No reliable and not enough **ground truth** data
  - Supervised learning no longer possible — *can we use transfer learning?*
  - Evaluation of any model will be problematic

- **Unsupervised learning**
  - Which search queries to use?
  - How do we know our model is related to COVID-19 and not other infectious diseases?
  - How do we know our signal is not affected by other factors such as concern, curiosity, and media coverage rather than by infection?

First few hundred (FF100) patient survey (NHS & UKHSA)



Boddington *et al.* (2021), *Bull. WHO*

Symptom-related search terms — *English*

- **cough:** *cough*, coughing
- **fatigue:** *fatigue*
- **fever:** chills, *fever*, high temp fever, high temperature
- **headache:** head ache, *headache*, headaches, migraine
- **muscle ache:** *muscle ache*, muscular pain
- **appetite loss:** *appetite loss*, loss of appetite, lost appetite
- **shortness of breath:** breathing difficulties, breathing difficulty, cant breathe, *shortness of breath*, short breath
- ...
- **loss of the sense of smell:** anosmia, loss of smell, loss smell
- **COVID-19 terms:** coronavirus, covid, covid-19, covid19

Symptom-related search terms — *Italian*

- **cough:** tosse, tossire
- **fatigue:** affaticamento, fatica, spossatezza, stanchezza
- **fever:** alta temperatura, brividi, febbre
- **headache:** emicrania, mal di testa
- **muscle ache:** dolore muscolare, dolori muscolari, male ai muscoli, mialgia
- **appetite loss:** appetito perso, inappetenza, perdita appetito, perdita di appetito
- **shortness of breath:** difficoltà respiratoria, difficoltà respiratorie, fiato corto, mancanza di respiro, respiro corto
- ...
- **loss of the sense of smell:** perdita olfatto
- **COVID-19 terms:** coronavirus, covid, covid-19, covid19

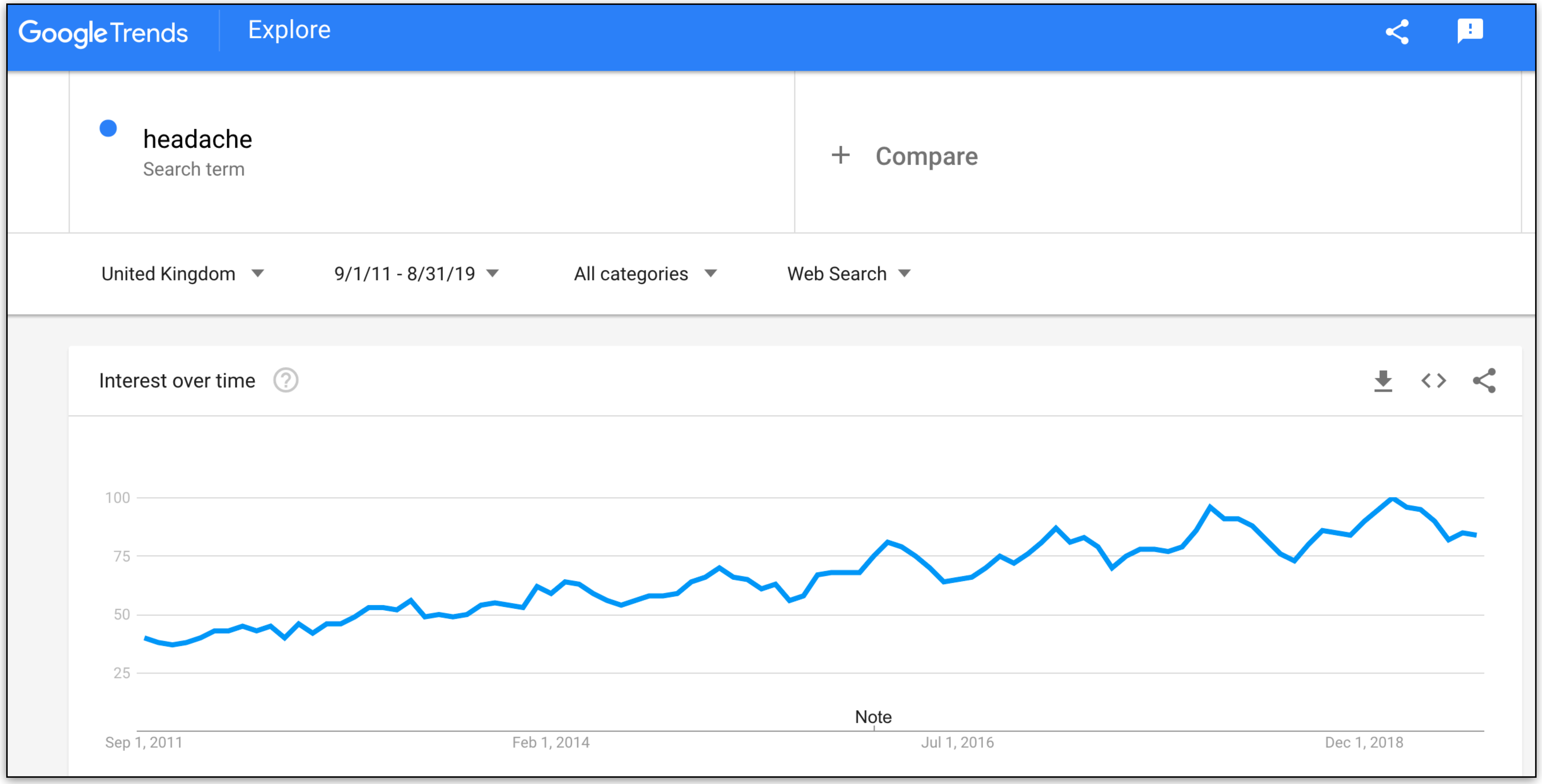
Symptom-related search terms — *Locations (countries) & languages*

Our analysis considered the following countries and corresponding languages:

- United States of America (US), United Kingdom (UK), Australia, Canada — **English**
- France — **French**
- Italy — **Italian**
- South Africa — **Zulu**, **Afrikaans**, **English**, and many more
- Greece — **Greek**

A simple COVID-19 prevalence model (1/2)

1. Query frequencies are **noisy**
   — harmonic smoothing using the frequencies of the past 2 weeks
2. Query frequencies are **not stationary** (*increasing mean*)
   — linear detrending



A simple COVID-19 prevalence model (2/2)

3. For each symptom category, obtain the frequency sum across all its search terms (**cumulative symptom-related search frequency**) on a daily basis

4. Apply **min-max normalisation** on the cumulative frequency of each symptom category; values become from 0 to 1 and all categories now share units

5. Compute a **daily weighted score** using the FF100 symptom probabilities as weights

6. Use the previous 8 years (2011-2019) to obtain a **historical baseline** of this scoring function

Reducing the effect of news media coverage (1/2)

For a given *day* and *location*

– proportion of COVID-19-related **news** articles: $m \in [0,1]$

– **COVID-19 score** based on web searches: $g \in [0,1]$

Decompose g such that $g = g_p + g_c$

– g_p represents '*infection*'

– g_c represents '*concern*'

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

– $g_p = \gamma g$

– $g_c = (1 - \gamma)g$

Reducing the effect of news media coverage (2/2)

Linear autoregressive model to forecast COVID-19 score g at a time point t based on its past values

$$\underset{\mathbf{w},b_1}{\arg\min} \frac{1}{N} \sum_{t=1}^{N} \left(g_t - w_1 g_{t-1} - w_2 g_{t-2} - b_1\right)^2 \rightarrow \text{ prediction error } \epsilon_1$$

Linear autoregressive model to forecast COVID-19 score g at a time point t based on its past values **and** the current and past values of m

$$\underset{\mathbf{w},\mathbf{v},b_2}{\arg\min} \frac{1}{N} \sum_{t=1}^{N} \left(g_t - w_1 g_{t-1} - w_2 g_{t-2} - v_1 m_t - v_2 m_{t-1} - v_3 m_{t-2} - b_2\right)^2 \rightarrow \text{ prediction error } \epsilon_2$$

- $\epsilon_1 < \epsilon_2$: the media signal does not help COVID-19 score predictions $\rightarrow \gamma \approx 1$, i.e. the media is expected to not have a causal effect on the estimated COVID-19 scores
- $\epsilon_1 \geq \epsilon_2 : \gamma = \epsilon_2 / \epsilon_1$ (*crude estimation of % of impact of news media*)

News media coverage corpus

- Data obtained from the **Media Cloud** database — mediacloud.org
- Number of **news media sources** per country

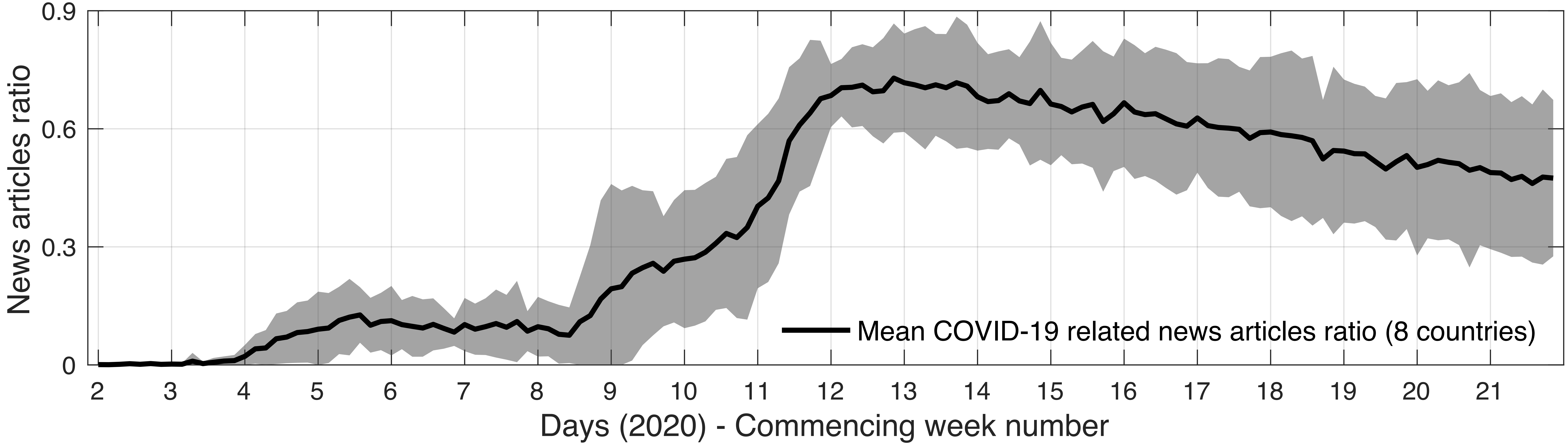
US	225
UK	93
Australia	61
Canada	79
France	360
Italy	178
Greece	75
South Africa	135

- Obtain the **daily ratio** of articles that include basic COVID-19-related keywords in their title or main text
  e.g. “covid” or “coronavirus”

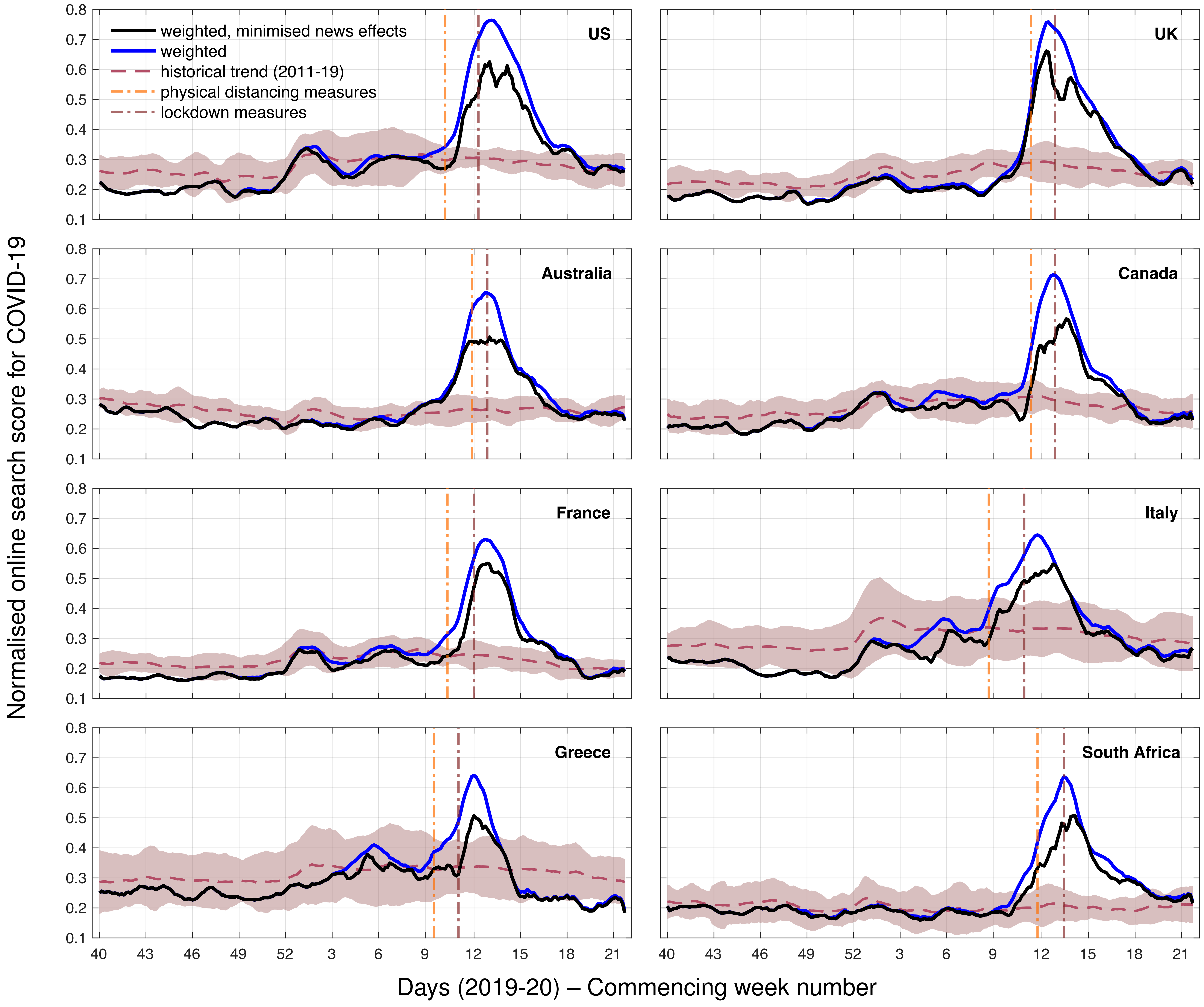
News media coverage corpus

- Data obtained from September 30, 2019 to May 24, 2020
- > 0 frequency from ~January, 2020 onwards
- **~2.5 million** COVID-19-related articles from a total of ~10 million

