Tracking COVID-19 using online search

Vasileios Lampos

Computer Science, UCL

@lampos

lampos.net

Presentation structure

A. Estimate 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

B. Transfer a disease model for one country to another country, based on web search activity (*transfer learning*)

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

Slides: lampos.net/talks

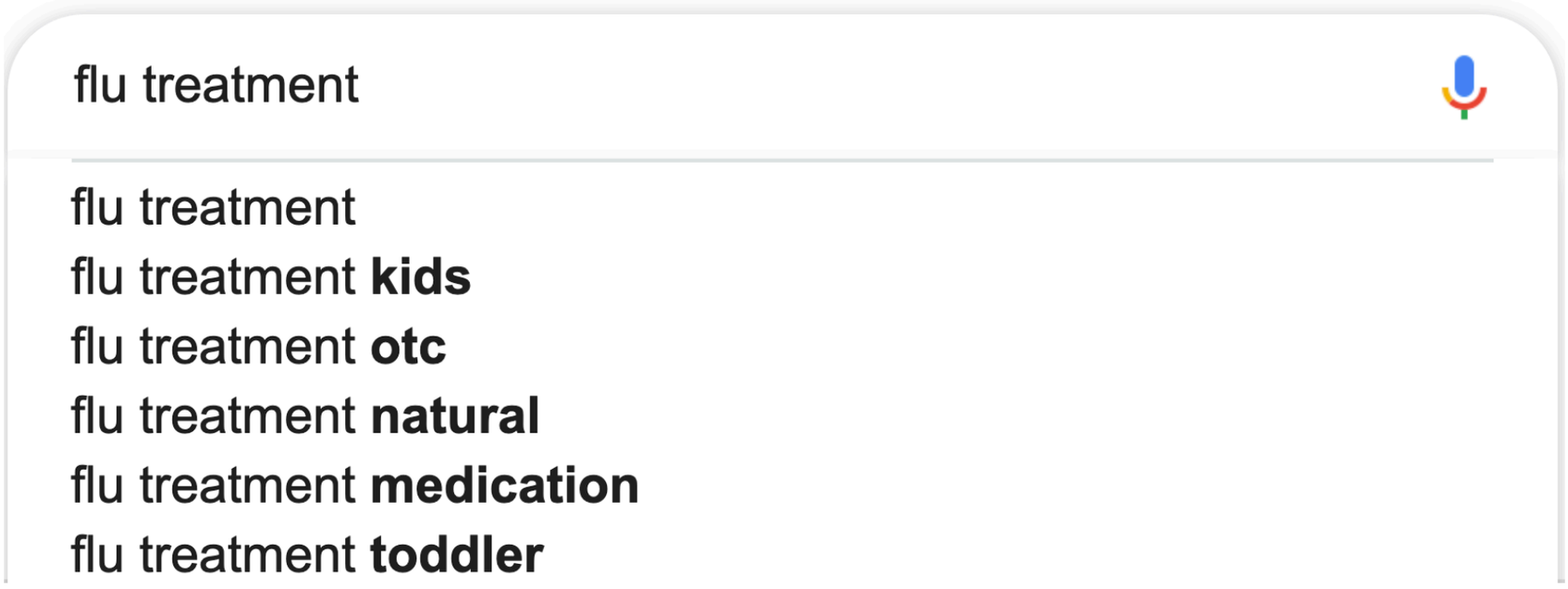
Part A

Estimating flu prevalence using web search activity

@lampos

lampos.net

From web searches to influenza rates



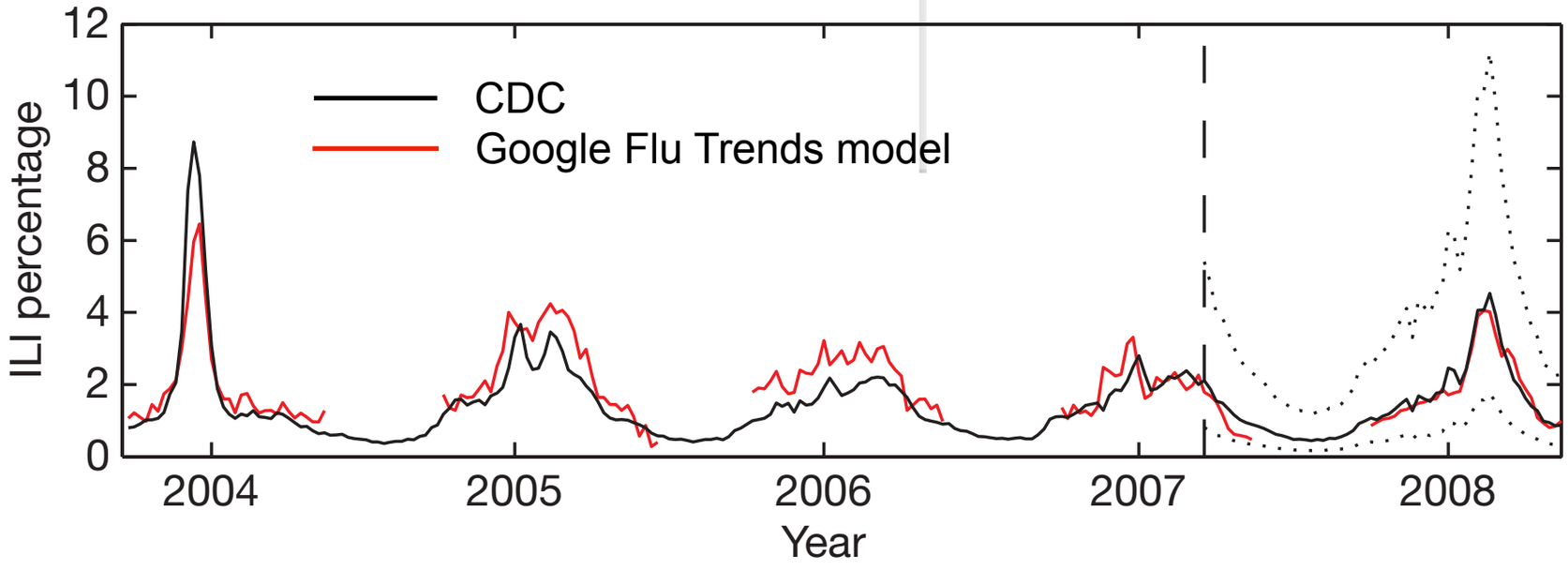
Eysenbach (2006), *AMIA*; Polgreen *et al.* (2008), *Clin. Infect. Dis.*; Ginsberg *et al.* (2009), *Nature*

From web searches to influenza rates

Google

flu treatment

flu treatment
flu treatment **kids**
flu treatment **otc**
flu treatment **natural**
flu treatment **medication**
flu treatment **toddler**



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*
  - broader, more granular *geographic coverage*
  - not affected by *closure days* and other *temporal biases*
  - *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?

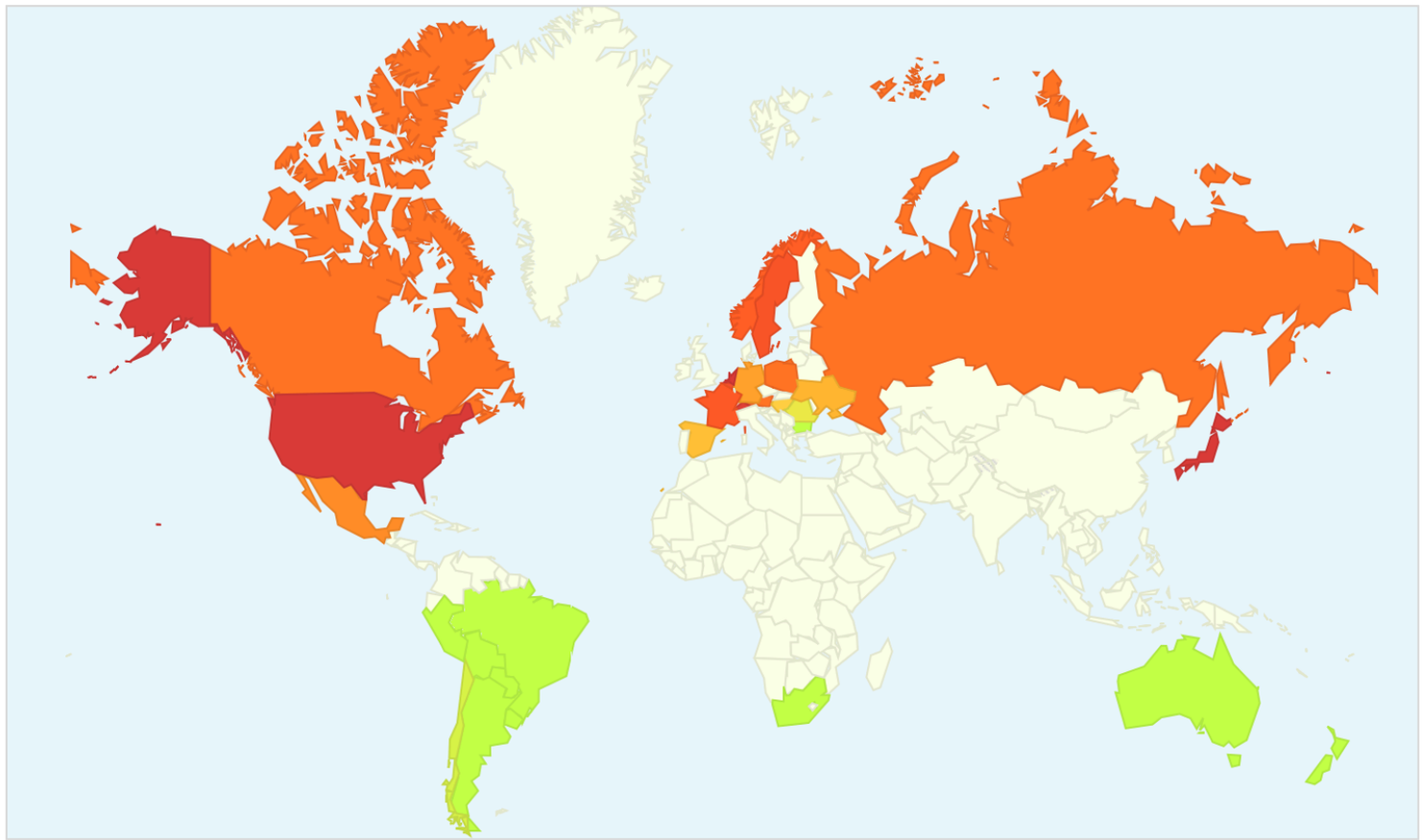
FAQ

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)

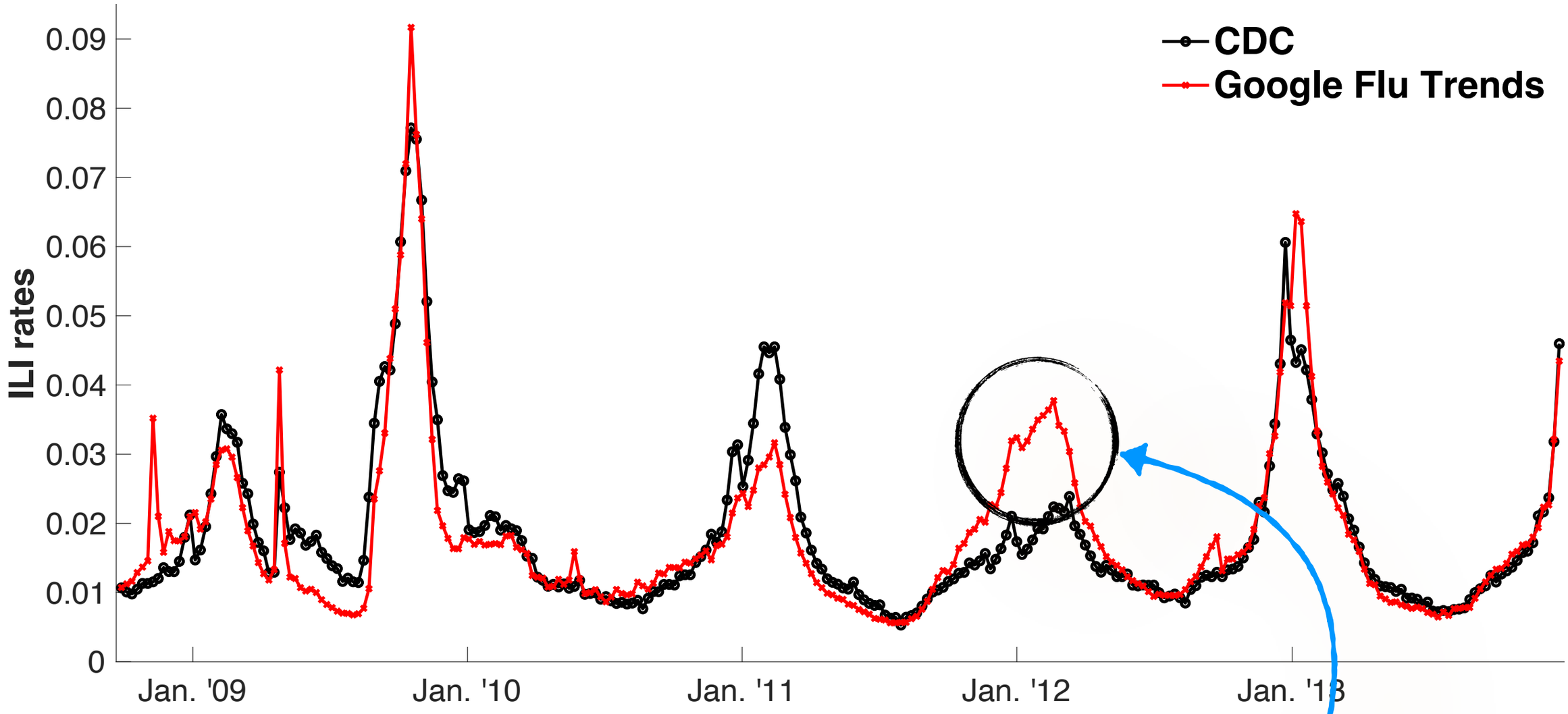
ϵ : independent, zero-centered noise

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.