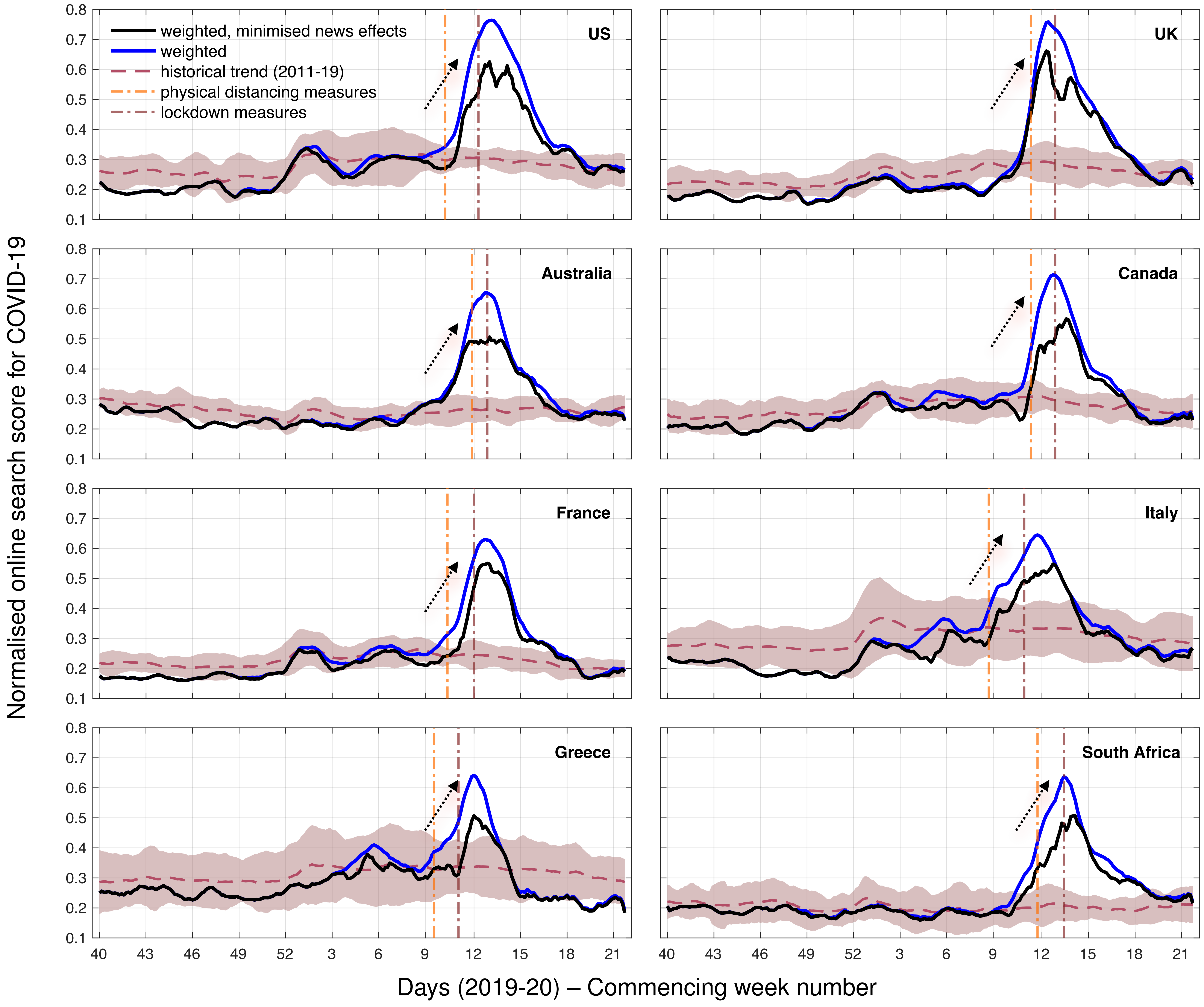
Average proportion of COVID-19-related news articles in the 8 countries of our analysis



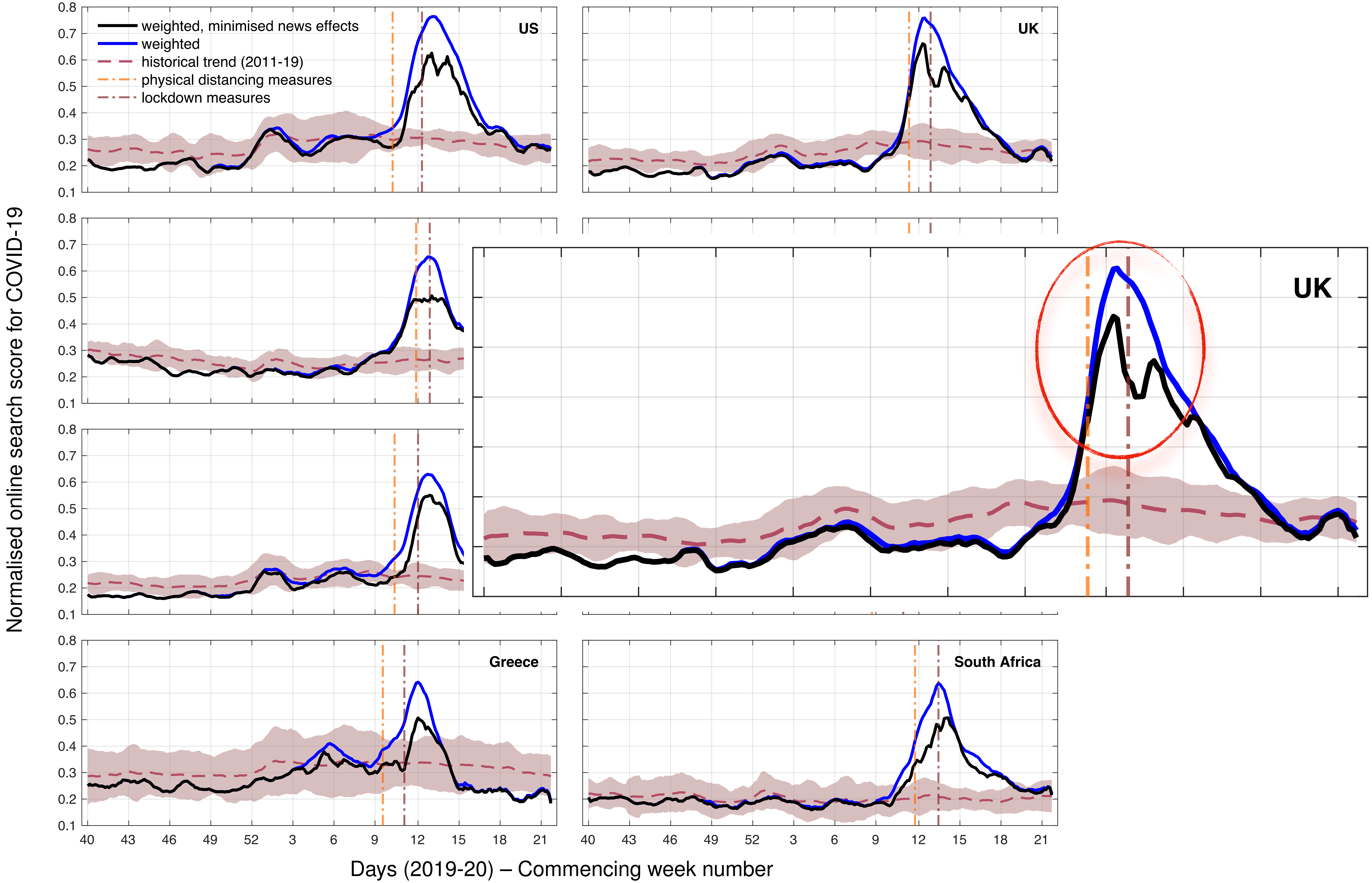
Unsupervised COVID-19 models in 8 countries based on web search



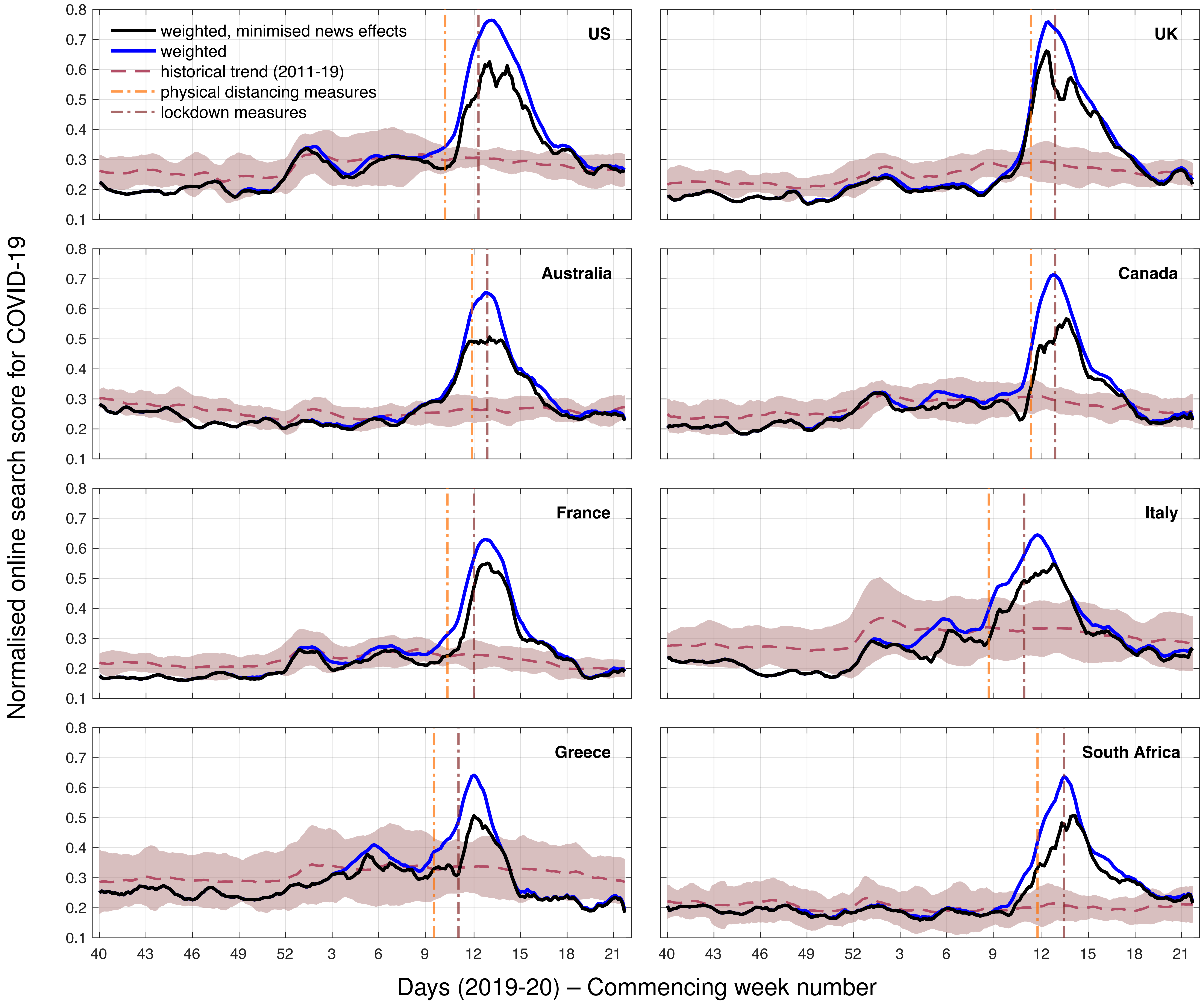
Unsupervised COVID-19 models in 8 countries based on web search