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

Google Flu Trends (GFT) — *why / how did it fail?*

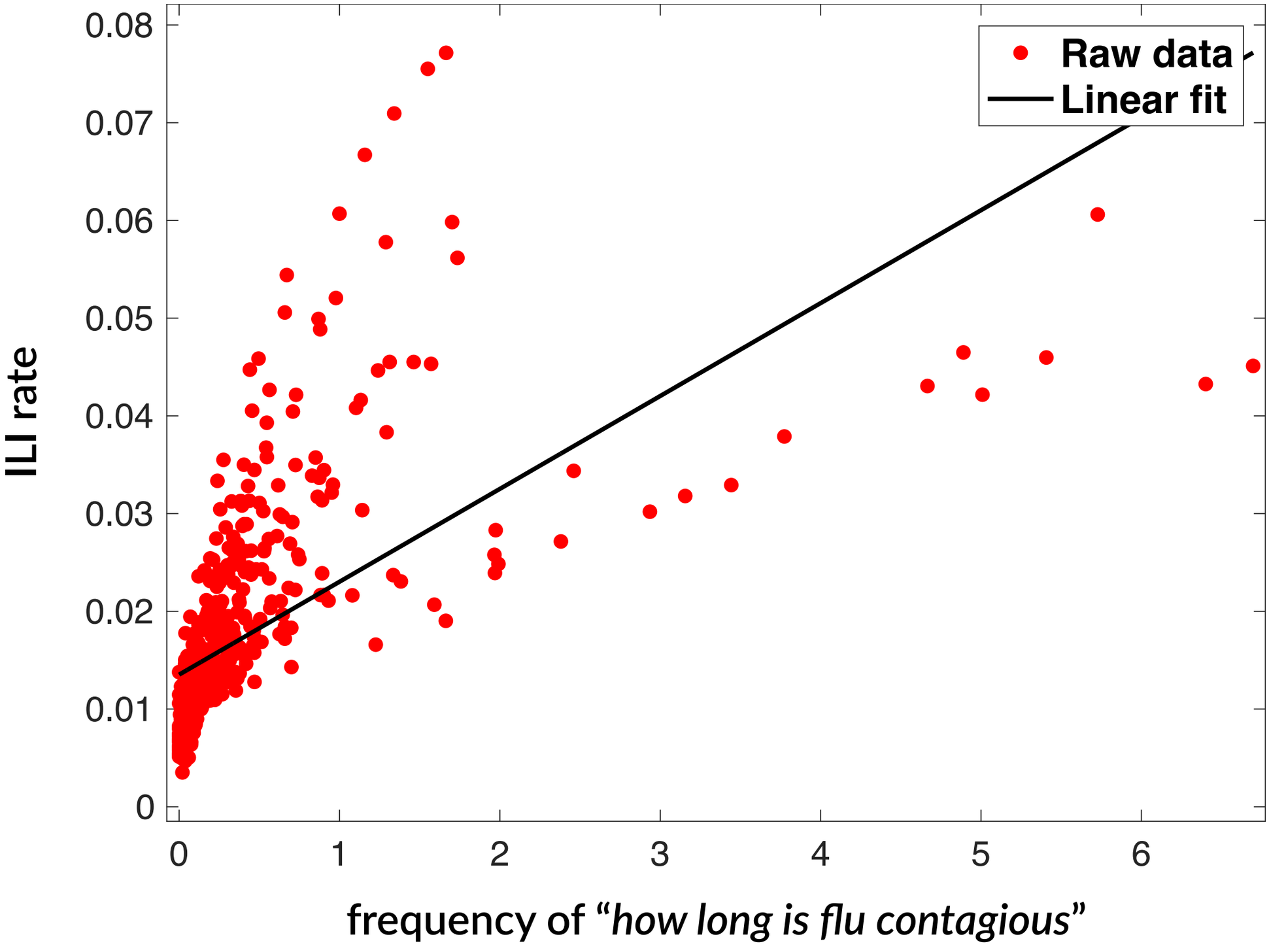


In the original Nature paper, the GFT model was **evaluated** on just ~1 flu season. *That should never be the case*!

rsv — 25%
flu symptoms — 18%
benzonatate — 6%
symptoms of pneumonia — 6%
upper respiratory infection — 4%

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

Nonlinearities



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

Multivariate kernels on search query clusters

Composite Gaussian Process (GP) kernel

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

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

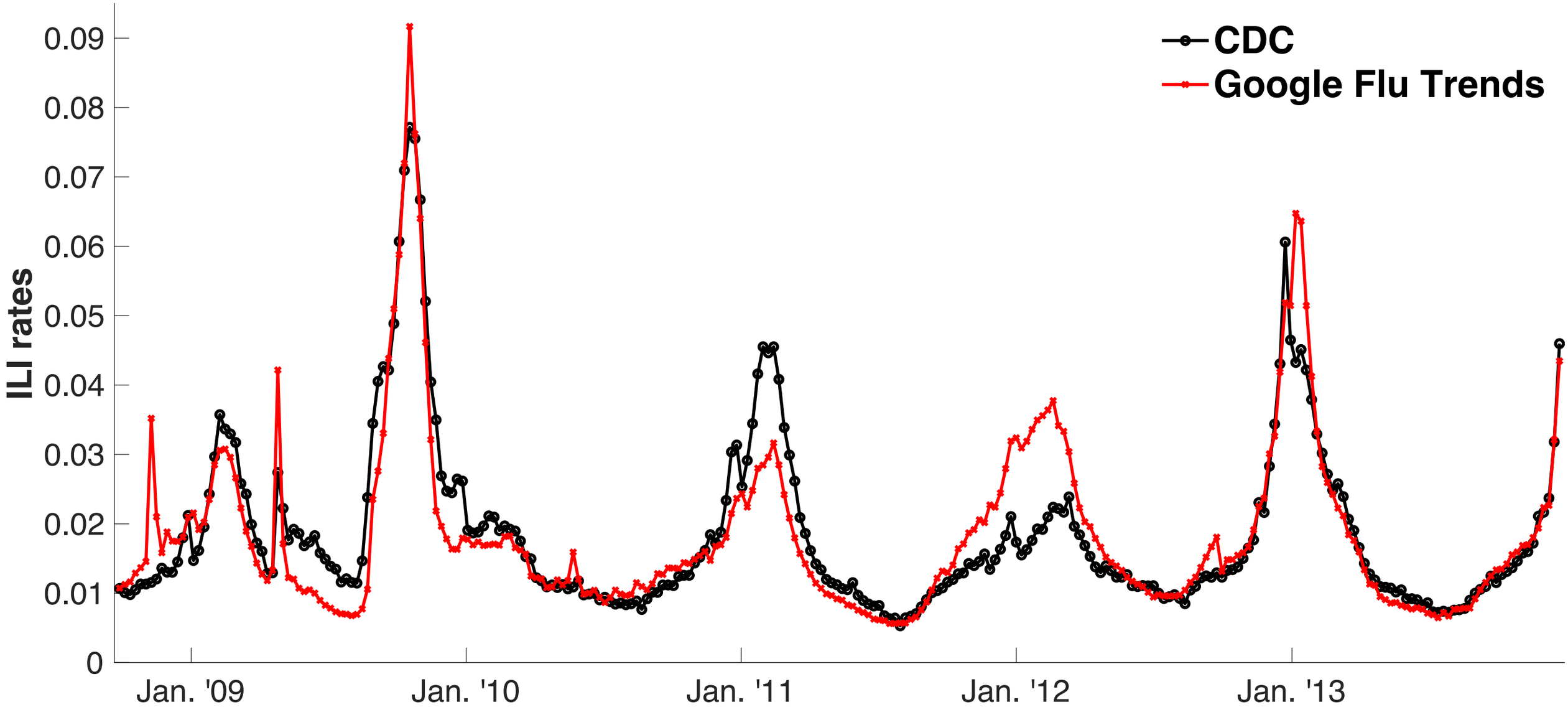
$\mathbf{c}_i, \mathbf{c}_i' \in \mathbb{R}_{\geq 0}^{z}, z < m, C$ query clusters based on frequency time series

Squared Exponential (SE) kernel

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

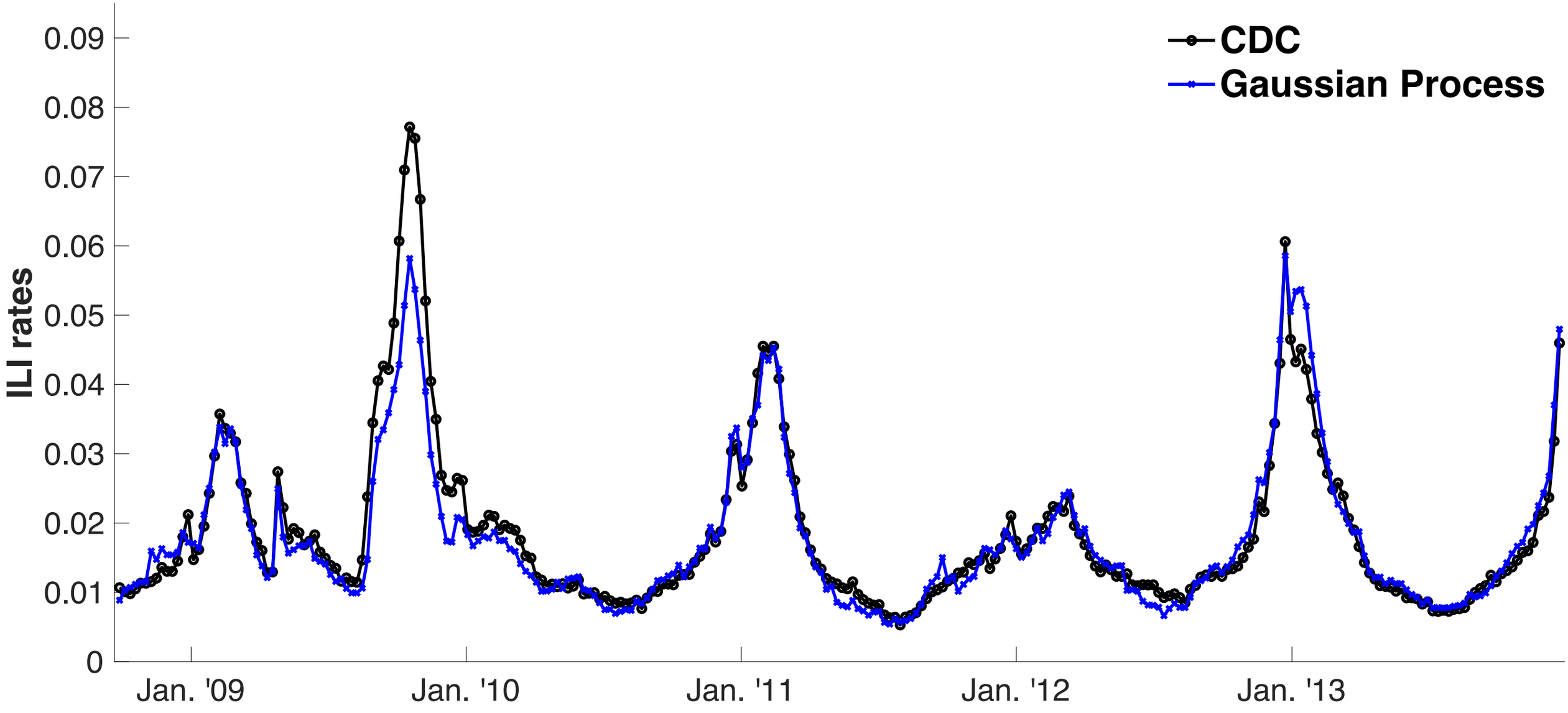
Lampos *et al.* (2015), *Sci. Rep.*; Rasmussen, Williams (2006), *MIT Press*

Modelling ILI rates with GP kernels



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

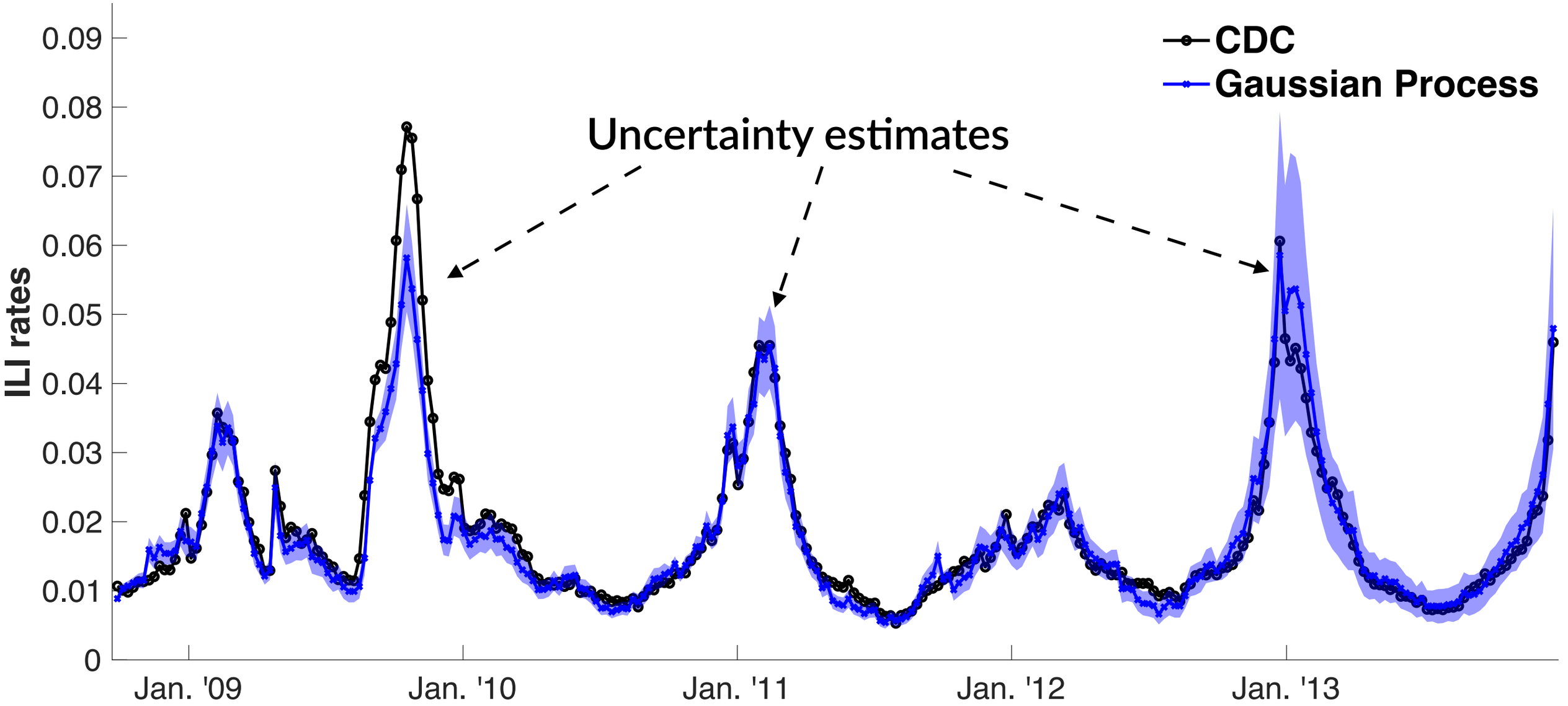
Modelling ILI rates with GP kernels



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

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

Modelling ILI rates with GP kernels



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

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

Autoregression (ARIMAX)