Unsupervised COVID-19 models in 8 countries based on web search



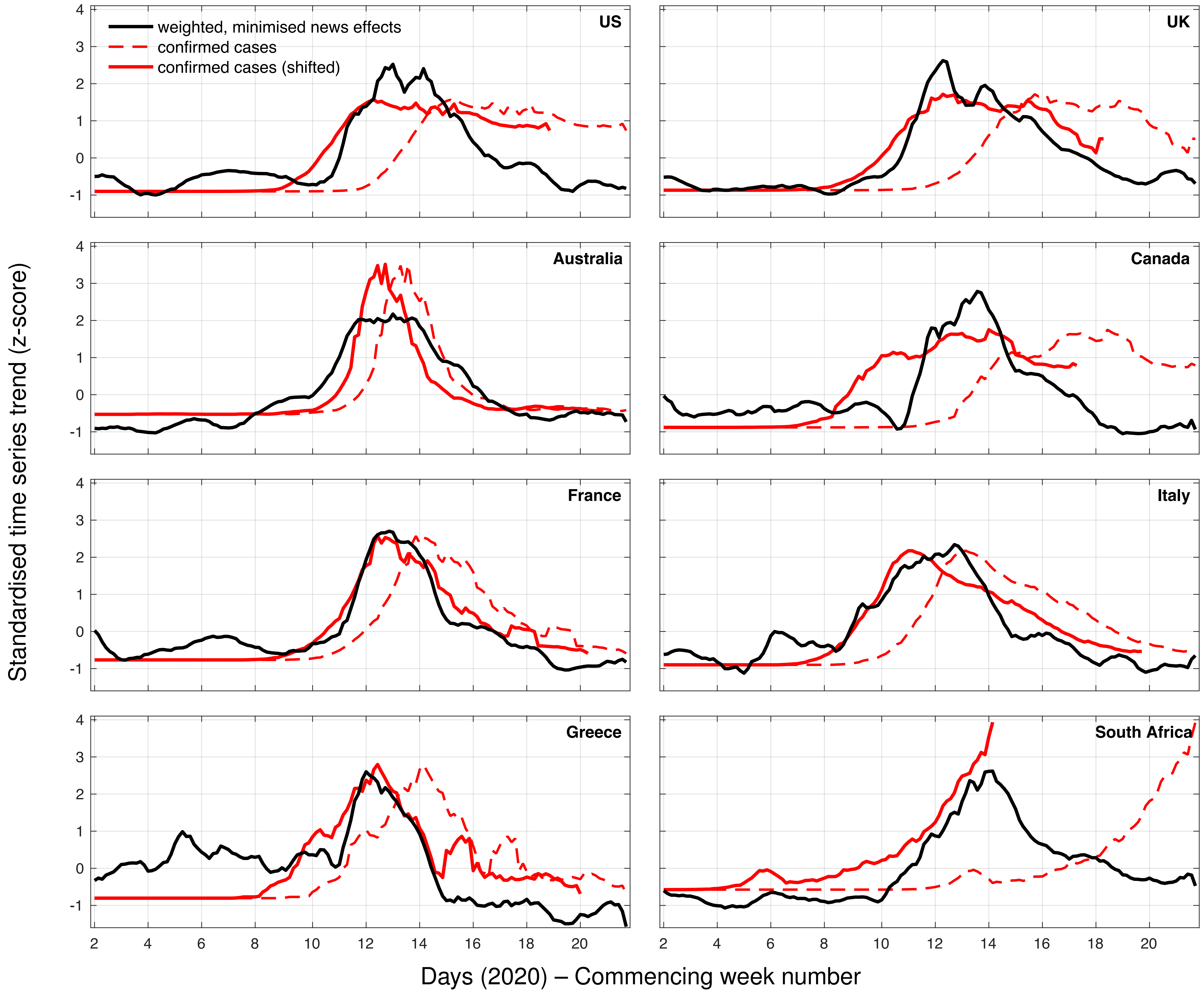
Unsupervised COVID-19 models in 8 countries based on web search



Reducing news media effects:

- **Altered trend** during peak periods
- Average reduction by **16.4%** (14.2%—18.7%) in a period of 14 days prior and after their peak moments, *r* = .822 (.739—.905)
- Reduction of **3.3%** (2.7%—4%) outside peak periods

Comparison with *confirmed COVID-19 cases*



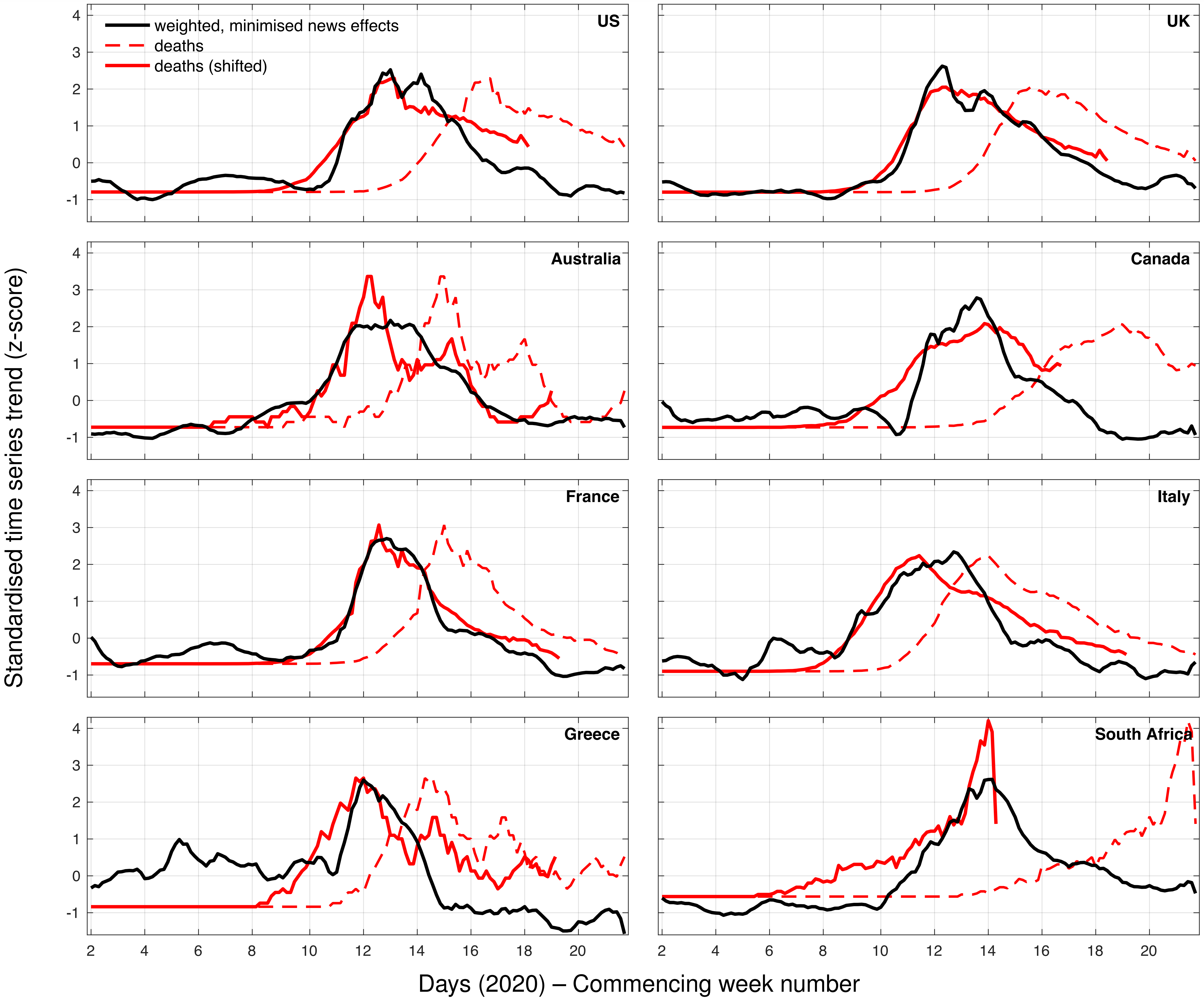
Web search activity based models provide an **early warning**

r_{max} = .83 (.74—.92)

when cases are brought forward by **16.7** (10.2—23.2) **days**

(*South Africa is excluded*)

Comparison with *deaths of people with COVID-19*



Web search activity based models provide an **early warning**

r_{max} = .85 (.70—.99)

when deaths of people with COVID-19 are brought forward by **22.1** (17.4—26.9) **days**

(*South Africa is excluded*)

Transfer learning for COVID-19 incidence models

- **Transfer** an incidence model — **trained on web search activity** — for a *source* country that has already experienced a COVID-19 epidemic to other *target* countries that are on earlier stages of the epidemic

- “**Supervised**” **learning** approach
  - corroborate our previous unsupervised findings
  - will also transfer characteristics/biases of the source country, and especially of its clinical reporting system

- Source country: **Italy**
  - first major outbreak in Europe and among the countries in our study

Transfer learning for COVID-19 incidence models

- **Source model**: regularised regression (*elastic net*)
  - use daily search query frequencies to estimate confirmed cases
  - Italy is our source country

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

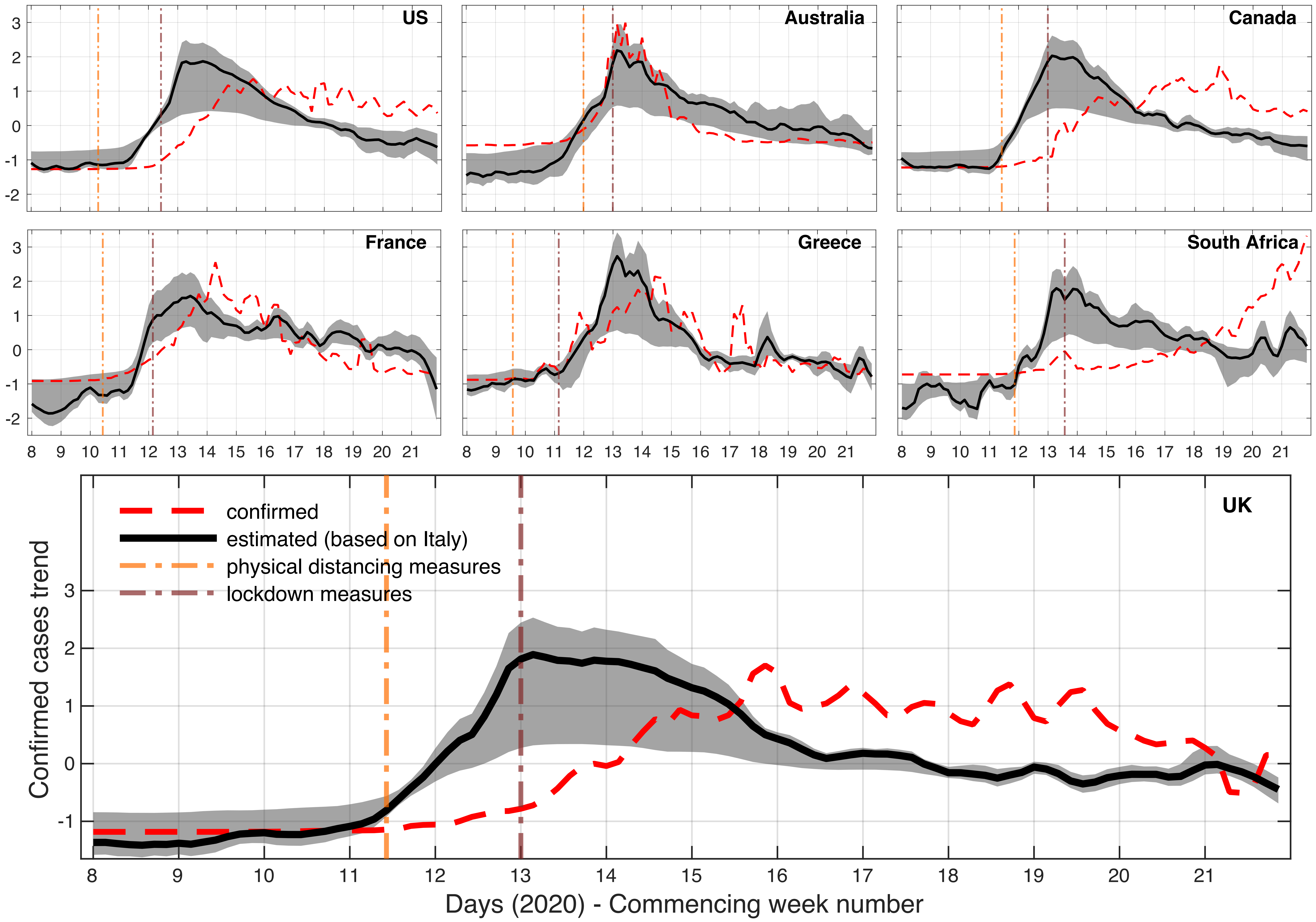
$\mathbf{S} \in \mathbb{R}^{M \times N}$: M daily frequencies of N search terms
$\mathbf{w} \in \mathbb{R}^N, \beta \in \mathbb{R}$: regression weights and intercept
$\lambda_1, \lambda_2 \in \mathbb{R}_{\geq 0}$: regularisation parameters

- **Many regression models** (~80K) — different regularisation amount
  - sparsity levels from 5.5% to 91%
    3 to 49 selected queries from the 54 we considered for Italy
  - use this as crude quantification of **model's uncertainty**

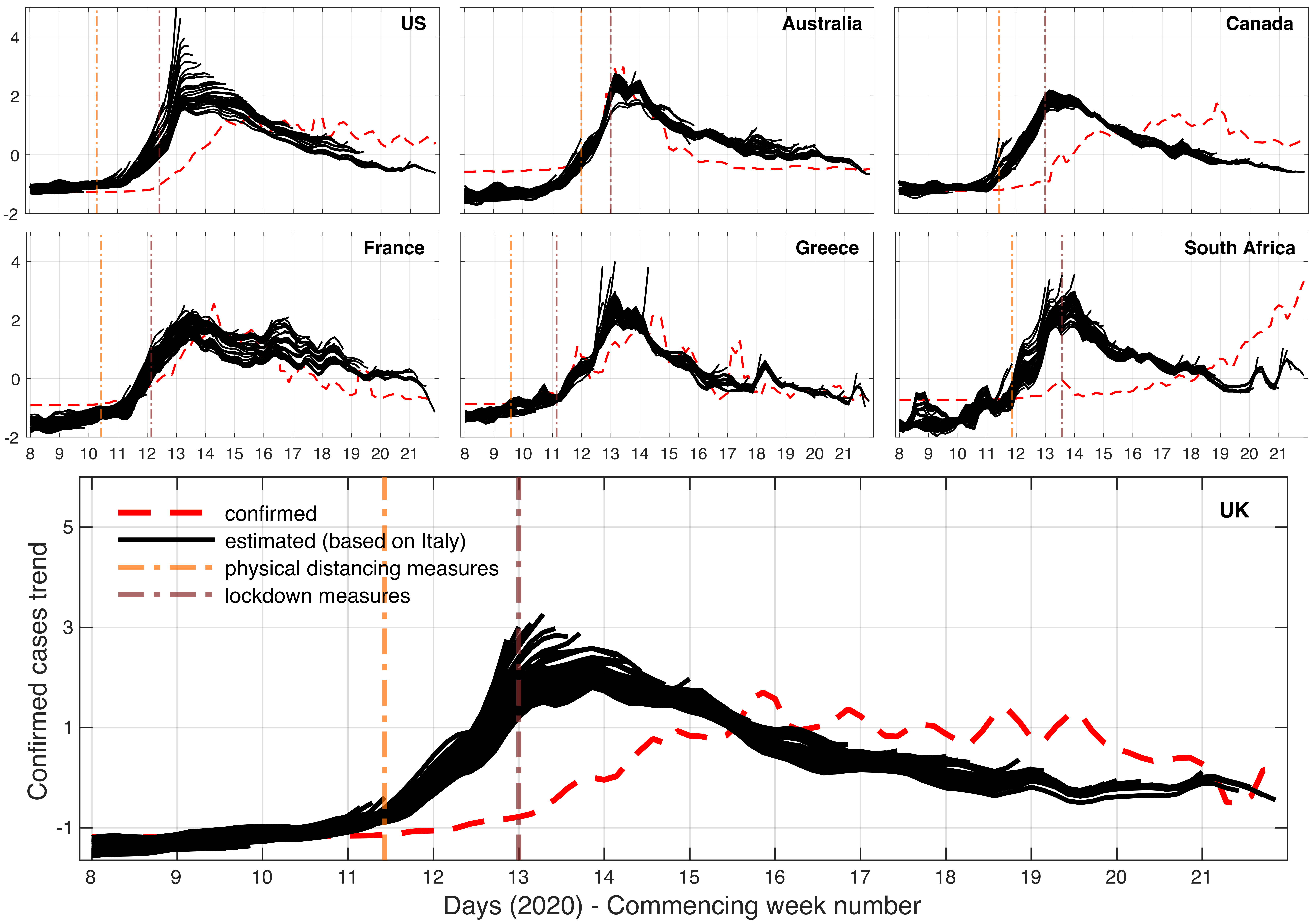
Transfer learning for COVID-19 incidence models

- Establish search query pairs between the source and the target countries
  - lookup for query pairs **within the same symptom category**
  - pair a source query to the target query with the greatest **bivariate correlation**, after identifying an optimal shifting period
- **Transfer** the regression weights from the source to the target feature space for all ~80K elastic net models
  - Final estimate of COVID-19 incidence is the **mean** over all elastic net models
  - **.025 and .975 quantiles** are used to form 95% confidence intervals
- Perform this daily from Feb. 17 to May 24, 2020, training models on increasing data from the source country

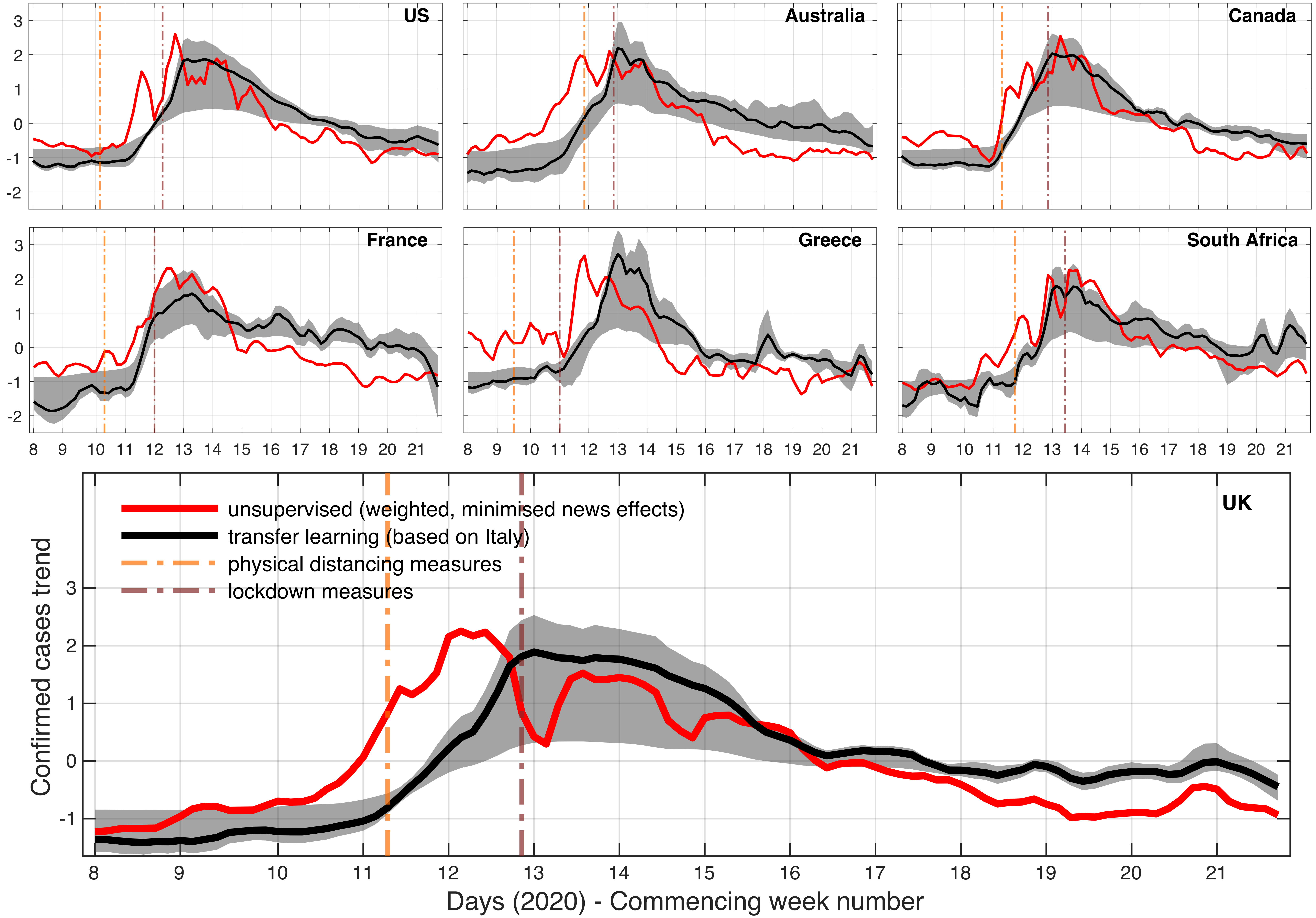
Transfer learning for COVID-19 incidence models



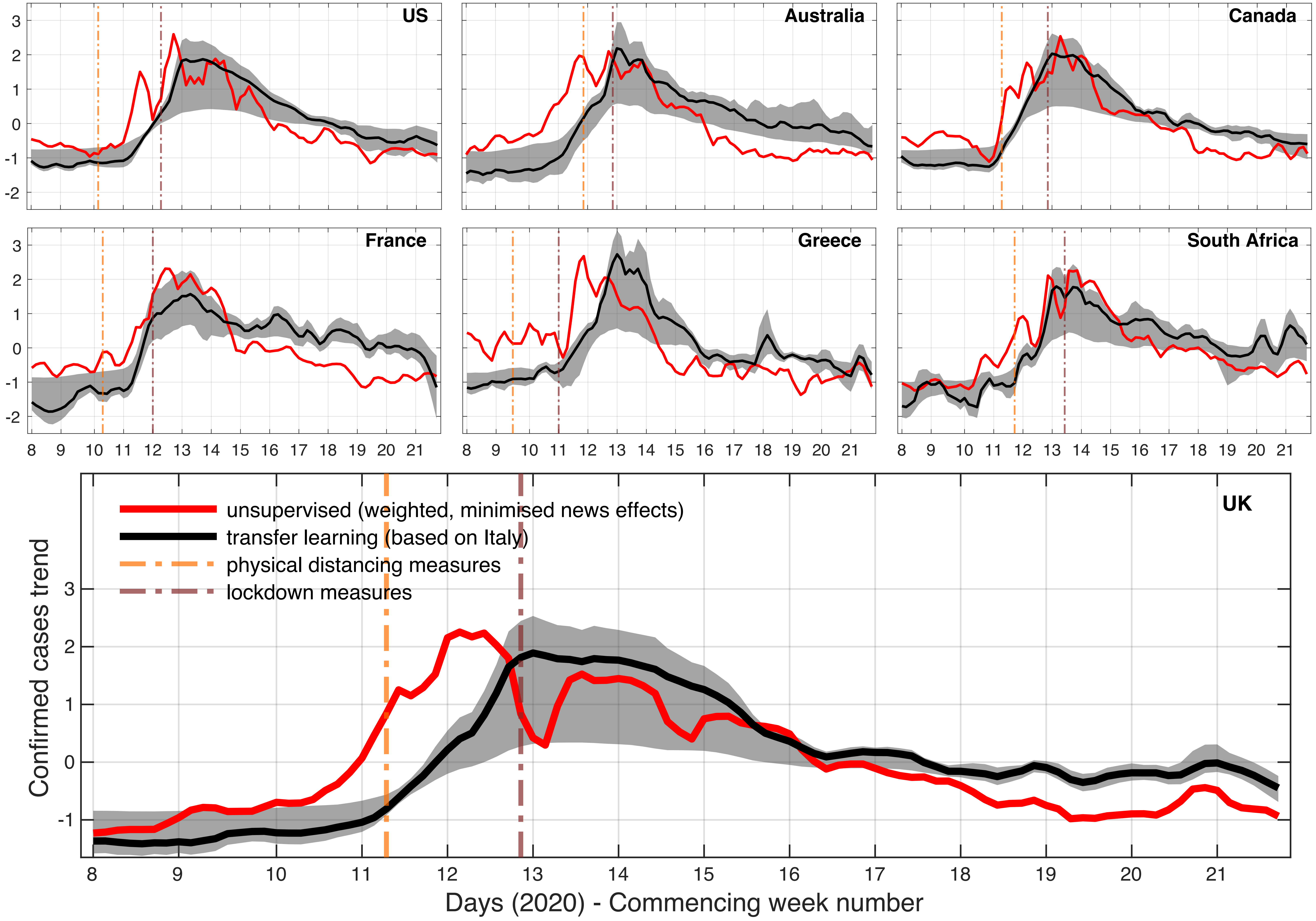
Transfer learning for COVID-19 incidence models — *In practice*



Transfer learning vs. unsupervised learning



Transfer learning vs. unsupervised learning



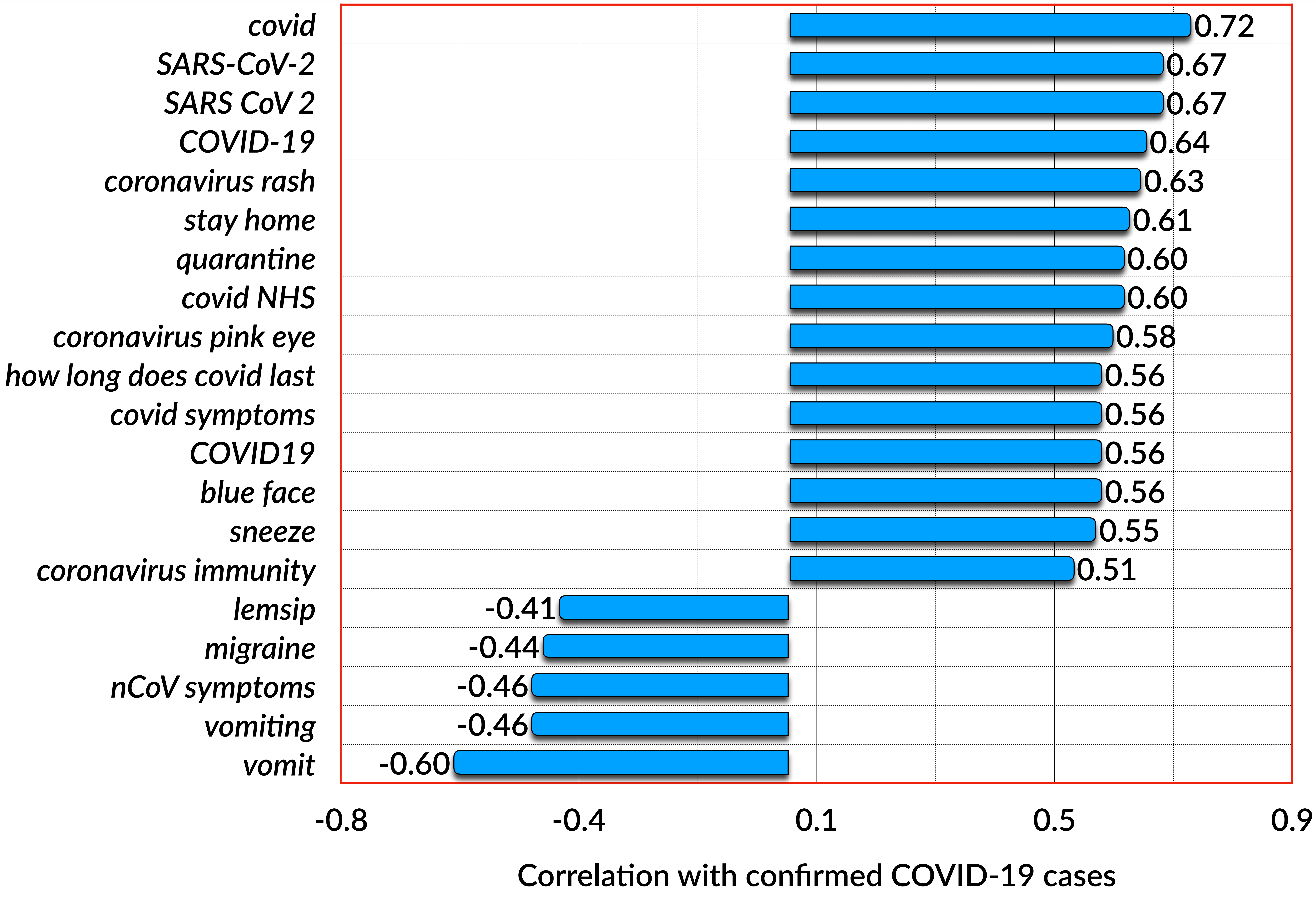
Correlation between the *transferred* models and the *unsupervised* models with *reduced media effects*