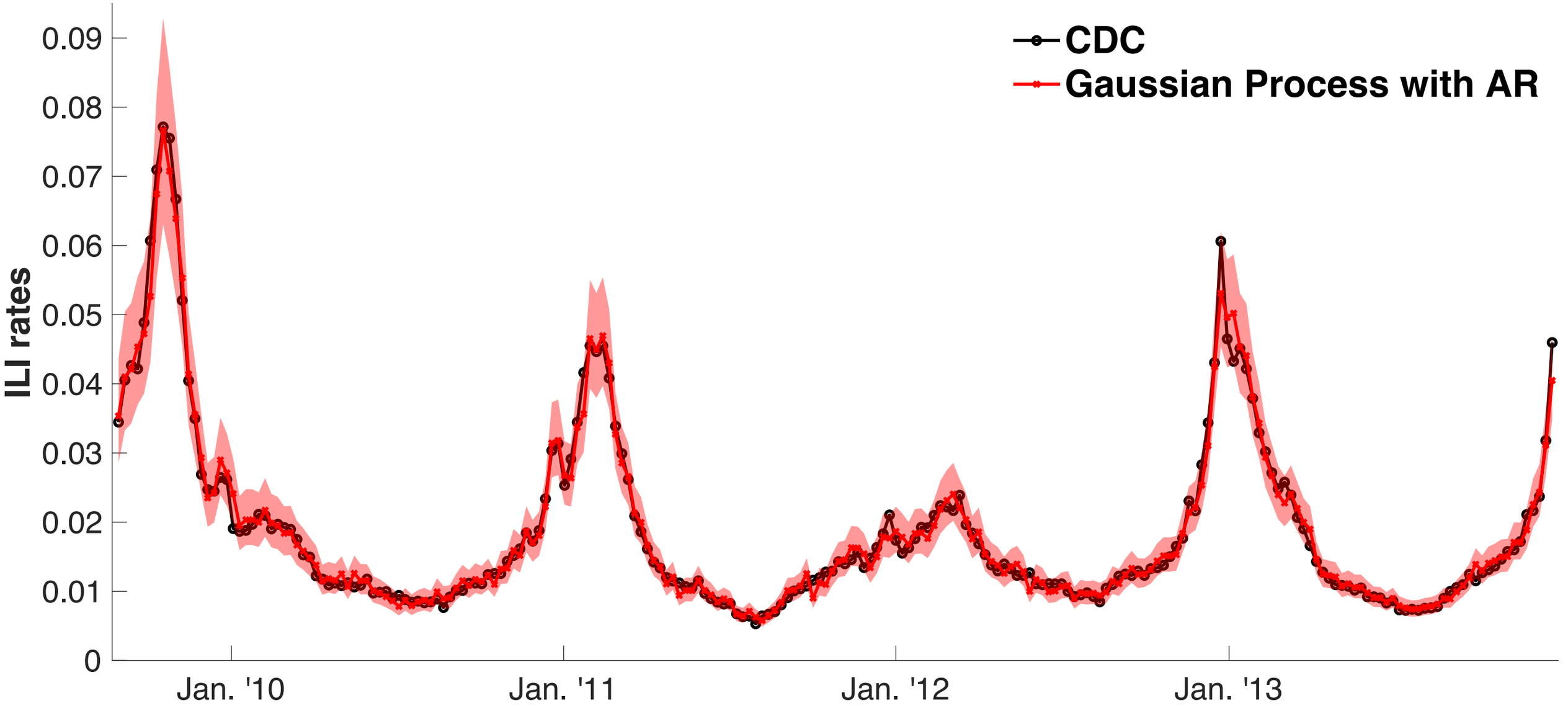
$$y_t = \underbrace{\sum_{i=1}^{p} \phi_i y_{t-d} + \sum_{i=1}^{J} \omega_i y_{t-52-i}}_{\text{AR and seasonal AR}} + \underbrace{\sum_{i=1}^{q} \theta_i \epsilon_{t-d} + \sum_{i=1}^{K} \nu_i \epsilon_{t-52-i}}_{\text{MA and seasonal MA}} + \cdots$$

$$\underbrace{\sum_{i=1}^{D} w_i h_{t,i}}_{\text{GP estimates}} + \epsilon_t$$

- d weeks delay in including past ILI rates as reported by CDC
- Choose model parameters based on the 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 GP kernels and ARIMAX



- 1 week delay in incorporating historical CDC estimates into an autoregressive (AR) formulation using ARMAX
- **27%** MAE reduction compared to GFT with AR, **52%** over the GP model without AR (*benefits increase as CDC data incorporation delay increases*)
- **.99** bivariate correlation with CDC

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

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**
- No need to use a supervised solution (*hard to obtain labels*)
- **Hybrid combination** this with previous feature selection regimes

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

Query selection based on distributional semantics

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

$\mathbf{e}_{(\cdot)}$: 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 $\mathbb{C}$

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

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

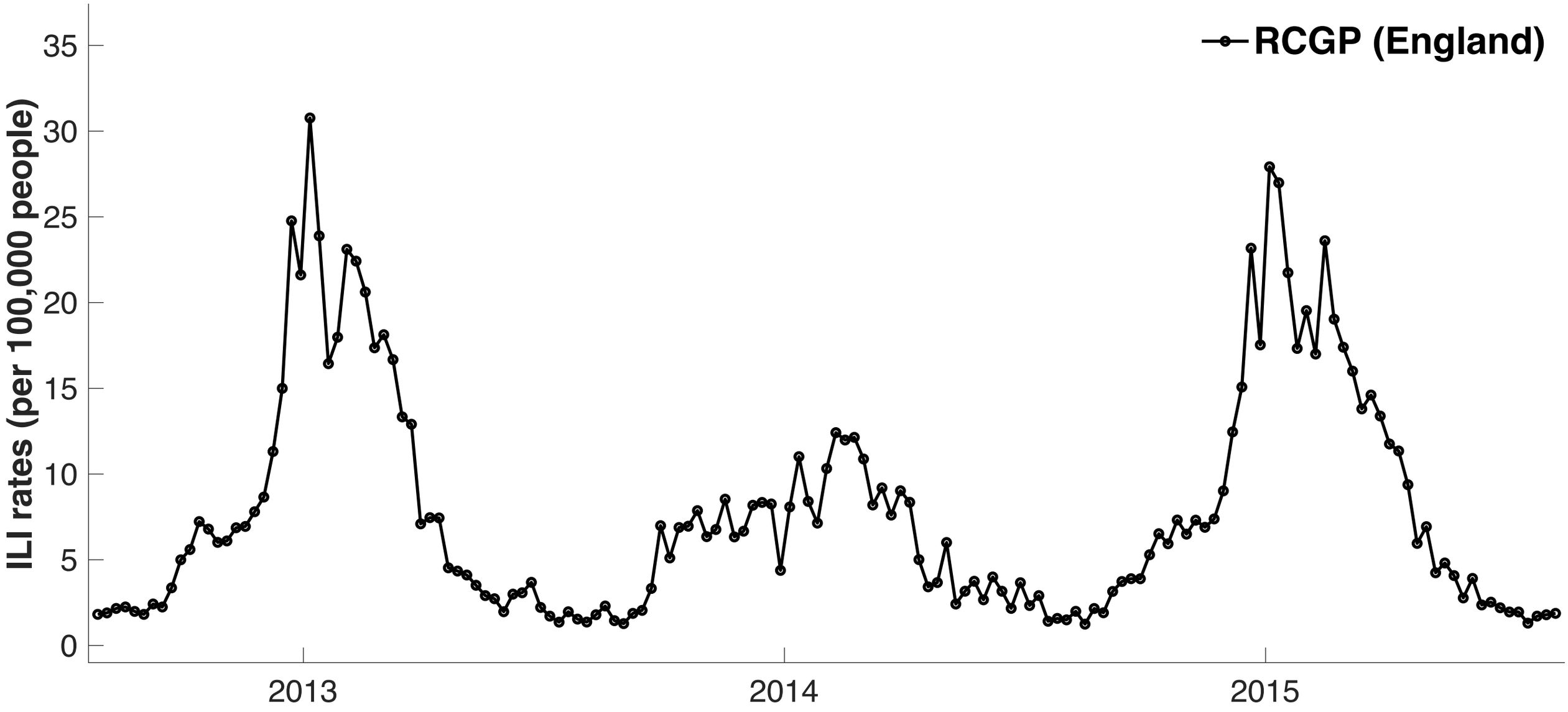
Lampos, Zou, Cox (2017), *WWW '17*; Levy, Goldberg (2014), *CoNLL '14*

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

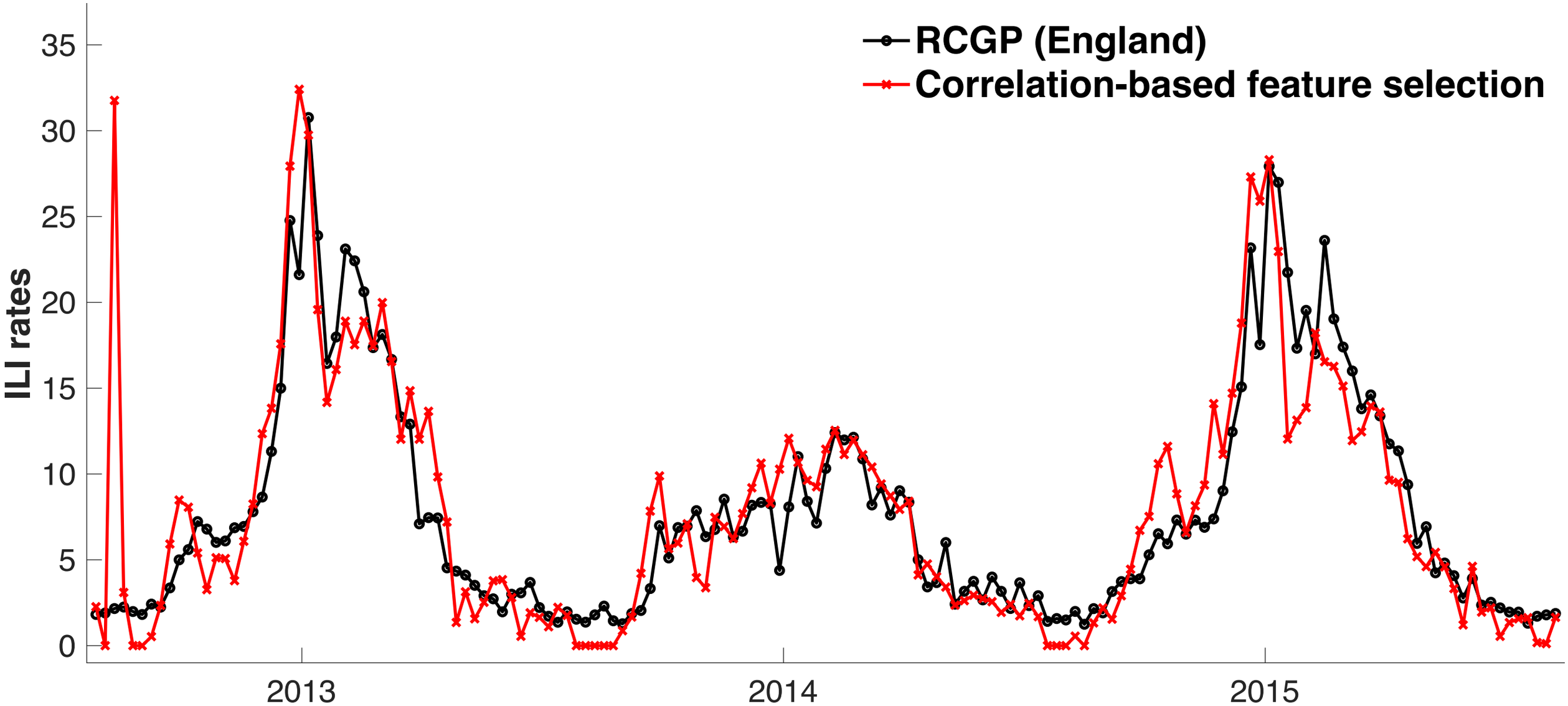
Lampos, Zou, Cox (2017), *WWW '17*

Feature selection based on *r* and reg. regression



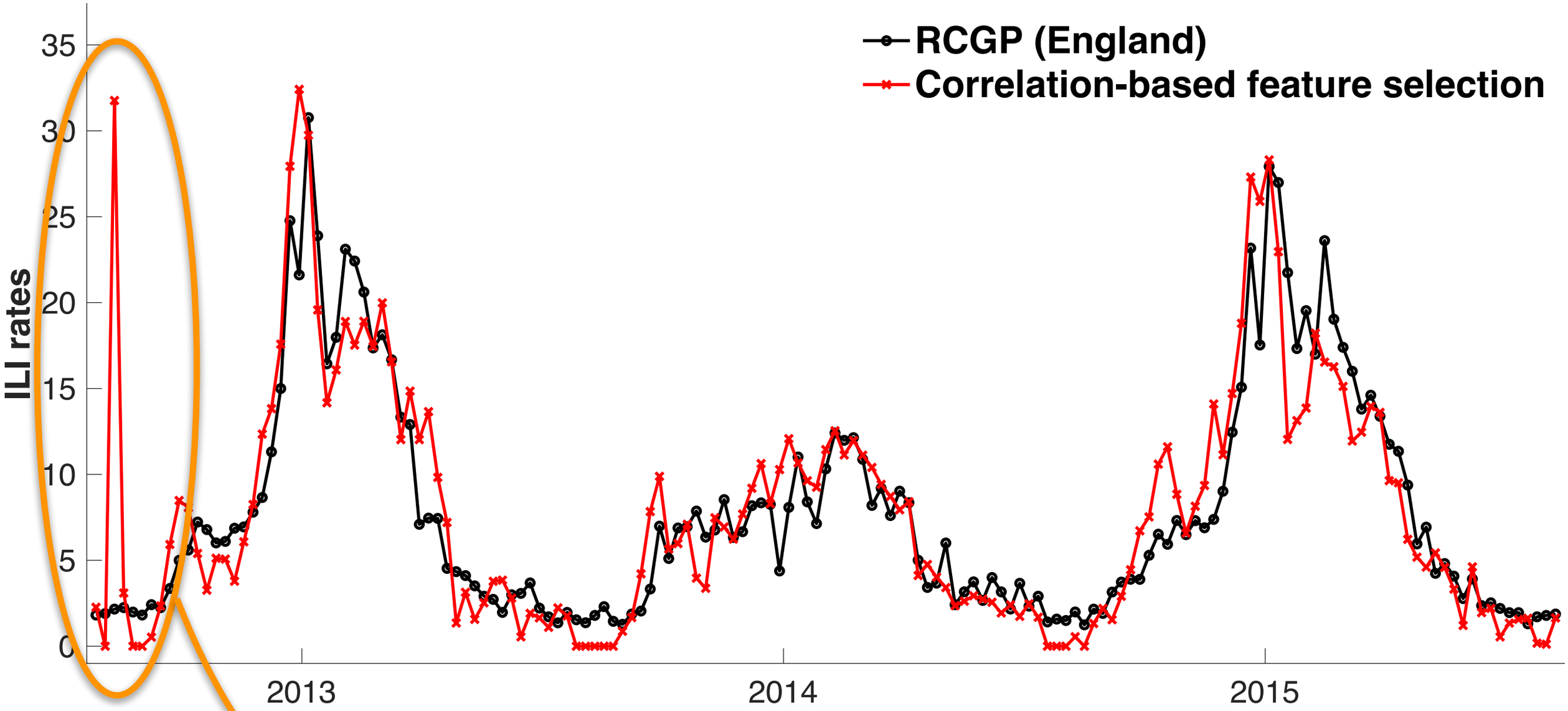
Lampos, Zou, Cox (2017), *WWW '17*

Feature selection based on *r* and reg. regression



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

Feature selection based on *r* and reg. regression



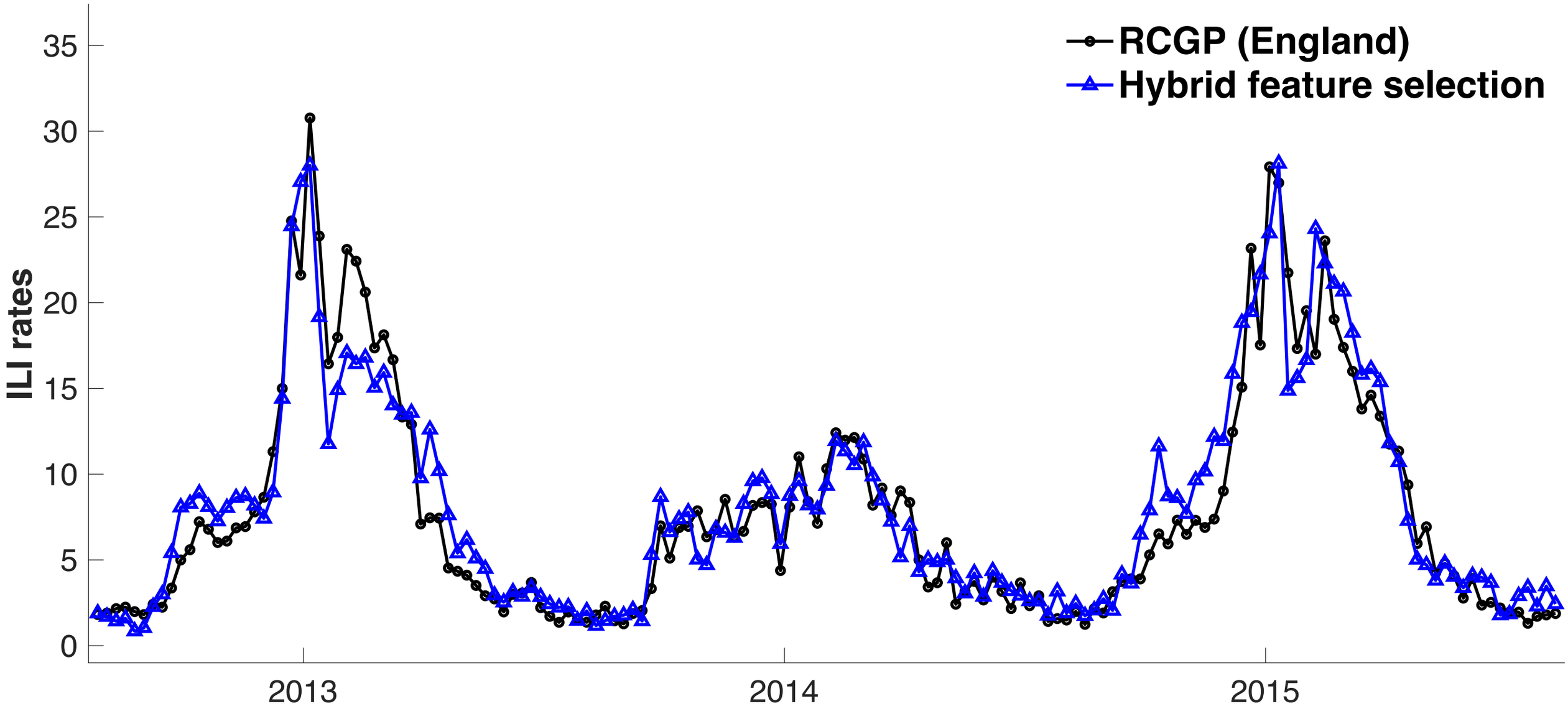
Examples of **problematic** query selections

prof. *surname*: 70%
name surname: 27%
heating oil: 21%

name surname recipes: 21%
blood game: 12.3%
swine flu vaccine side effects: 7.2%

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

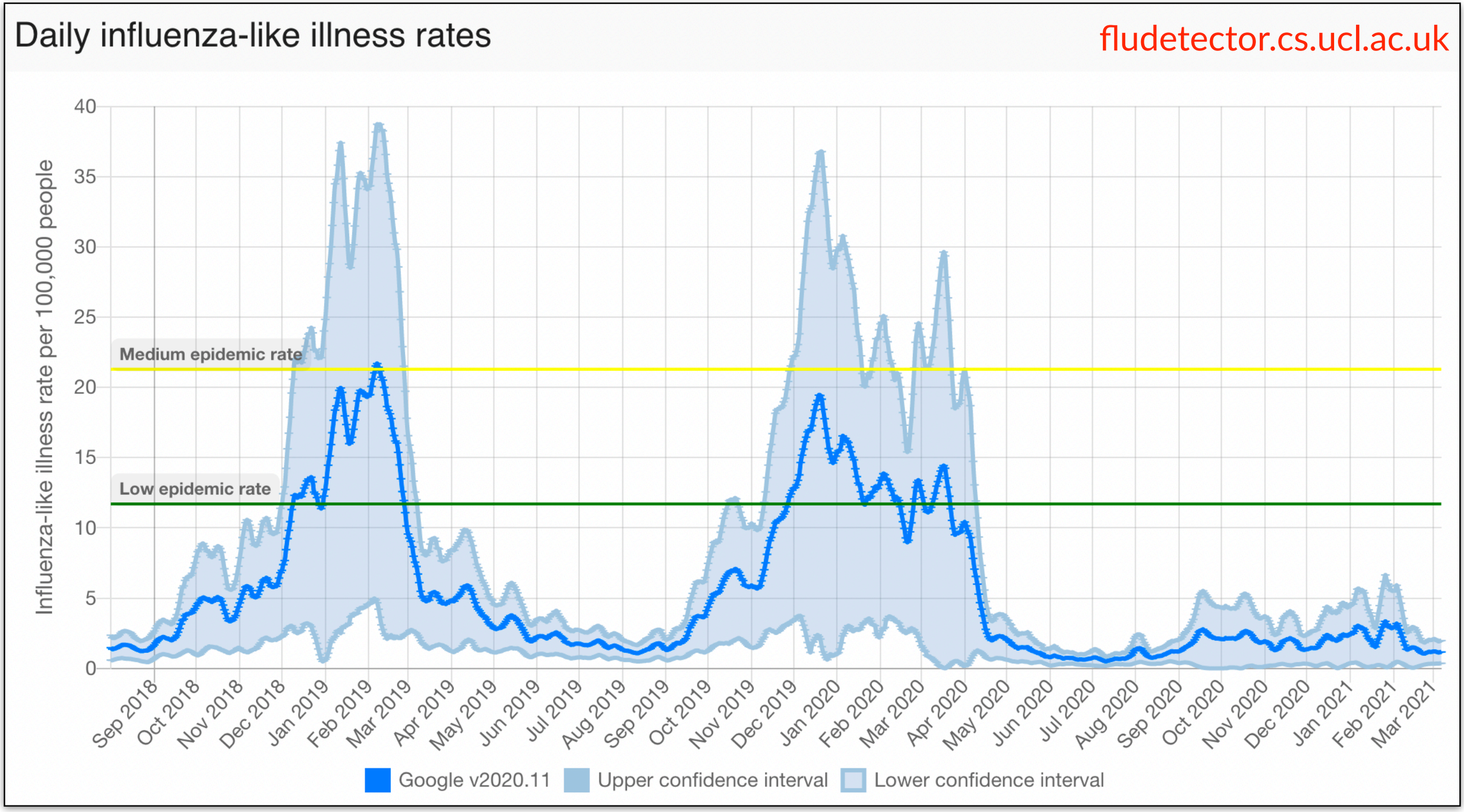
Hybrid feature selection



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

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

Flu detector



Public Health England

gov.uk/government/statistics/national-flu-and-covid-19-surveillance-reports

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*
- 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.*

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 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

@lampos

lampos.net

Transfer learning for flu models from web searches

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

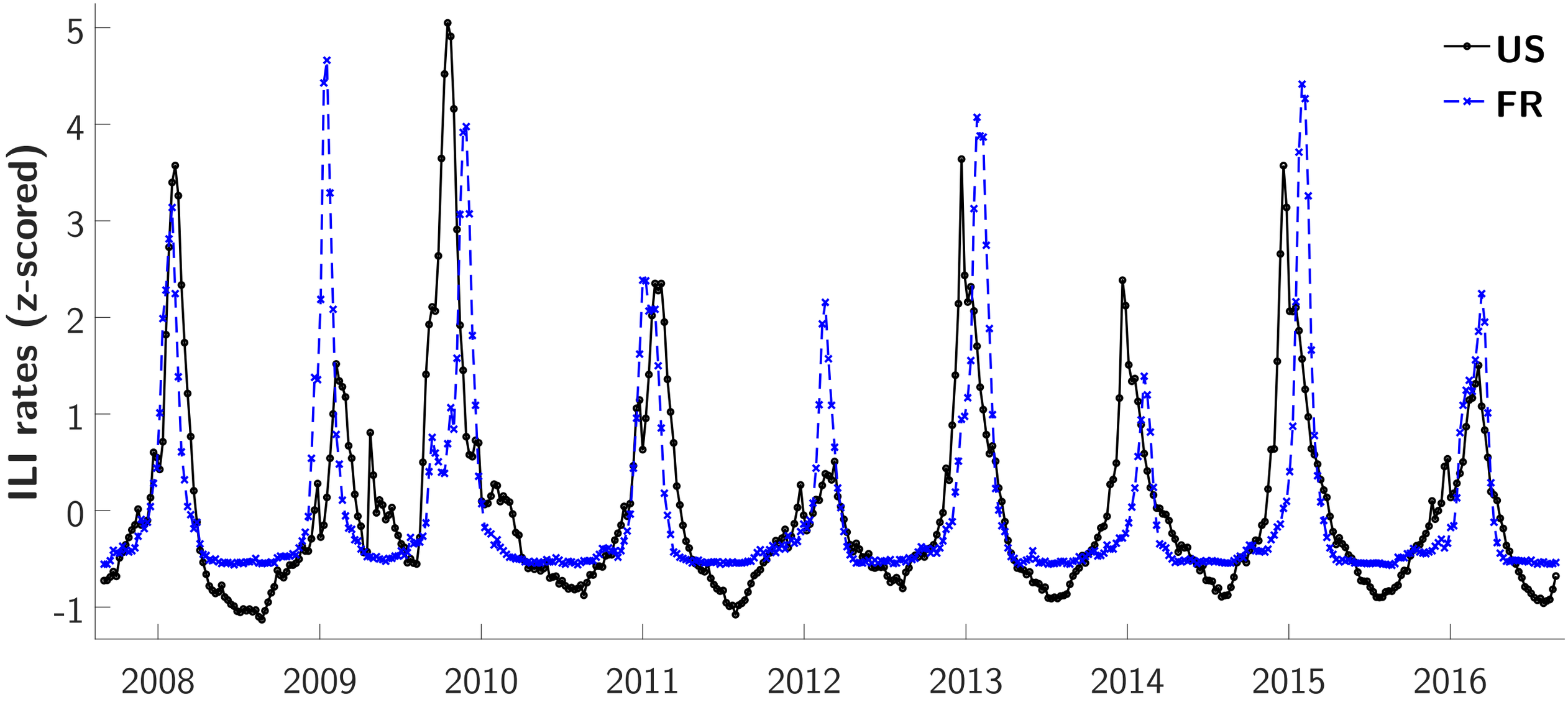
- **Transfer learning** *for estimating flu rates*
  - Locations: **source** (*no missing data*), **target** (*no disease rates*)
  - regularised **regression model** for a source location based on web search activity and historical disease rates
  - **map search queries** from the source to the target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