- r_{avg} = .66
- $r_{max\text{-}avg}$ = .80, when the *transferred* time series are brought **5 days forward**

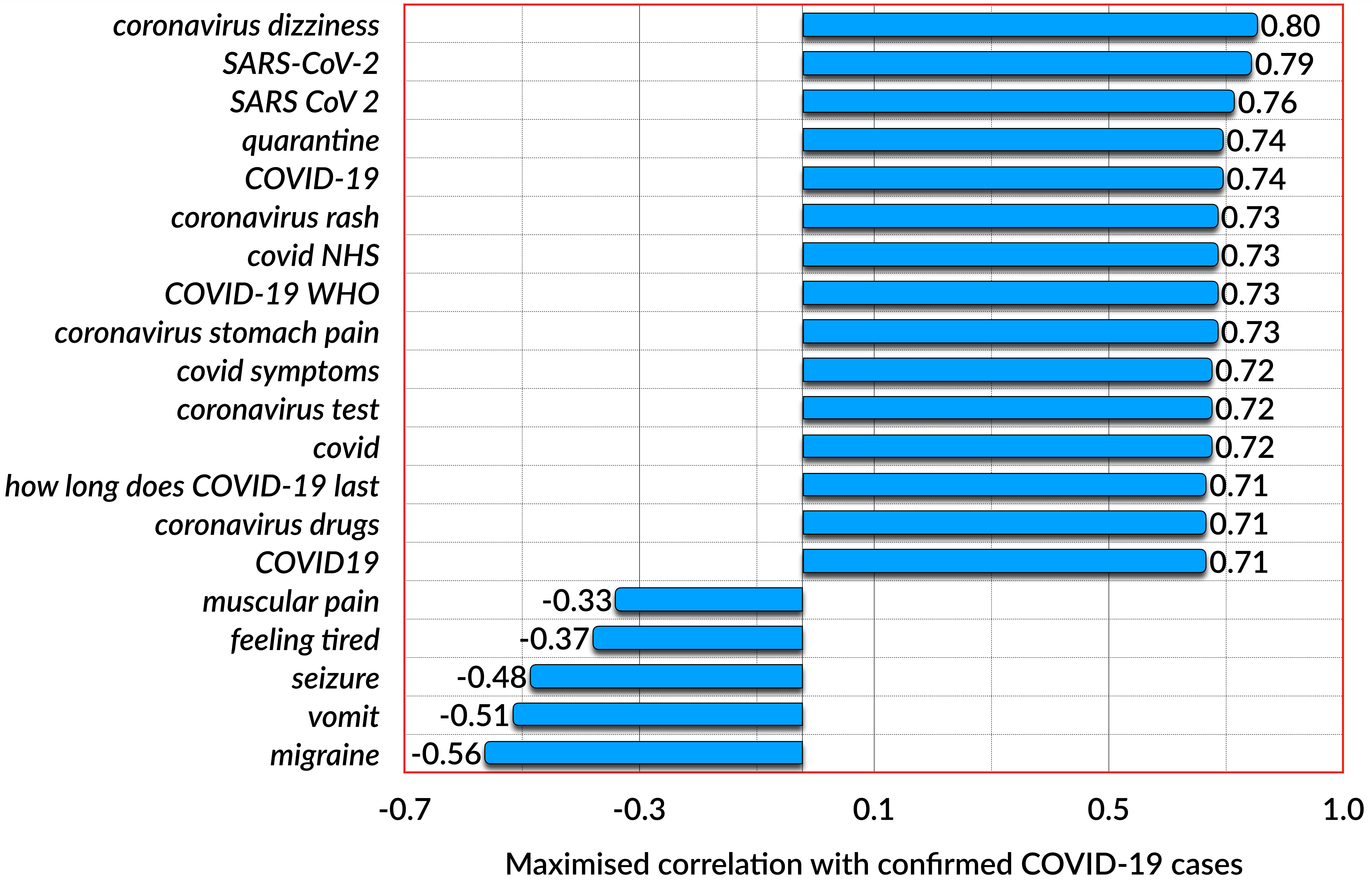
Correlation analysis

- Examine the statistical relationship between web search frequencies and confirmed COVID-19 cases (or deaths)
- Jointly for 4 English-speaking countries (US, UK, Australia, Canada)
  - attempt to reduce the bias of clinical endpoints in these different countries
  - focus on English-speaking countries for more comprehensive outcomes (without the need to translate searches)
- Use a broader set of search terms, not just symptom-related
  — figshare.com/projects/Tracking_COVID-19_using_online_search/81548
- Compute the joint bivariate correlation between search frequency and clinical indicators (cases or deaths) without any shifting and after shifting data so as to maximise it

Correlation between web searches and COVID-19 cases



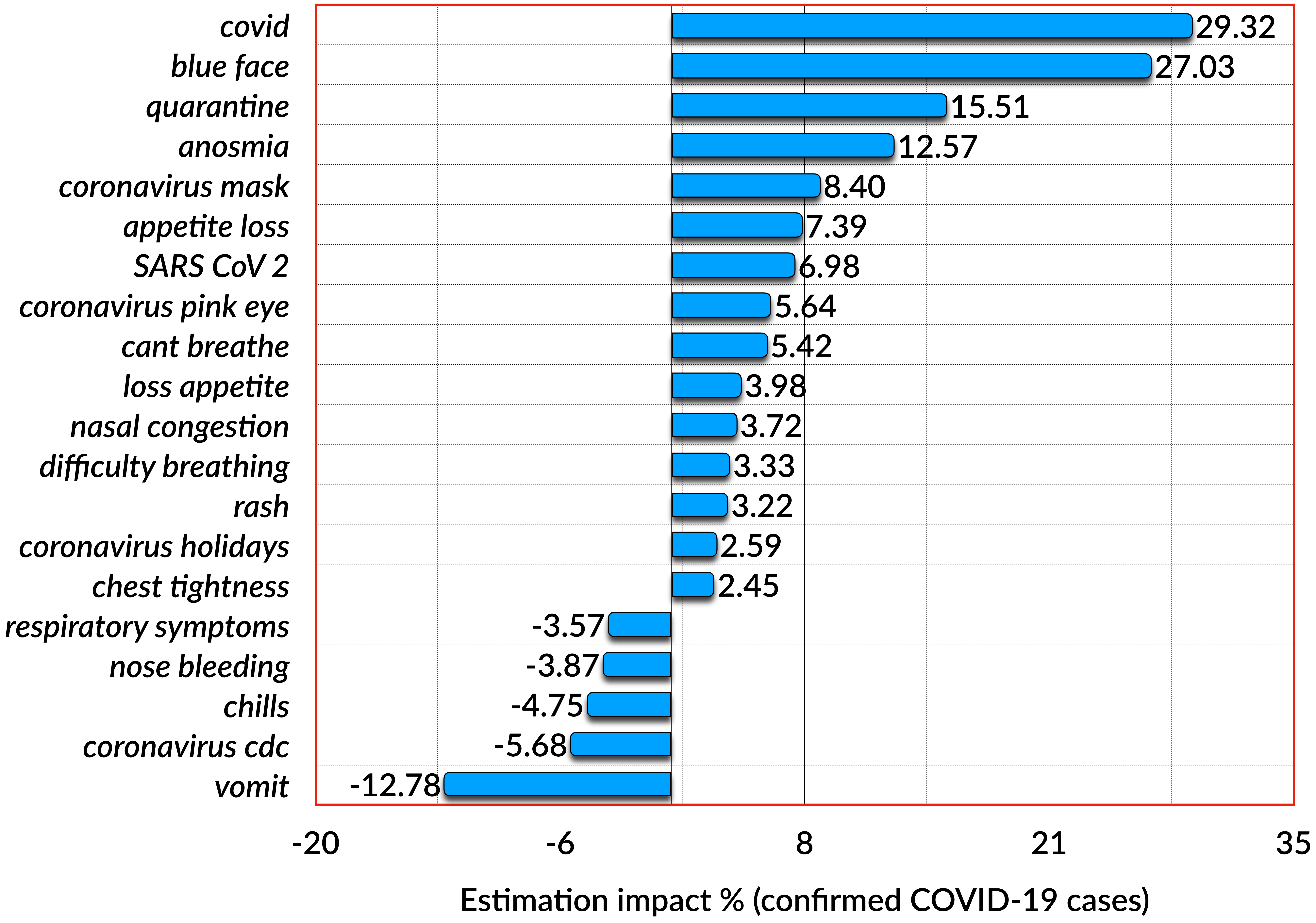
Maximised correlation between web searches and COVID-19 cases



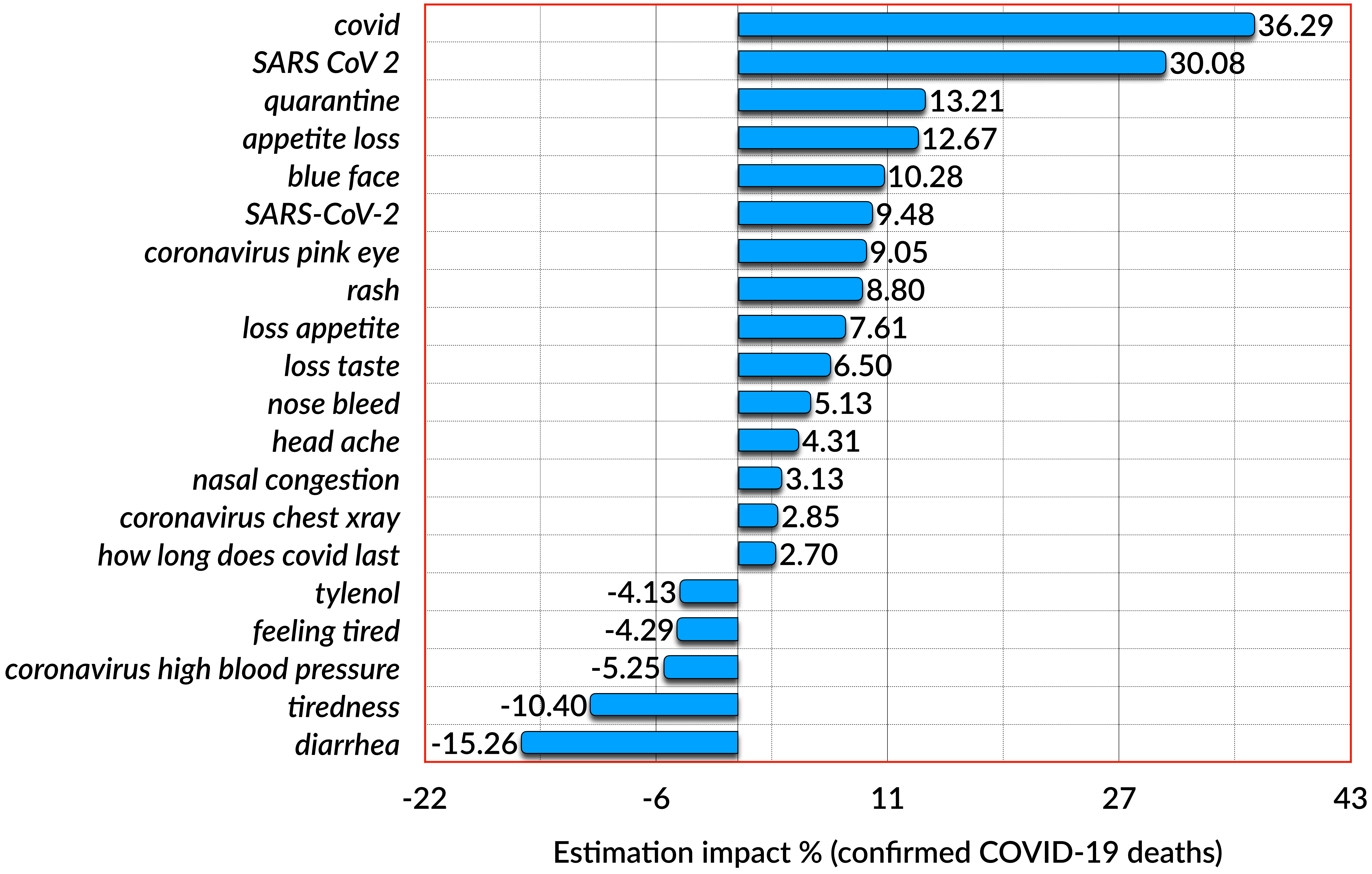
Regression analysis

- Same **4 English speaking countries** (US, UK, Australia, Canada)
- **Joint** approach again
- **Multivariate regression analysis**
  - Learn many elastic net models for different levels of sparsity (50%-99% to reduce the chance of *overfitting*) to **jointly** estimate cases or deaths based on web search data in these 4 countries
  - Train on data up to day *d*, test performance on the next day, *d*+1
  - Repeat this daily from the 2nd of March to the 24th of May, 2020
  - Use ground truth to find the best solution at each sparsity level
  - Compute the impact (average across all days) of each search term in the best solution at each density level