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

Transferring a flu model from US to France

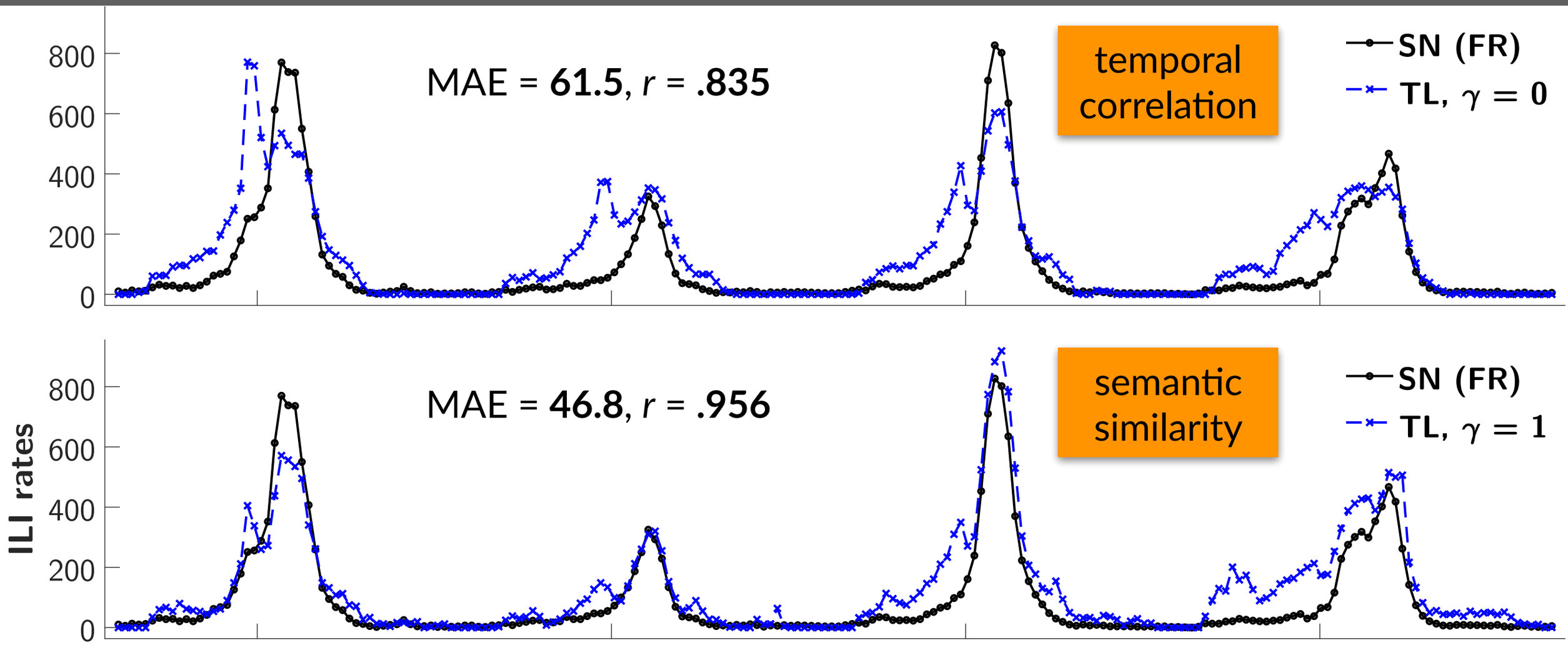
How similar are the flu rates in the US and **France**?

- temporal differences (e.g. different onset/peak moments)
- intensity differences



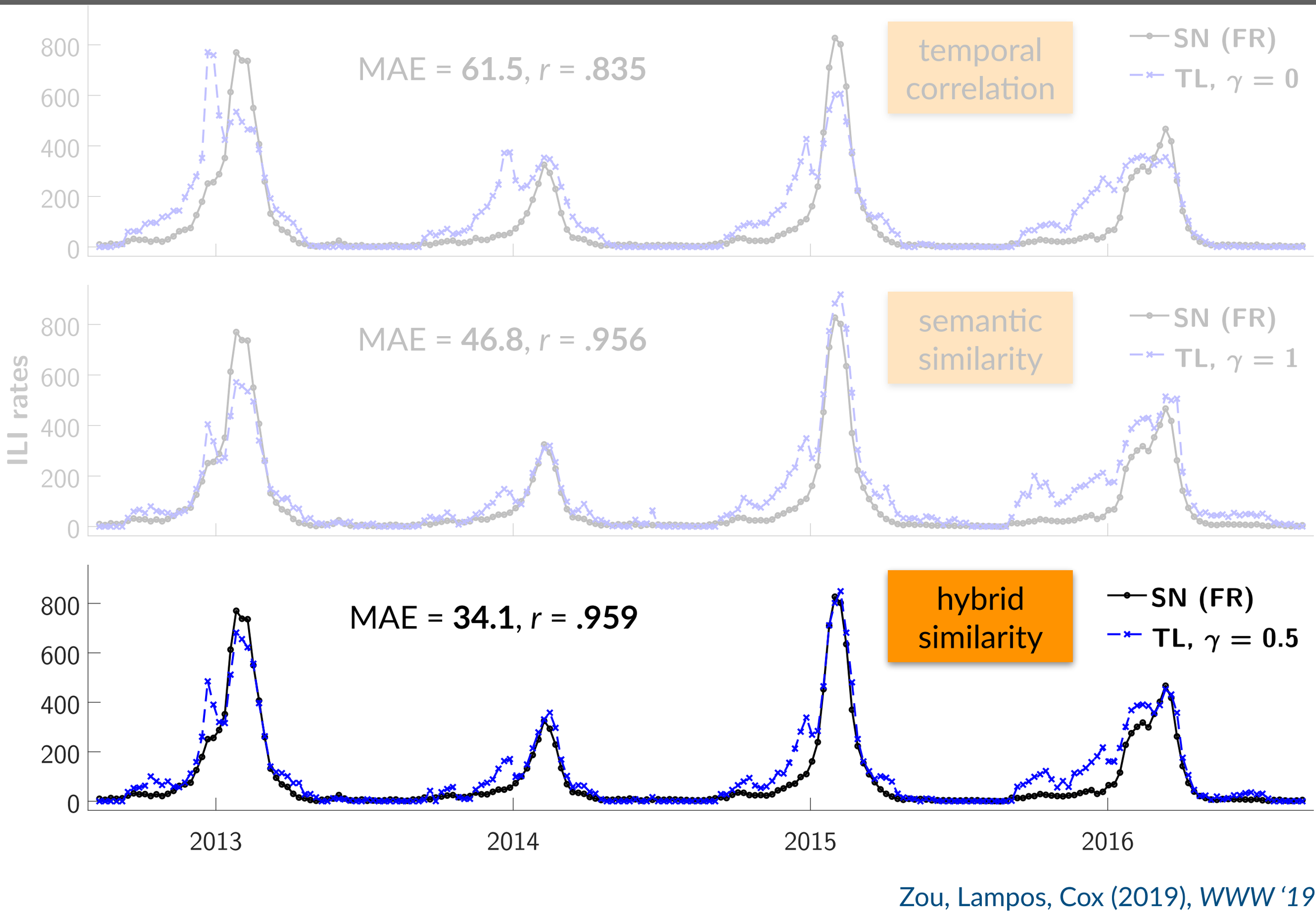
Zou, Lampos, Cox (2019), *WWW '19*

Transferring a flu model from US to France



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

Transferring a flu model from US to France

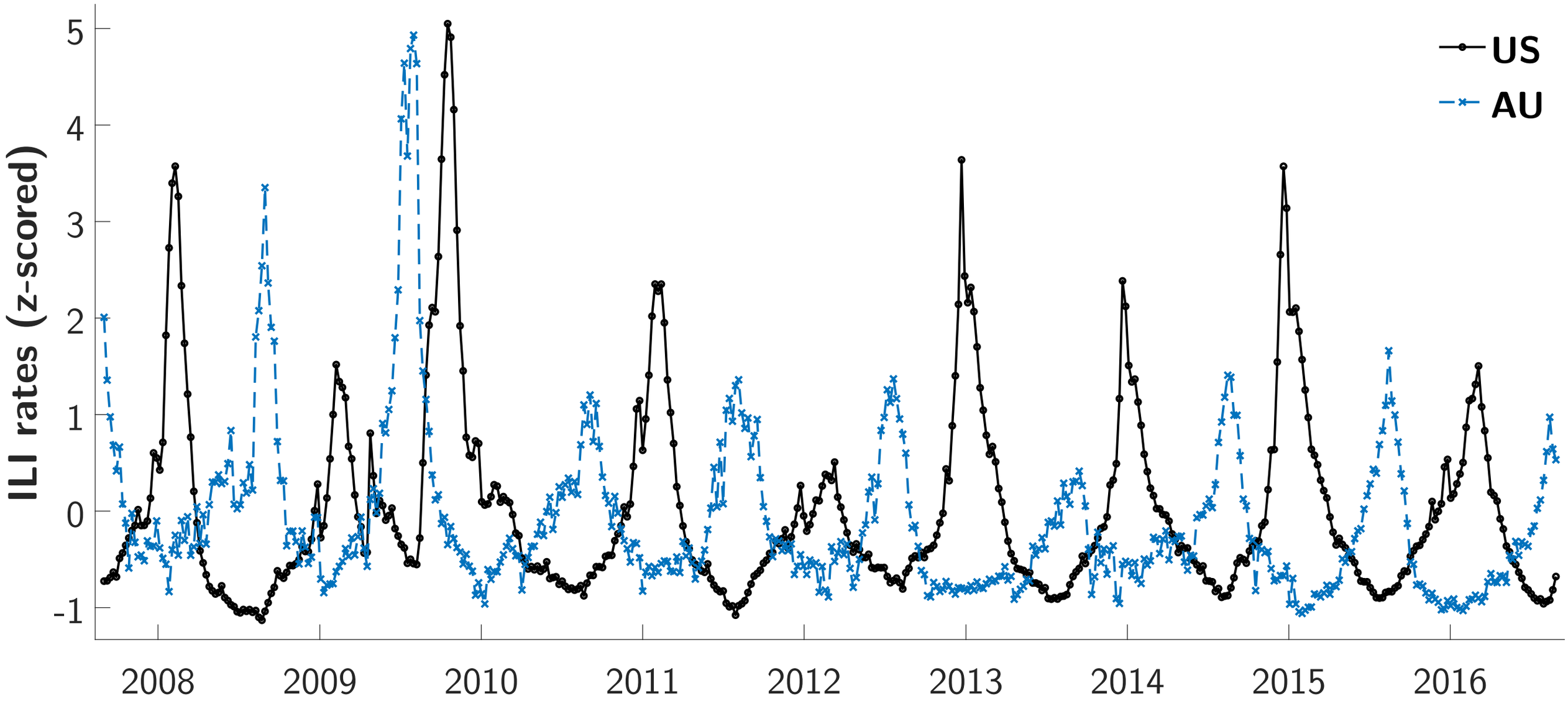


Zou, Lampos, Cox (2019), *WWW '19*

Transferring a flu model from US to Australia

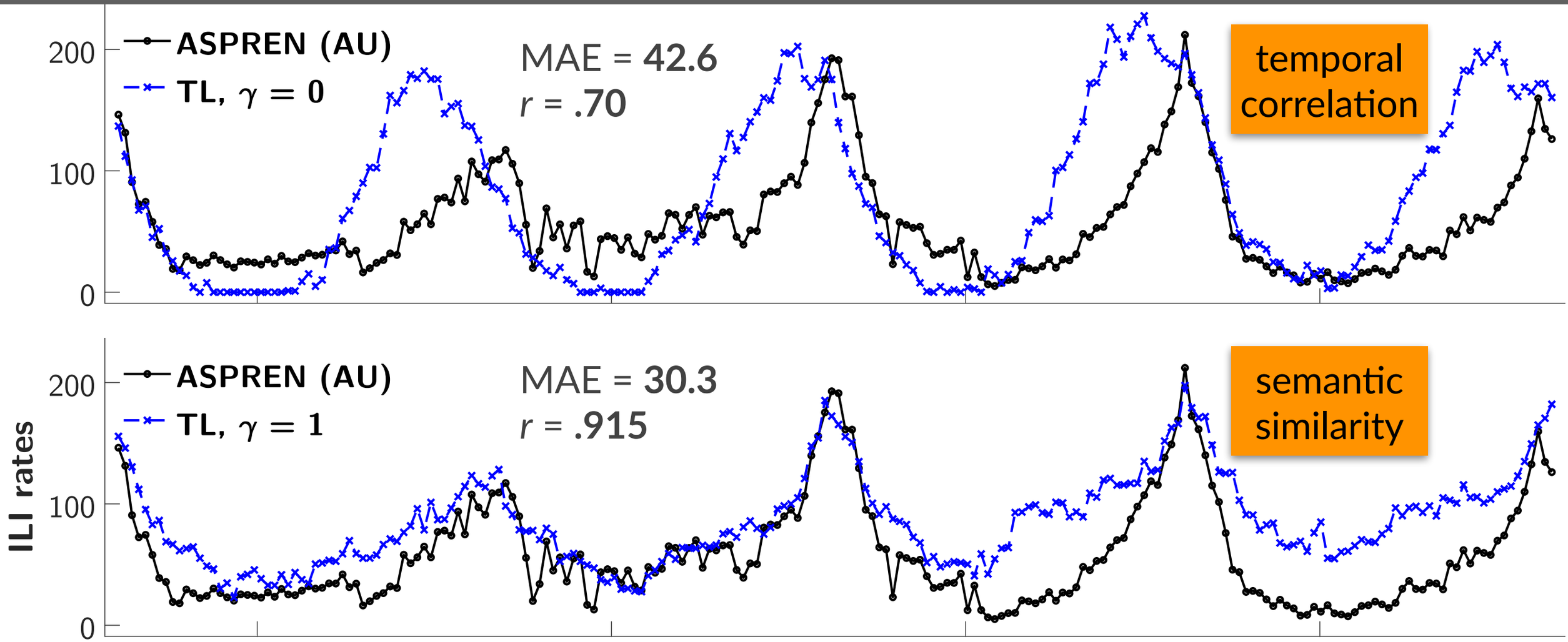
How similar are the flu rates in the US and **Australia**?

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



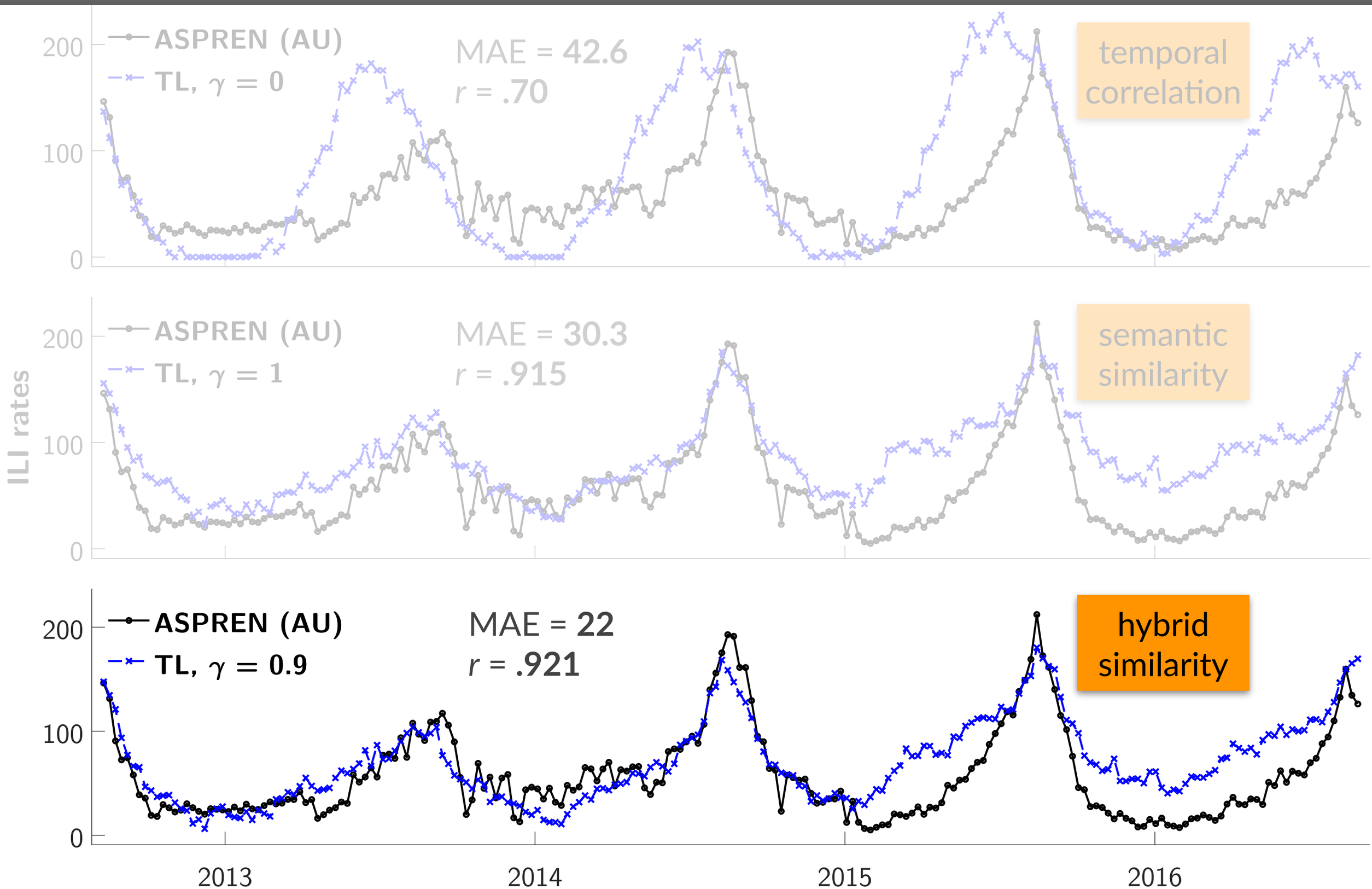
Zou, Lampos, Cox (2019), *WWW '19*

Transferring a flu model from US to Australia



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

Transferring a flu model from US to Australia



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

Part C

Tracking COVID-19 using online search

@lampos

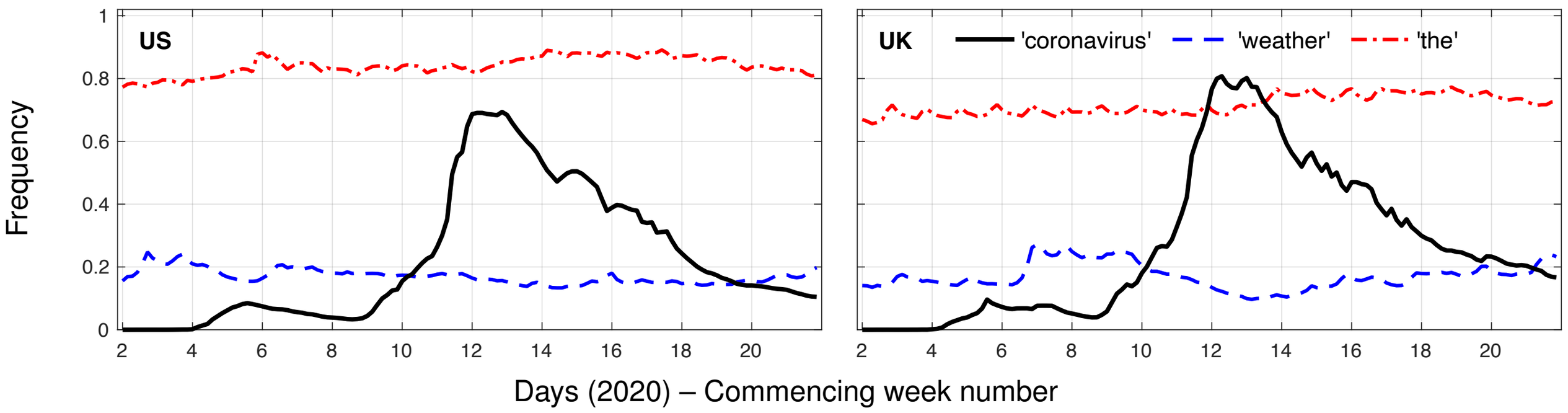
www lampos.net

Google search activity

Google Health Trends: daily frequency of web searches for a location

$$\text{frequency} = \frac{\text{\# times a search term was issued}}{\text{total \# of searches}}$$

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



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

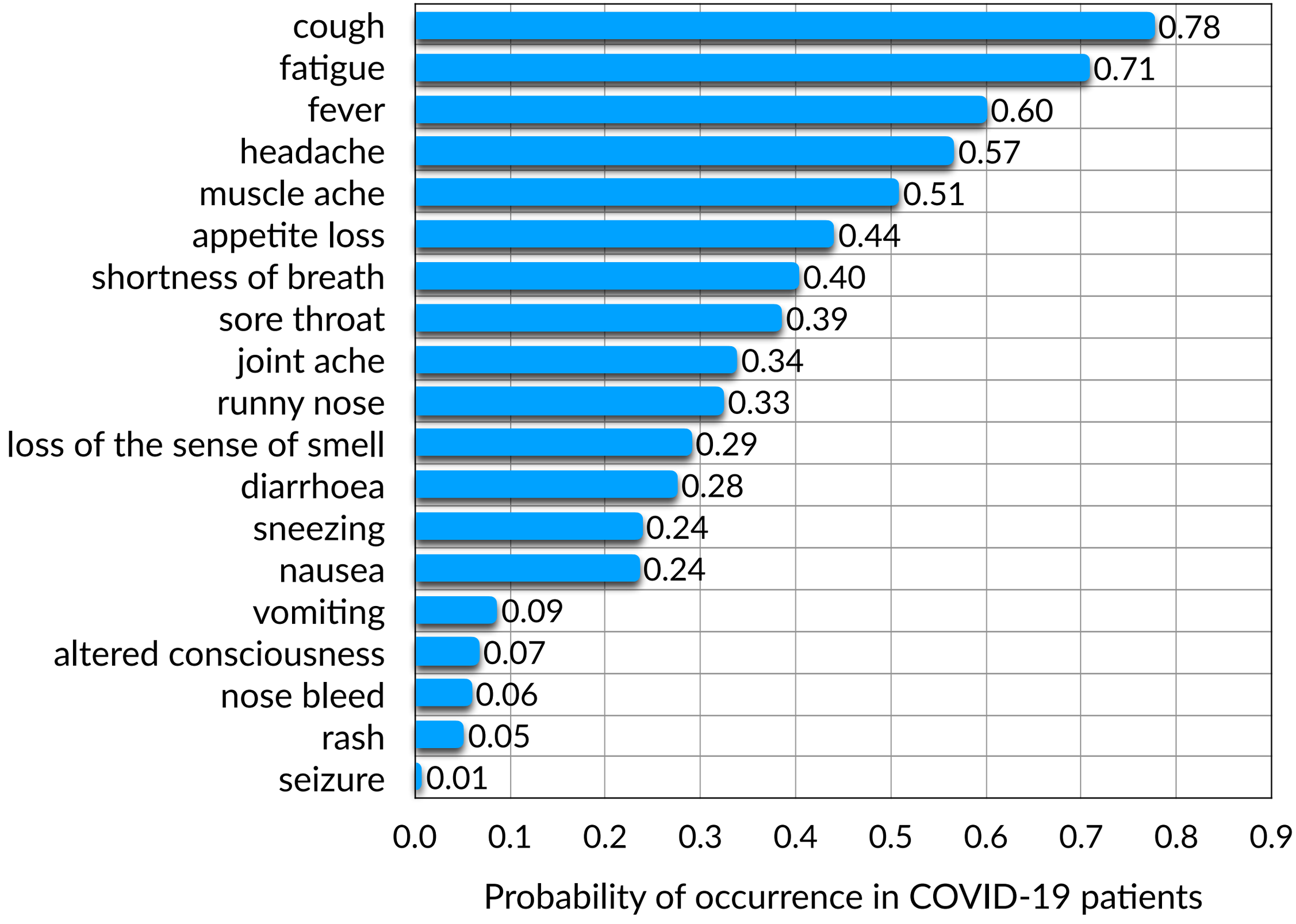
Challenges in modelling COVID-19 using web search

- No reliable and not enough **ground truth** data
  - Supervised learning no longer possible — transfer learning (?)
  - Evaluation of any model is 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 (concern, curiosity, media coverage) rather than by infection?

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

First Few 100 (FF100) patients survey (NHS/PHE)



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

Symptom-related search terms — *English*

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

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

Symptom-related search terms — *Italian*

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

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

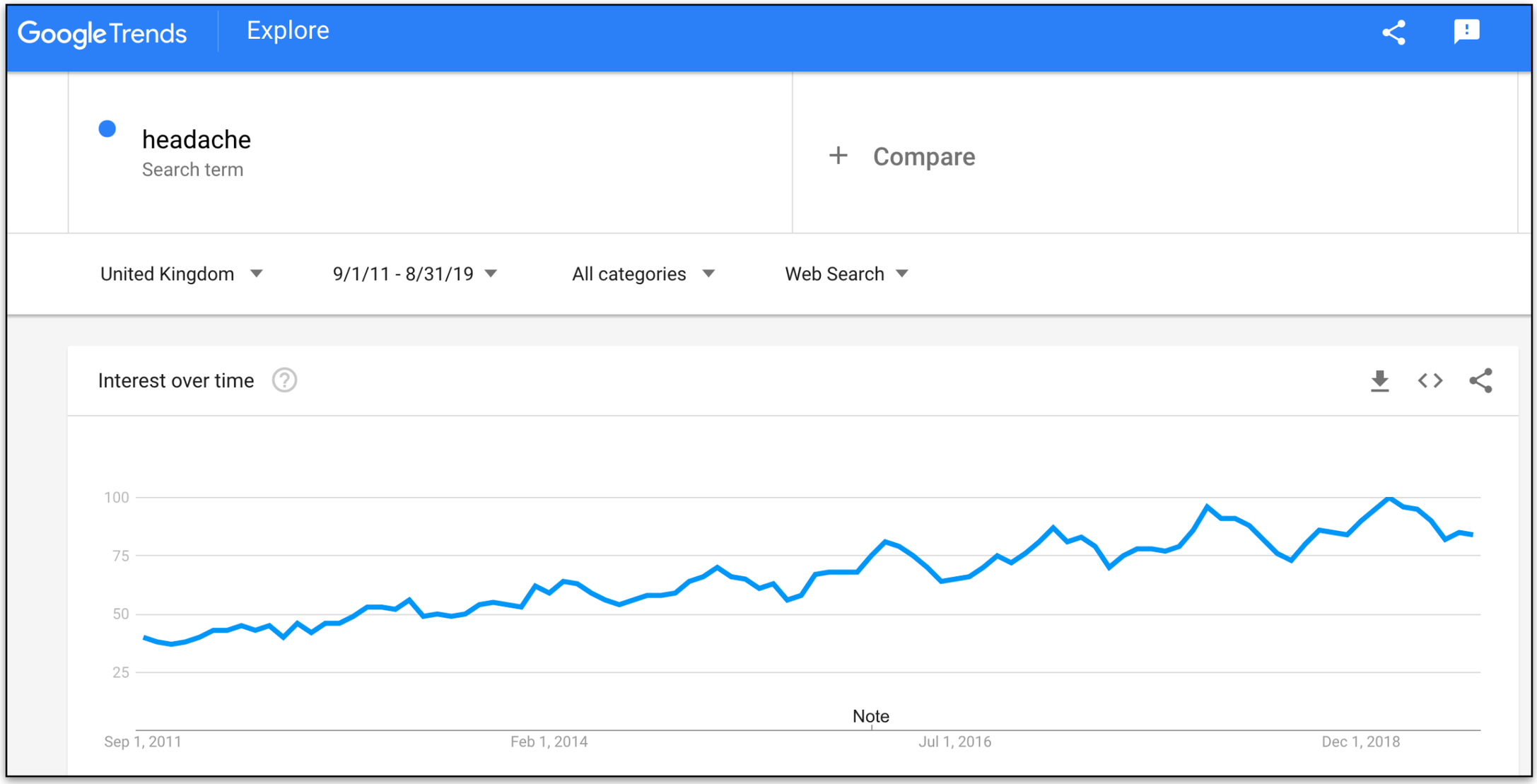
Symptom-related search terms — *languages & countries*

- English (*US*, *UK*, *Australia*, *Canada*)
- French (*France*)
- Italian (*Italy*)
- Zulu, Afrikaans, English, and many more (*South Africa*)
- Greek (*Greece*)

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

A very 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



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

A very 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

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

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

– $g_p = \gamma g$

– $g_c = (1 - \gamma)g$

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

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

Linear AR model to forecast g at a time point t based on its past values

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

Linear AR model to forecast g at a time point t based on its past values **and** the current and past values of m

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

$$\rightarrow \text{ prediction error } \epsilon_2$$

- $\epsilon_1 < \epsilon_2$: the media signal does not help $\rightarrow \gamma \approx 1$
- $\epsilon_1 \geq \epsilon_2 : \gamma = \epsilon_2/\epsilon_1$ (*crude estimation*)