Regression analysis — *confirmed COVID-19 cases*



Regression analysis — *deaths of people with COVID-19*



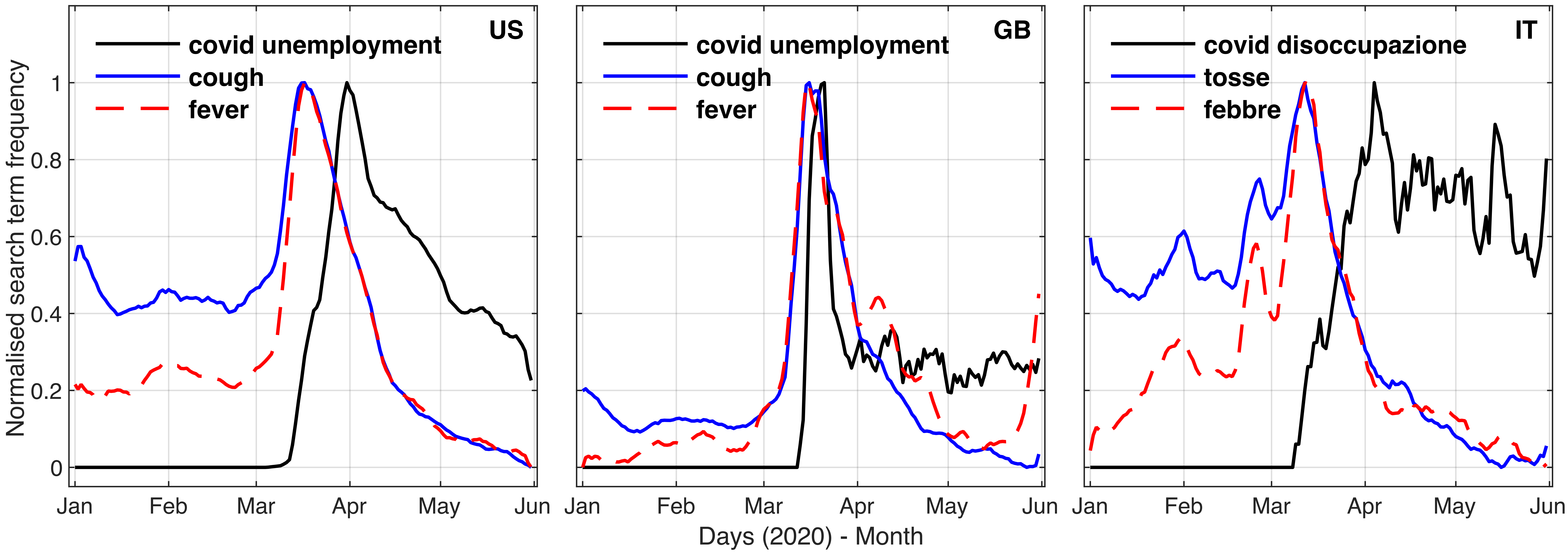
The effect of the COVID-19 outbreak in Italy in web search activity

Did the outbreak in Italy **cause** an increase in the frequency of the web searches (*the ones used in our analysis*) elsewhere?

- Test this hypothesis from Feb. 17 to April 19, 2020
  a 4-week period after the corresponding peak in confirmed cases or deaths in Italy is added
- Cases or deaths in Italy **Granger-caused < 27.5%** of the considered search terms across the 7 other countries in our analysis
- **> 70%** of the search terms used in our analysis are **not affected**
- This analysis does not account for the fact that cases and deaths might have been **rising in both locations** *at the same time*
- We have also attempted to **reduce news media effects** in the final signal
- For **Italy itself** the early-warning provided by the unsupervised signal with reduced media effects is 14 and 18 days compared to confirmed cases and deaths, respectively

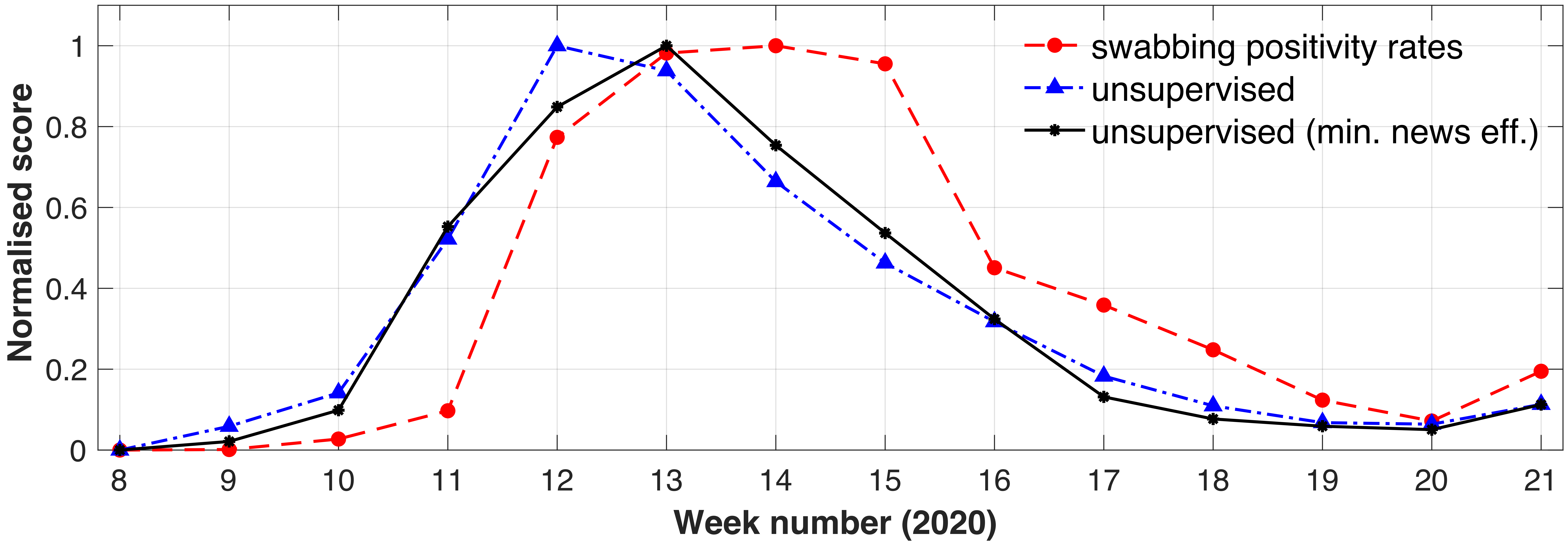
Symptom-related vs. general interest search terms

Search terms that are less likely to represent infection ("COVID unemployment") ***follow*** the corresponding trends of search terms about COVID-19-related symptoms ("cough", "fever")



RCGP swabbing scheme for estimating COVID-19 prevalence in England

The Royal College of General Practitioners (**RCGP**) **swabbing scheme** included people with **no** COVID-19-related symptoms → better capturing community-level spread



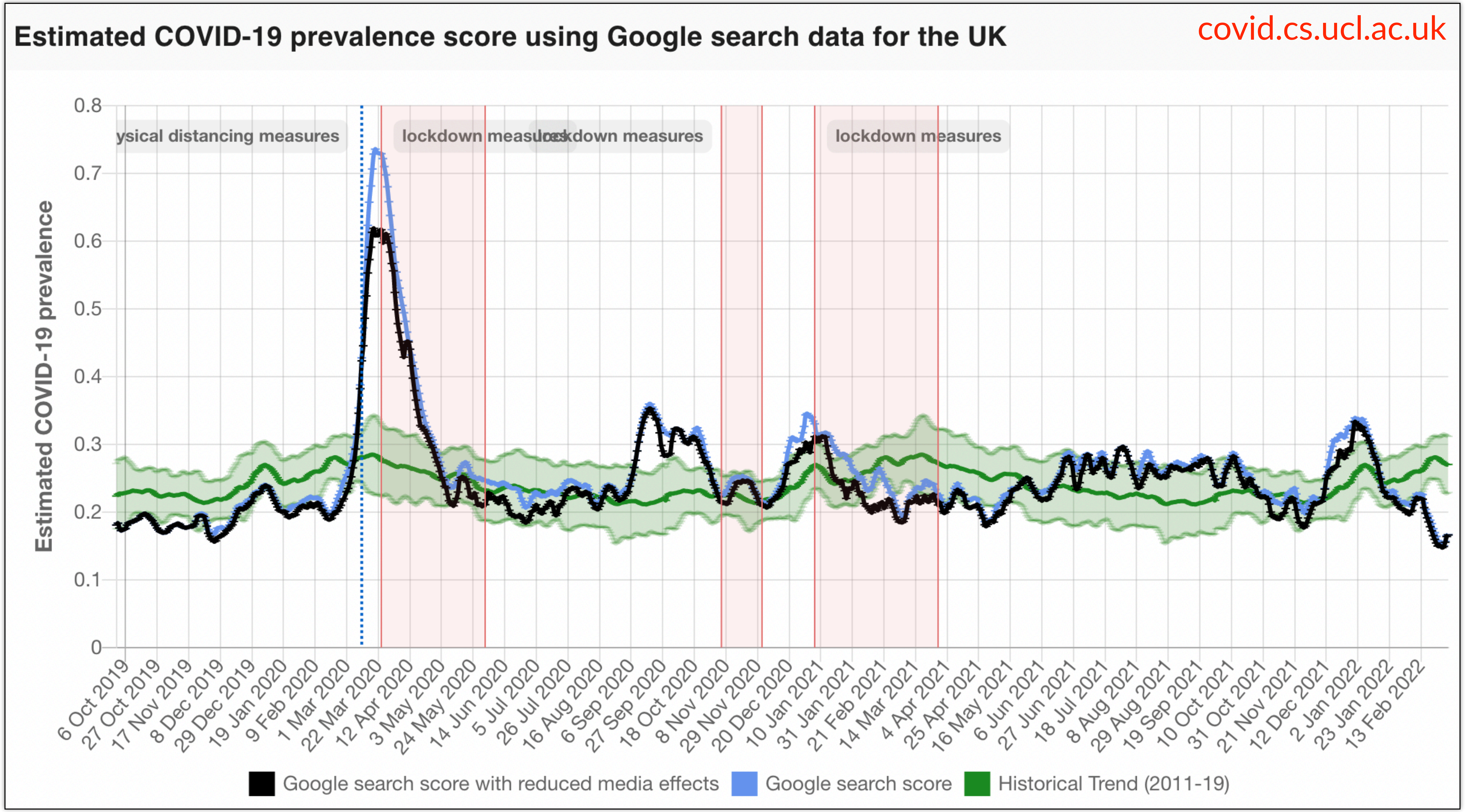
Limitations

- A thorough **evaluation** of our findings, *no matter our efforts to mitigate against confounding signals*, is not possible
  - No definitive ground truth exists
- Difficult to use national-level indicators for **policy making**
  - More *geographically granular models* are needed — there is data to support this now in some countries
    - pair-code.github.io/covid19_symptom_dataset
  - Better integration with conventional epidemiological models is required
- Limited applicability to locations with lower rates of **Internet access**

Translation and impact — Part of UK's COVID-19 surveillance

UK Health Security Agency

gov.uk/government/statistics/national-flu-and-covid-19-surveillance-reports-2023-to-2024-season