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

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

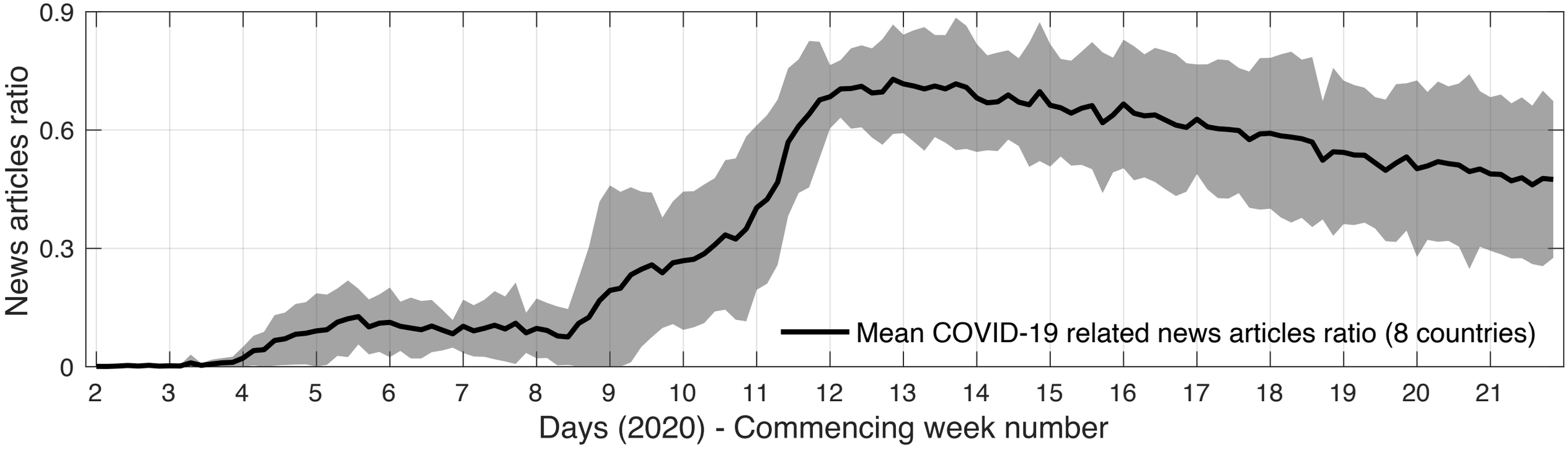
- Obtain the **daily ratio** of articles that include basic COVID-19-related keywords in their title or main text
e.g. “covid” or “coronavirus”

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

News media coverage corpus

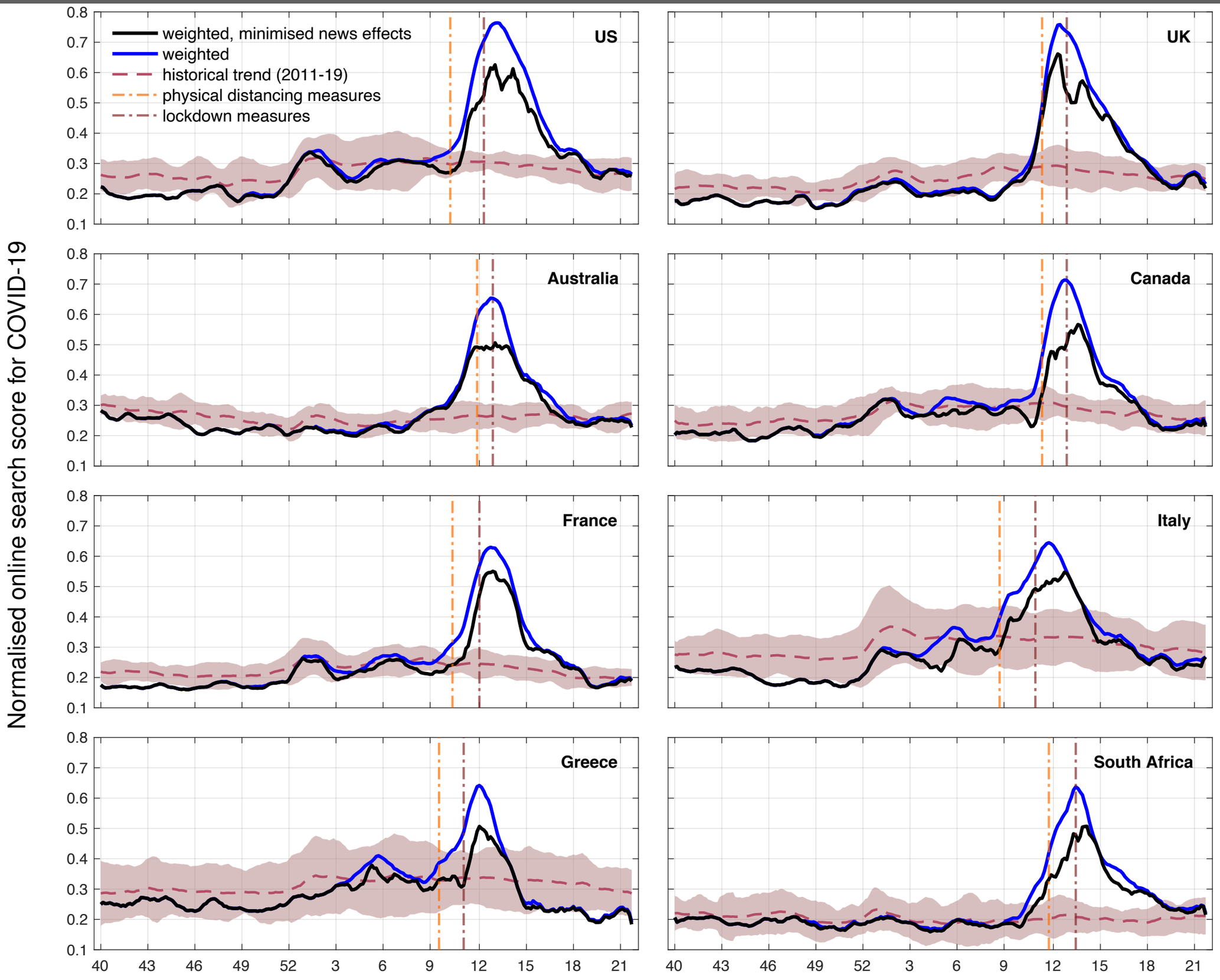
- Data obtained from September 30, 2019 to May 24, 2020
- > 0 frequency from ~January, 2020 onwards
- **~2.5 million** COVID-19-related articles from a total of ~10 million

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



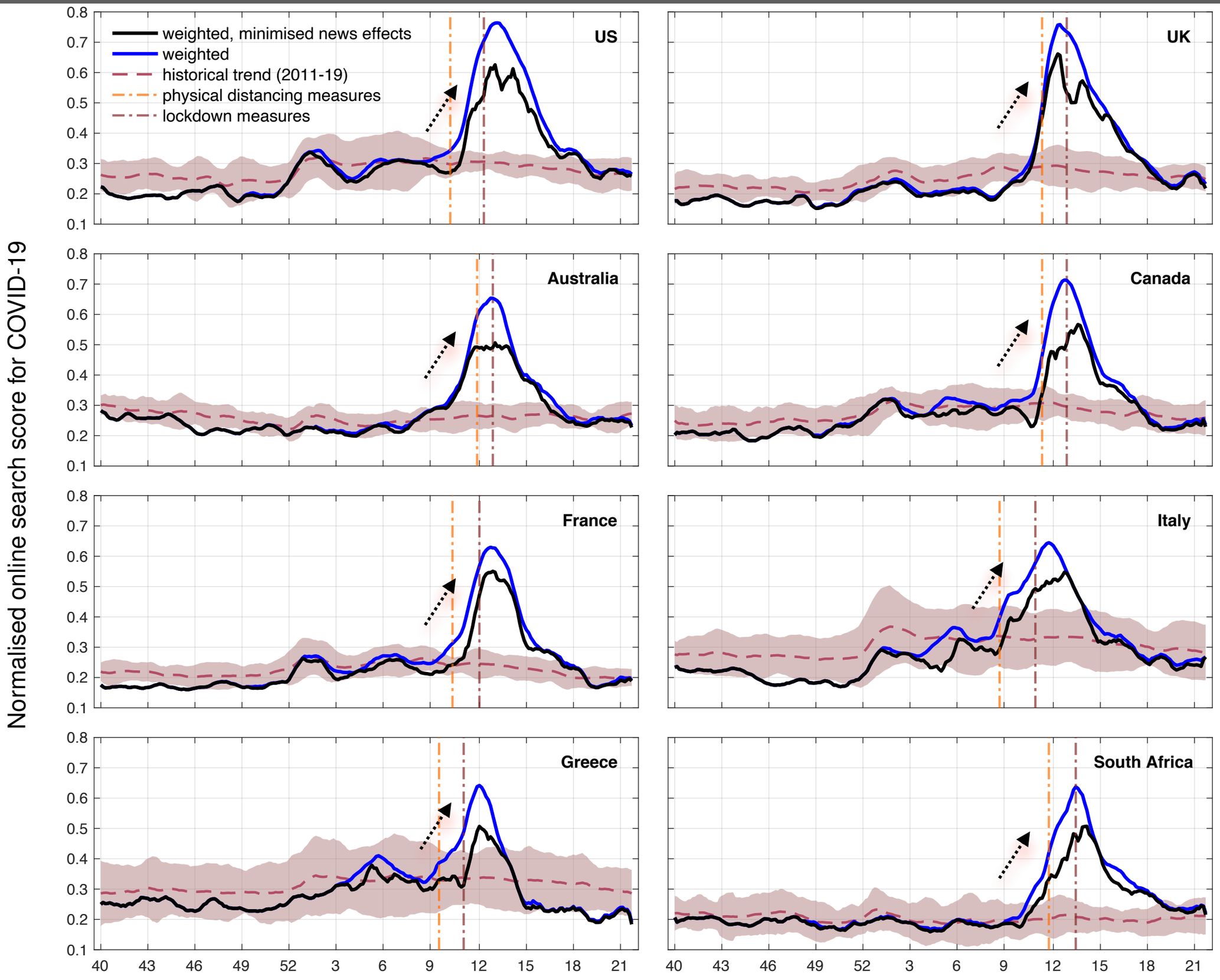
Lampos *et al*. (2021), *npj Digit. Med.*

Unsupervised COVID-19 models in 8 countries



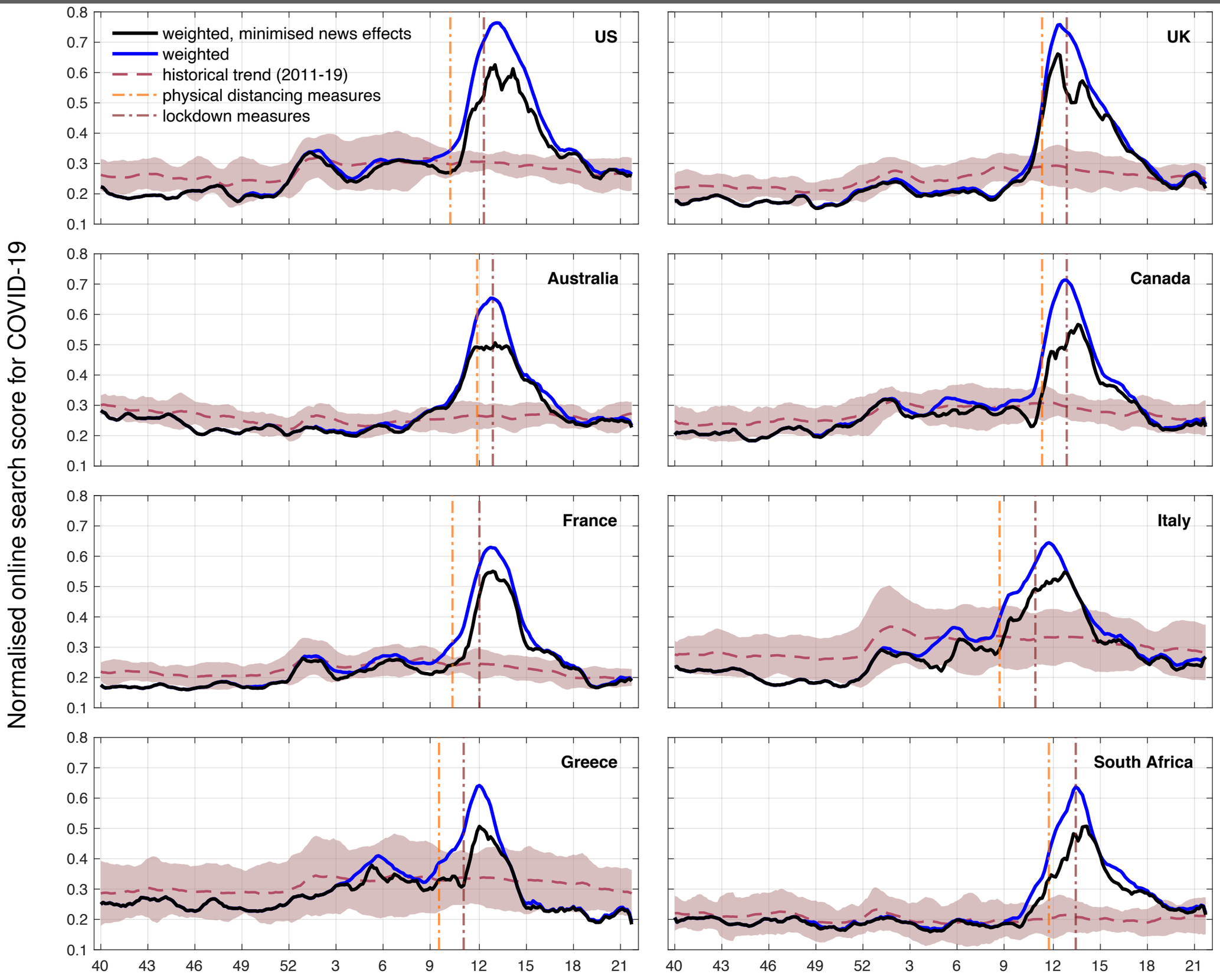
Lampos *et al.* (2021), *npj Digit. Med.*