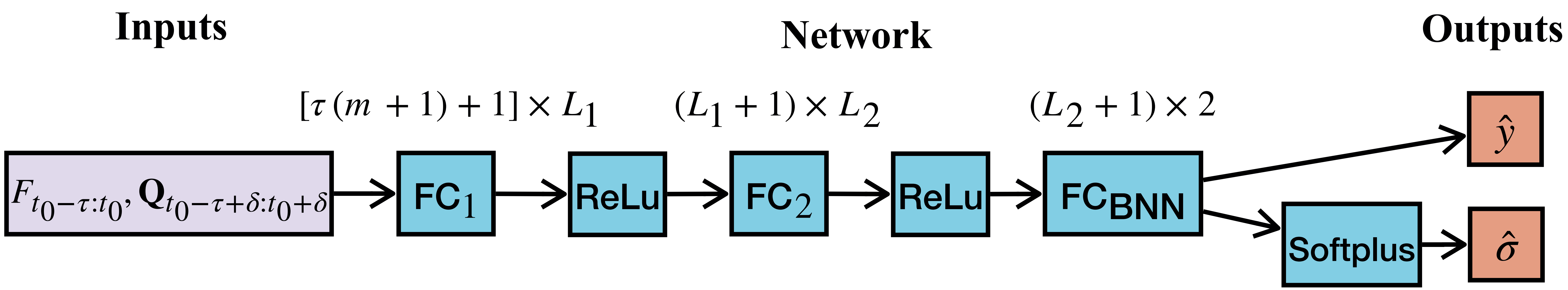


Part D

Disease forecasting with uncertainty

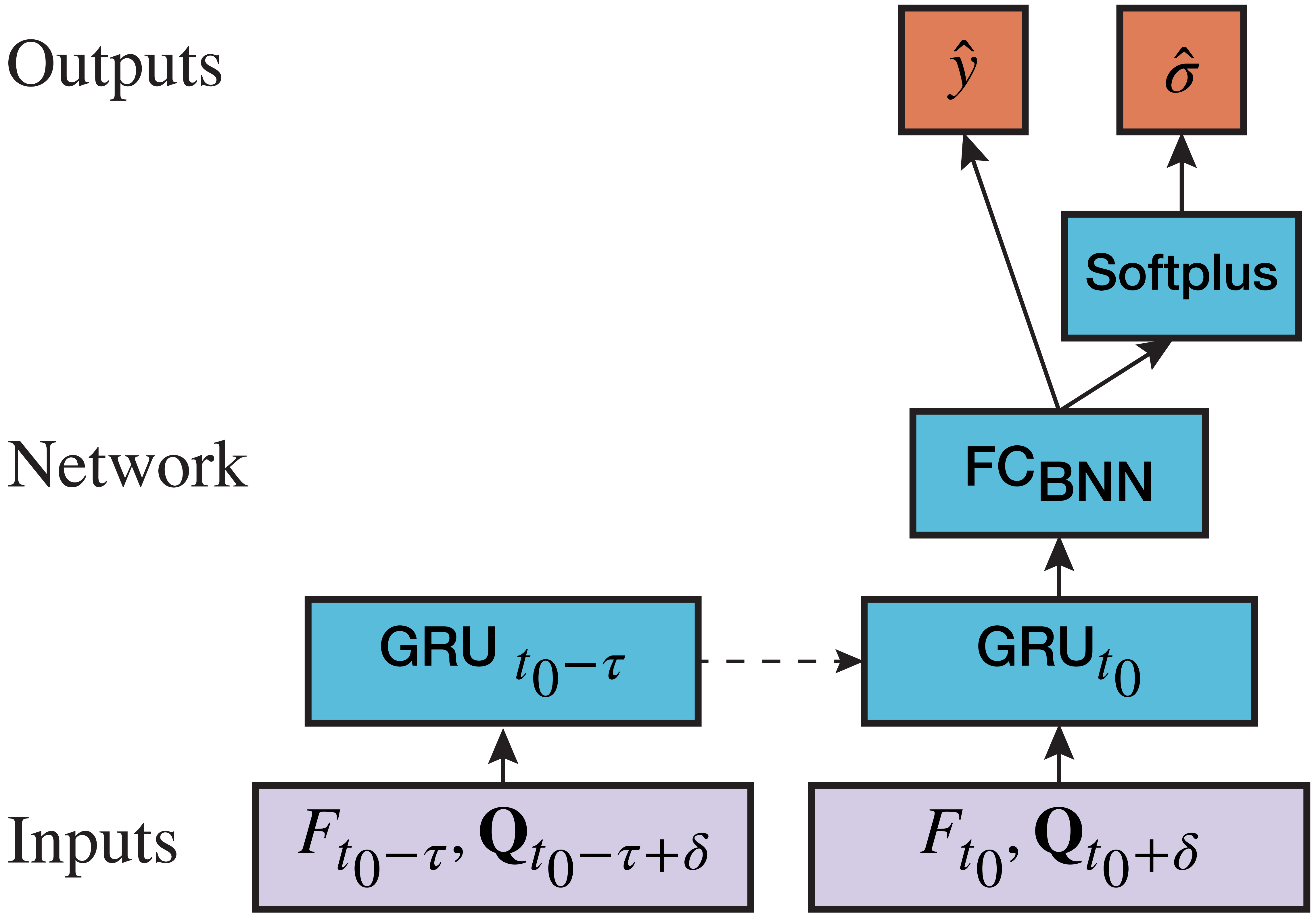
Morris, Hayes, Cox, Lampos (2023), *PLOS Comput. Biol.*

Neural networks for disease forecasting — Feedforward baseline



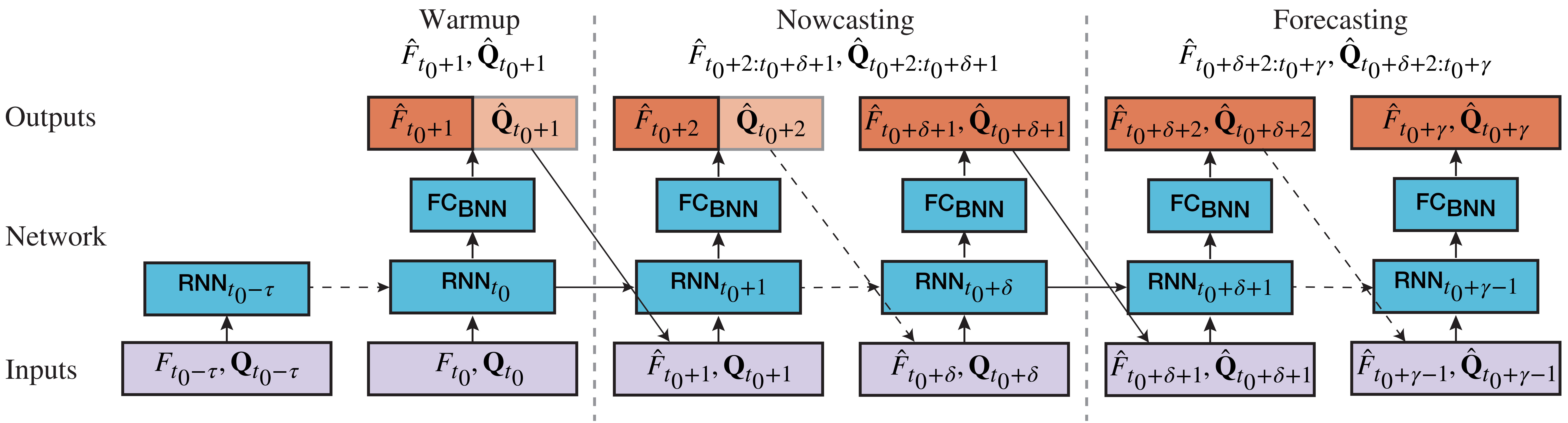
- **Input:** web search activity ($\mathbf{Q}$), previous ILI rates (F) with a temporal delay δ flattened over a window of days
- BNN denotes a fully connected Bayesian layer with a probability distribution over its weights
- **Data / aleatoric uncertainty:** by using negative log likelihood as our loss function to obtain a mean and a standard deviation for each forecast
- **Model / epistemic uncertainty:** by training the BNN layer using variational inference
- **Combine** data and model uncertainties by sampling the posterior of the NN's parameters
- Multiple **output estimates (samples)** are used to derive a forecast and its confidence intervals

Neural networks for disease forecasting — Simple RNN (SRNN)



- Replace FF layers with a GRU layer
- Input is not flattened as it becomes a time series sequence

Neural networks for disease forecasting — Iterative RNN (IRNN)



- Fully autoregressive, i.e. the network predicts all the input data for the next time step
- Feeds this data back to itself, unlimited forecasting horizon
- Initially for a certain some of the data (Google) is known to us (*we feed the actual data not the predicted ones*)
- **Limitation**: no way of understanding forecasting distance to calibrate uncertainty

Forecasting accuracy (ILI, US)

CRPS: Continuous Ranked Probability Score

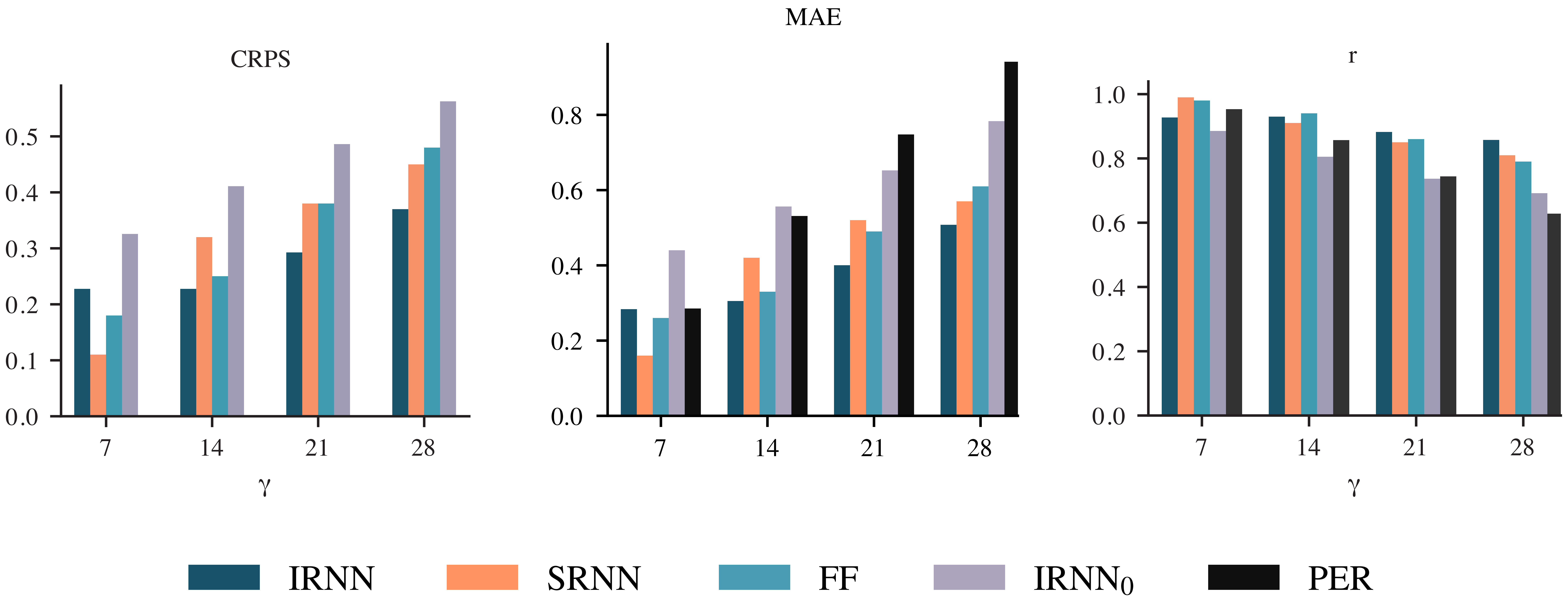
MAE: Mean Absolute Error

r: bivariate (linear) correlation

γ: days-ahead compared to the last ILI rate in the input, γ-14 days ahead compared the last search query frequency

Forecasting horizon	Accuracy metrics	FF	SRNN	IRNN
γ = 21	CRPS	0.39	0.41	**0.30**
	MAE	0.51	0.55	**0.42**
	r	0.85	0.83	**0.87**
γ = 28	CRPS	0.50	0.50	**0.38**
	MAE	0.63	0.64	**0.53**
	r	0.76	0.78	**0.84**

Forecasting accuracy (ILI, US)



IRNN without search query data

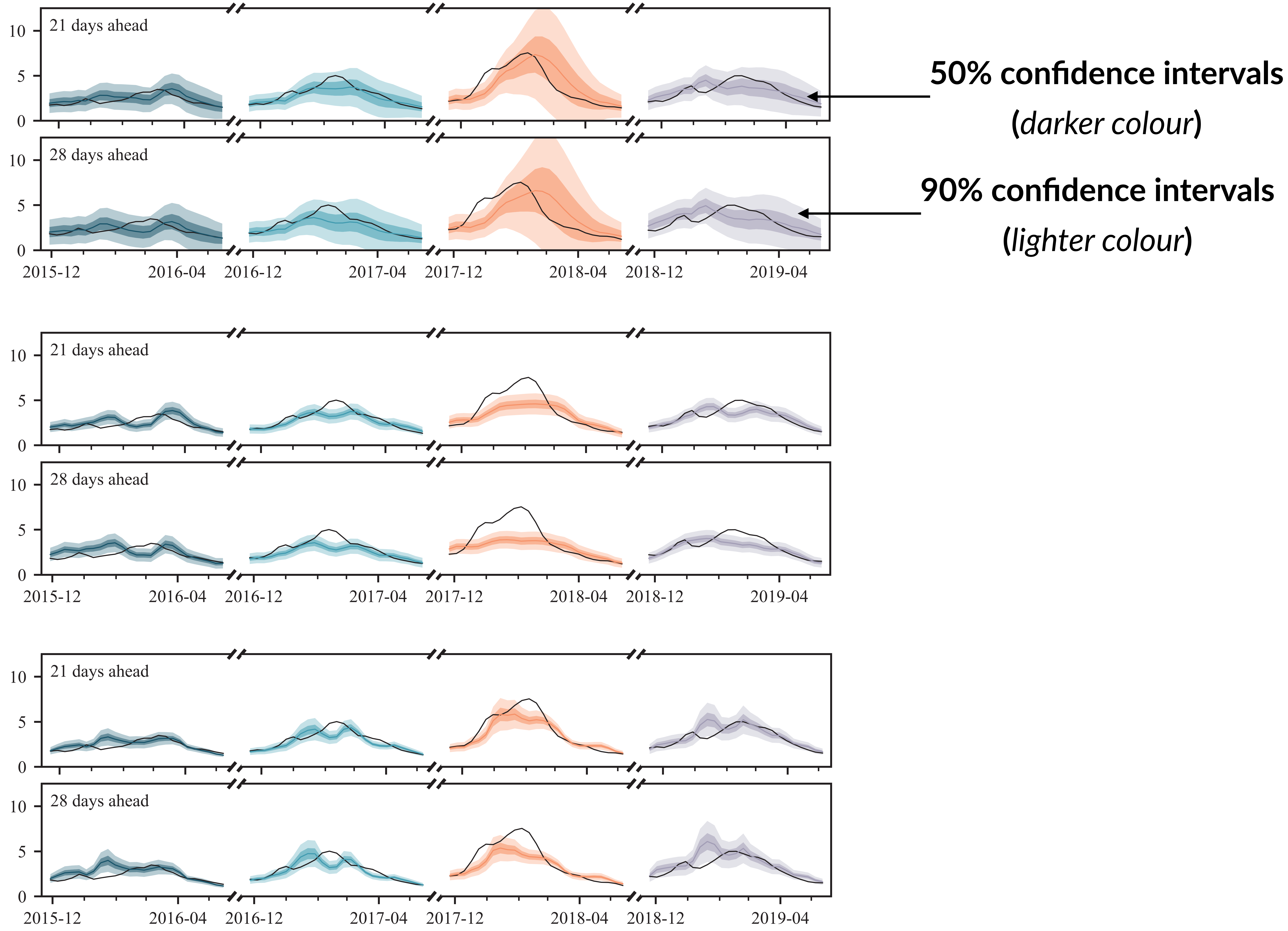
persistence model

ILI Forecasts (US)

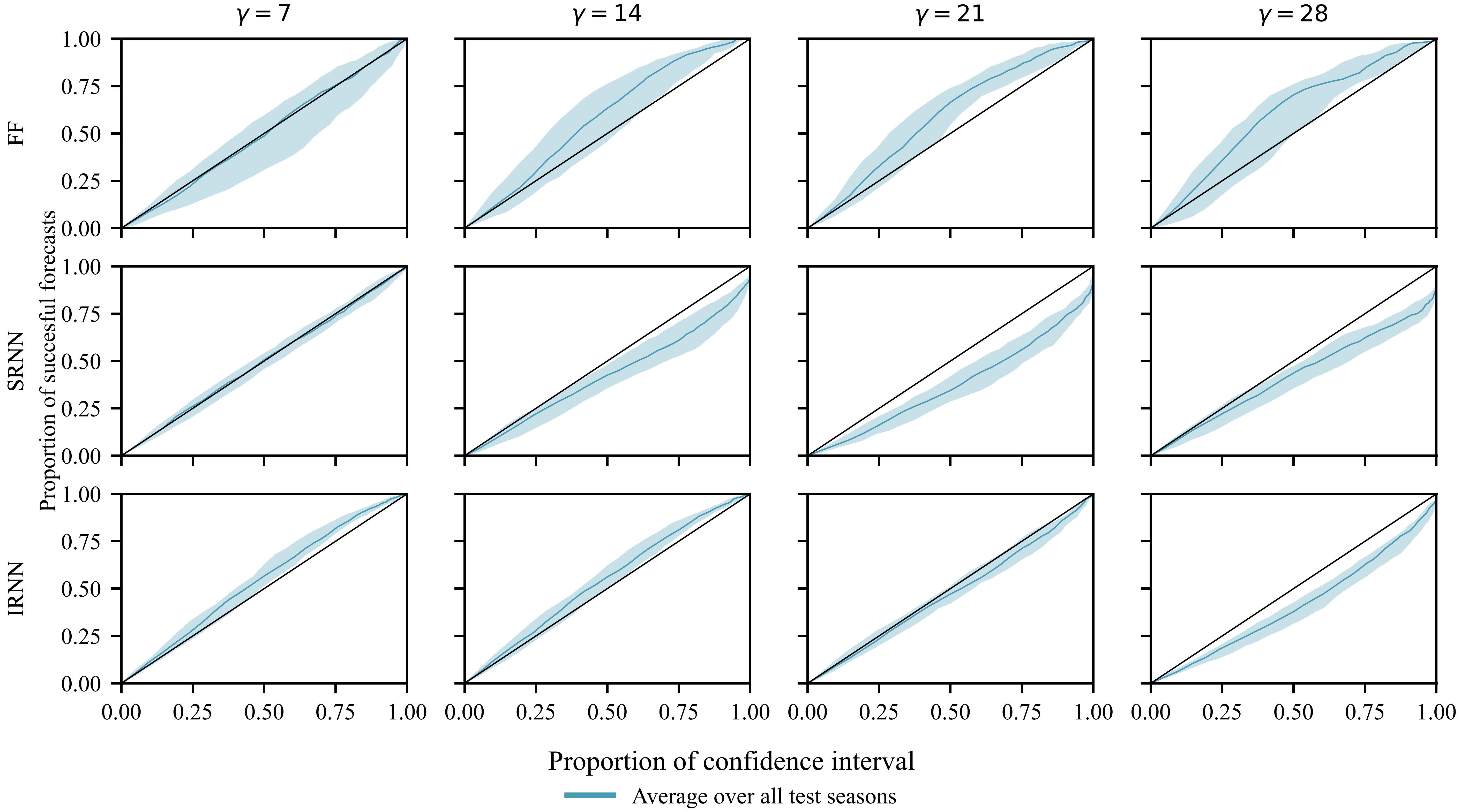
Feedforward NN

Simple RNN

Iterative RNN



Uncertainty calibration



How frequently (*probability*, *proportion*) the ground truth falls within a confidence interval of the same level: diagonal optimal calibration, above the diagonal signals an overestimation of uncertainty (*less confident*), below the diagonal signals an underestimation of uncertainty (*over confident*).

Comparison with a state-of-the-art mechanistic model ("*Dante*")

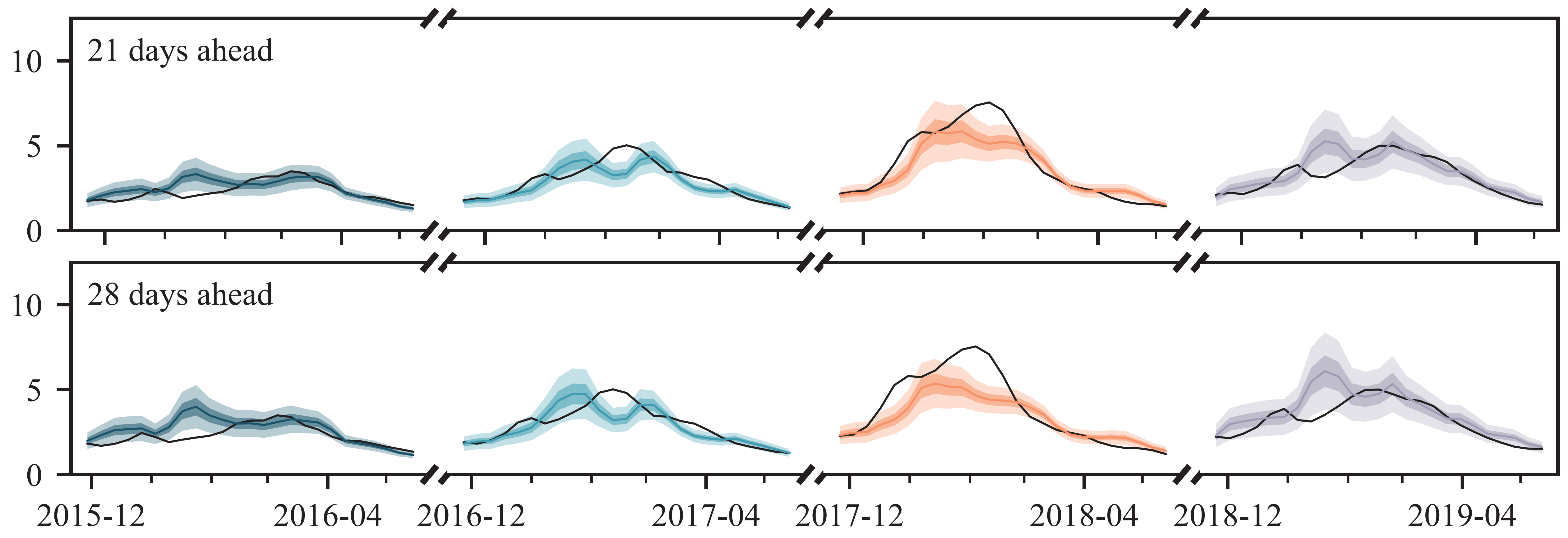
- State of the art performance based on a CDC competition
- Dante leverages information from US regions, our model does not
- Our model provides a much better accuracy, and you will see more meaningful uncertainty bounds as well

Forecasting horizon	**Accuracy metrics**	**Dante**	**IRNN**
$\gamma = 21$	MAE	0.53	**0.47**
	r	0.73	**0.81**
$\gamma = 28$	MAE	0.61	**0.60**
	r	0.68	**0.78**

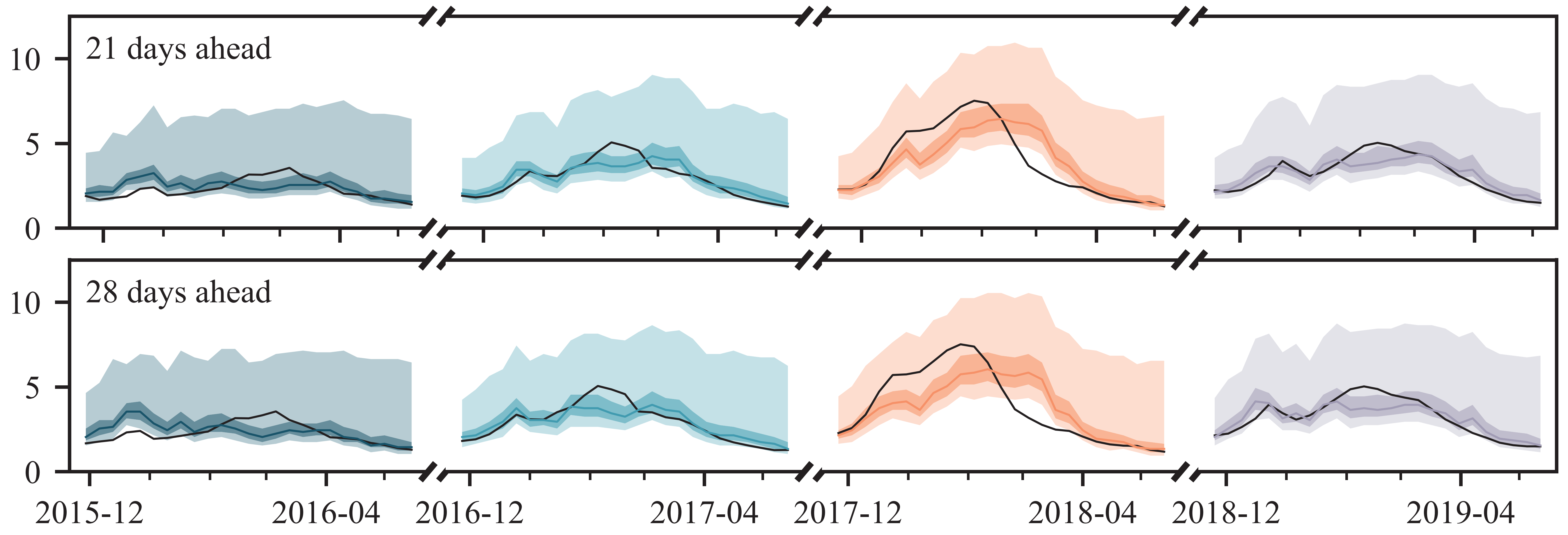
Osthus & Moran (2021), *Nat. Commun.*

Comparison with "*Dante*"

Iterative RNN



Dante



Take-aways, conclusions, future work

- **Web search activity** can be used for **infectious disease monitoring**
  - Google Flu Trends "*failed*" because of its methodological flaws
  - ML and NLP provide the tools to get this right
- We can **transfer** disease models based on web search data to locations that don't have (sufficient) syndromic surveillance data
- **Unsupervised models** based on web search activity
  - demand a **careful design**
  - could be very **informative** especially when nothing else works
- Searches about common COVID-19 symptoms are not necessarily great COVID-19 prevalence indicators
- Will we **continue to** use the plethora of data generated during this pandemic to develop better disease modelling techniques?

Take-aways and conclusions

- **Forecasting models** can provide invaluable insights that could inform policy
  - Lot of space for improvement in terms of accuracy
  - Evaluation needs to be more thorough
  - Models need to be constrained by our understanding of how an infectious disease spreads (*epidemiology*)
  - End-to-end architectures using SOTA developments in machine learning and NLP

- **These approaches need to be incorporated into public health systems**
  - Thorough evaluation across different locations and diseases
  - Development of accessible platforms and tools to share insights with experts
  - Knowledge transfer in collaboration with experts

References

1. Lampos, Miller, Crossan, Stefansen. *Advances in nowcasting influenza-like illness rates using search query logs*. Scientific Reports **5** (12760), 2015. doi:10.1038/srep12760
2. Zou, Lampos, Cox. *Transfer learning for unsupervised influenza-like illness models from online search data*. WWW '19, pp. 2505-2516, 2019. doi:10.1145/3308558.3313477
3. Lampos, Majumder, Yom-Tov *et al*. *Tracking COVID-19 using online search*. npj Digital Medicine **4** (17), 2021. doi:10.1038/s41746-021-00384-w
4. Eysenbach. *Infodemiology: tracking flu-related searches on the web for syndromic surveillance*. AMIA, pp. 244-248, 2006.
5. Polgreen, Chen, Pennock, Nelson. *Using internet searches for influenza surveillance*. Clinical Infectious Diseases **47** (11), pp. 1443-1448, 2008. doi:10.1086/593098
6. Ginsberg, Mohebbi, Patel *et al*. *Detecting influenza epidemics using search engine query data*. Nature **457**, pp. 1012–1014, 2009. doi:10.1038/nature07634
7. Wagner, Lampos, Cox, Pebody. *The added value of online user-generated content in traditional methods for influenza surveillance*. Scientific Reports **8** (13963), 2018. doi:10.1038/s41598-018-32029-6
8. Budd, Miller, Manning *et al*. *Digital technologies in the public-health response to COVID-19*. Nature Medicine **26**, pp. 1183-1192, 2020. doi:10.1038/s41591-020-1011-4
9. Rasmussen, Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
10. Lampos, Zou, Cox. *Enhancing feature selection using word embeddings: The case of flu surveillance*. WWW '17, pp. 695-704, 2017. doi:10.1145/3038912.3052622
11. Levy, Goldberg. *Linguistic regularities in sparse and explicit word representations*. CoNLL '14, pp. 171-180, 2014. doi:10.3115/v1/W14-1618
12. Boddington *et al*. *COVID-19 in Great Britain: epidemiological and clinical characteristics of the first few hundred (FF100) cases: a descriptive case series and case control analysis*. Bulletin WHO **99**, pp. 178-189, 2021. doi:10.2471/BLT.20.265603
13. Morris, Hayes, Cox, Lampos. *Neural network models for influenza forecasting with associated uncertainty using Web search activity trends*. PLOS Computational Biology **19** (8), 2023. doi:10.1371/journal.pcbi.1011392
14. Osthus, Moran. *Multiscale influenza forecasting*. Nature Communications **12** (2991), 2021. doi:10.1038/s41467-021-23234-5