Unsupervised COVID-19 models in 8 countries



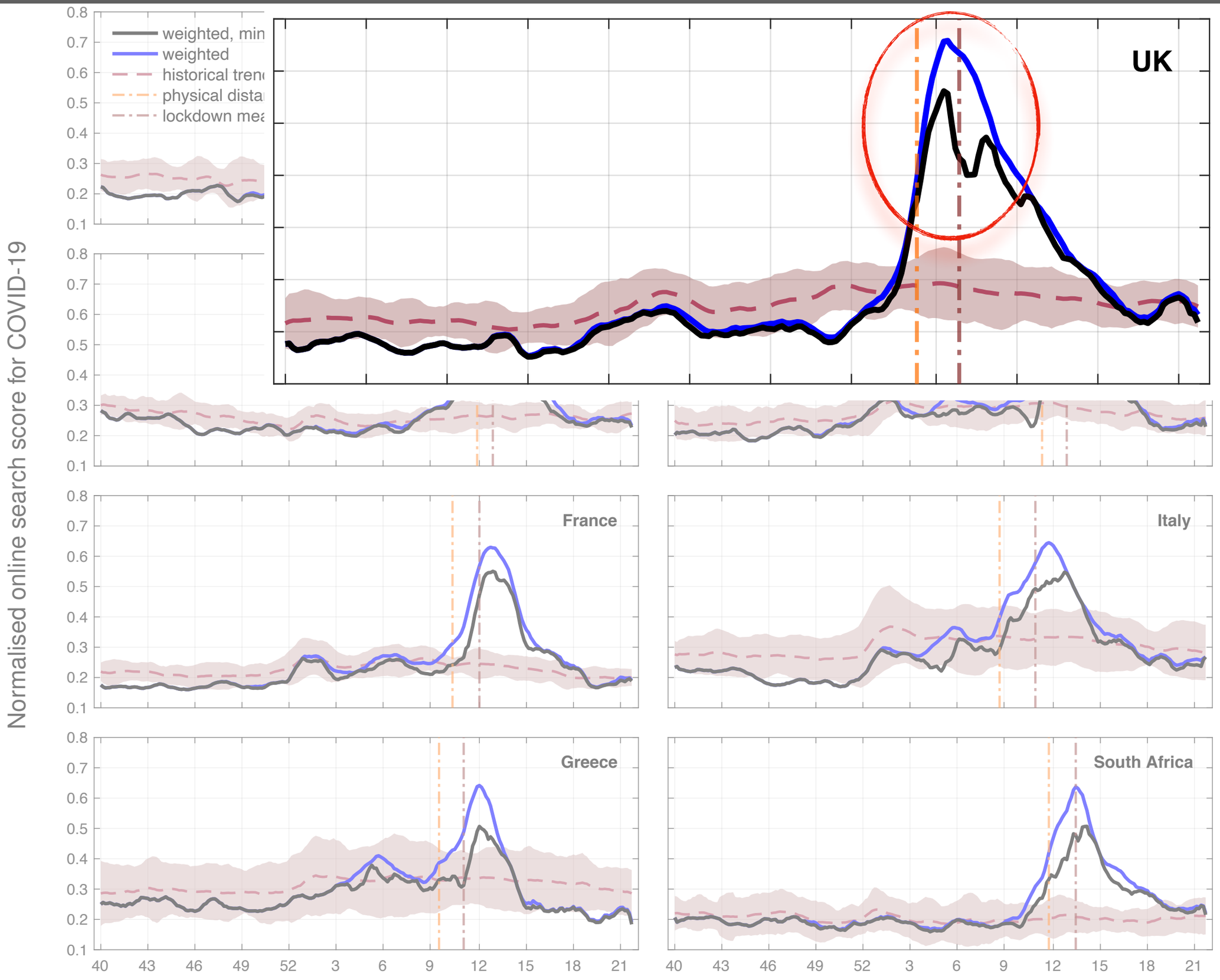
Lampos *et al.* (2021), *npj Digit. Med.*

Unsupervised COVID-19 models in 8 countries



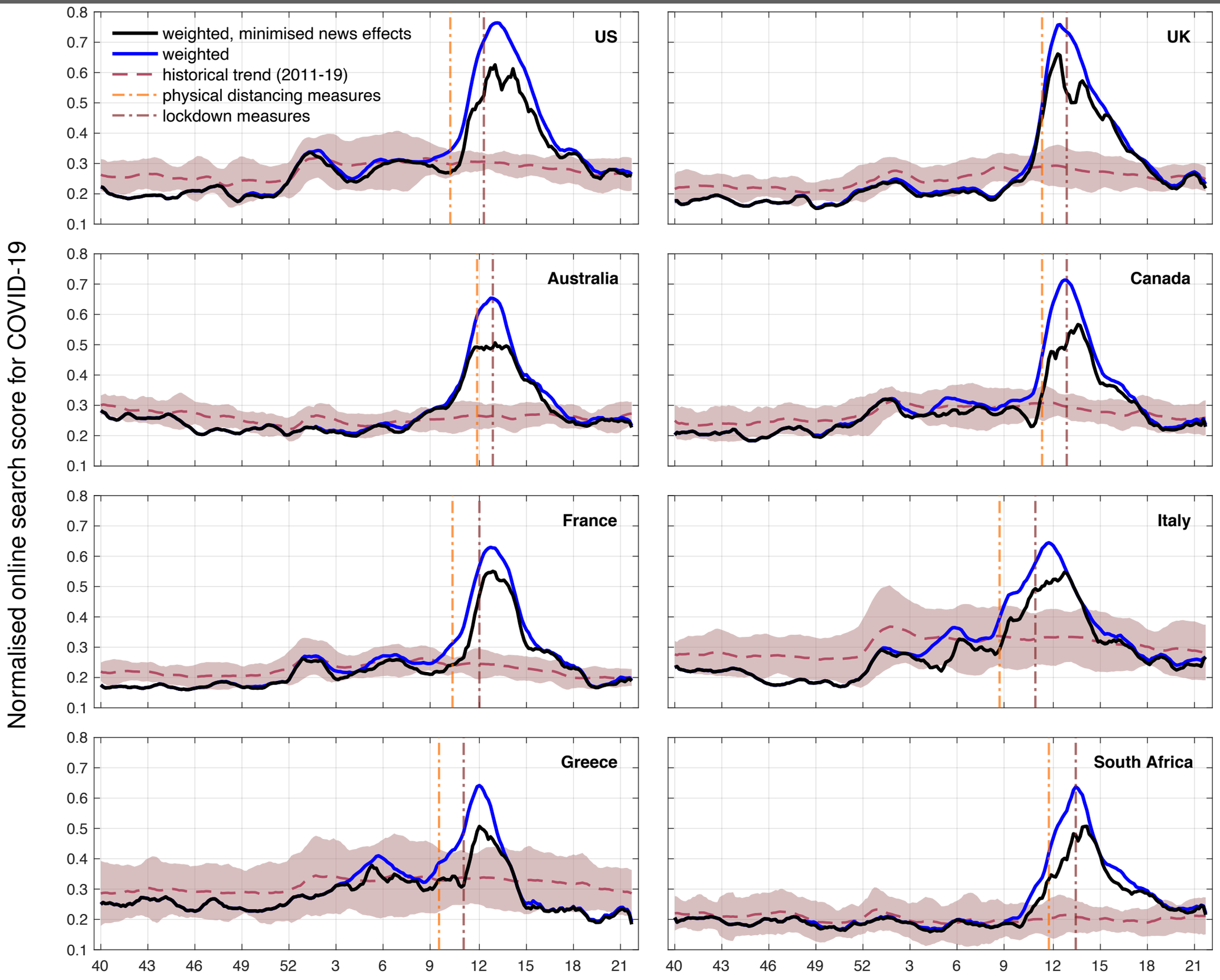
Lampos *et al.* (2021), *npj Digit. Med.*

Unsupervised COVID-19 models in 8 countries



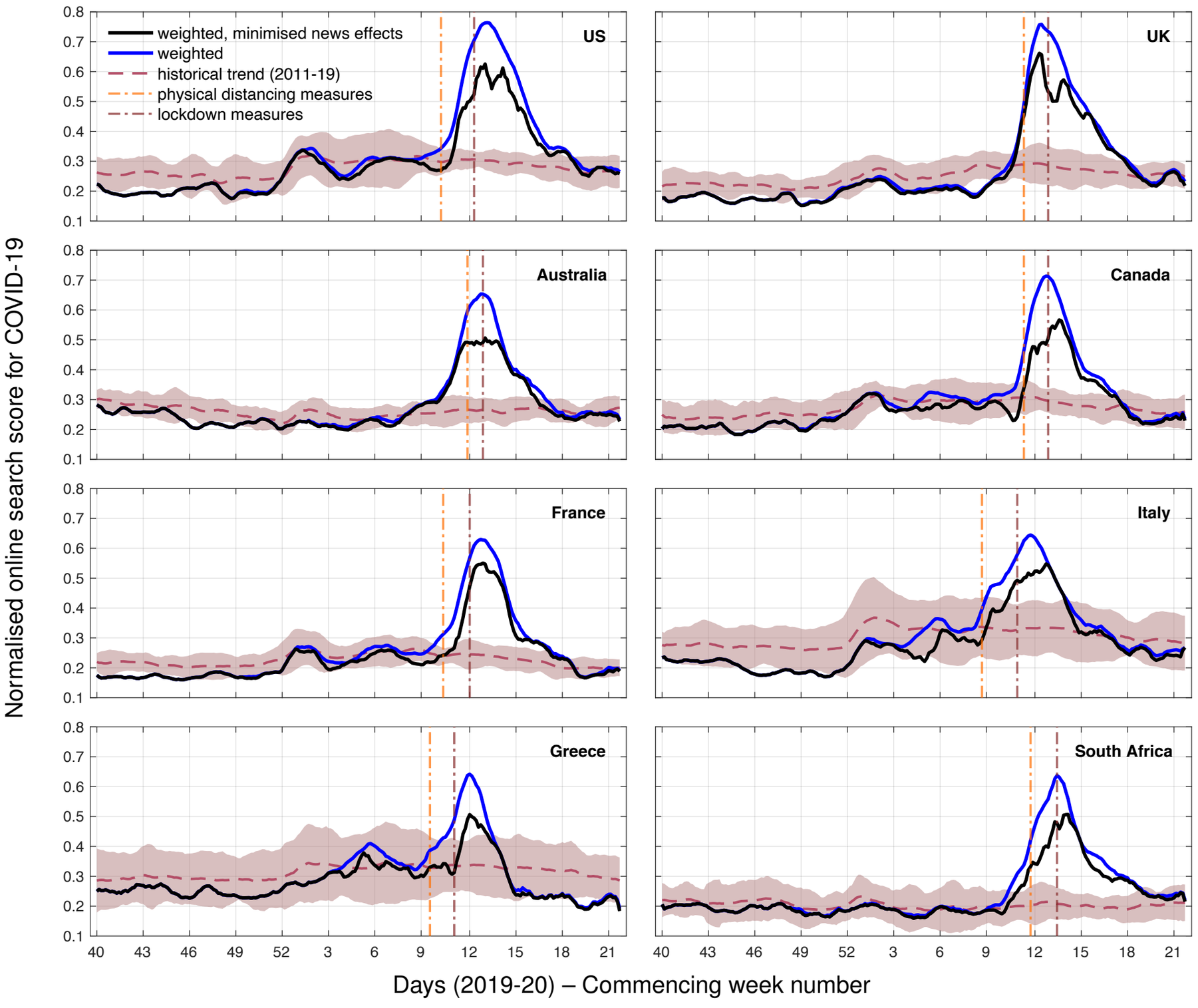
Lampos *et al*. (2021), *npj Digit. Med.*

Unsupervised COVID-19 models in 8 countries



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

Unsupervised COVID-19 models in 8 countries



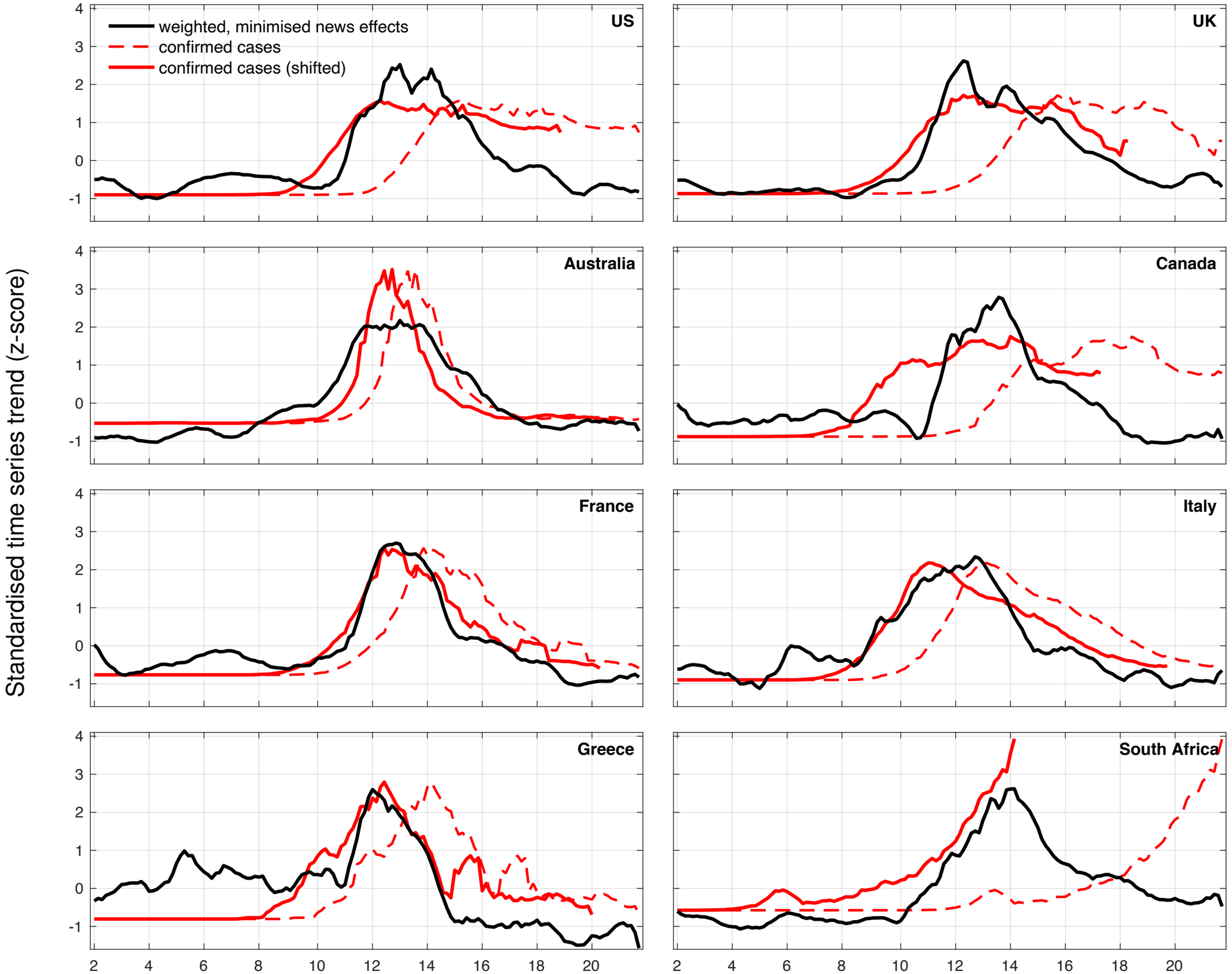
Altered trend during peak periods

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

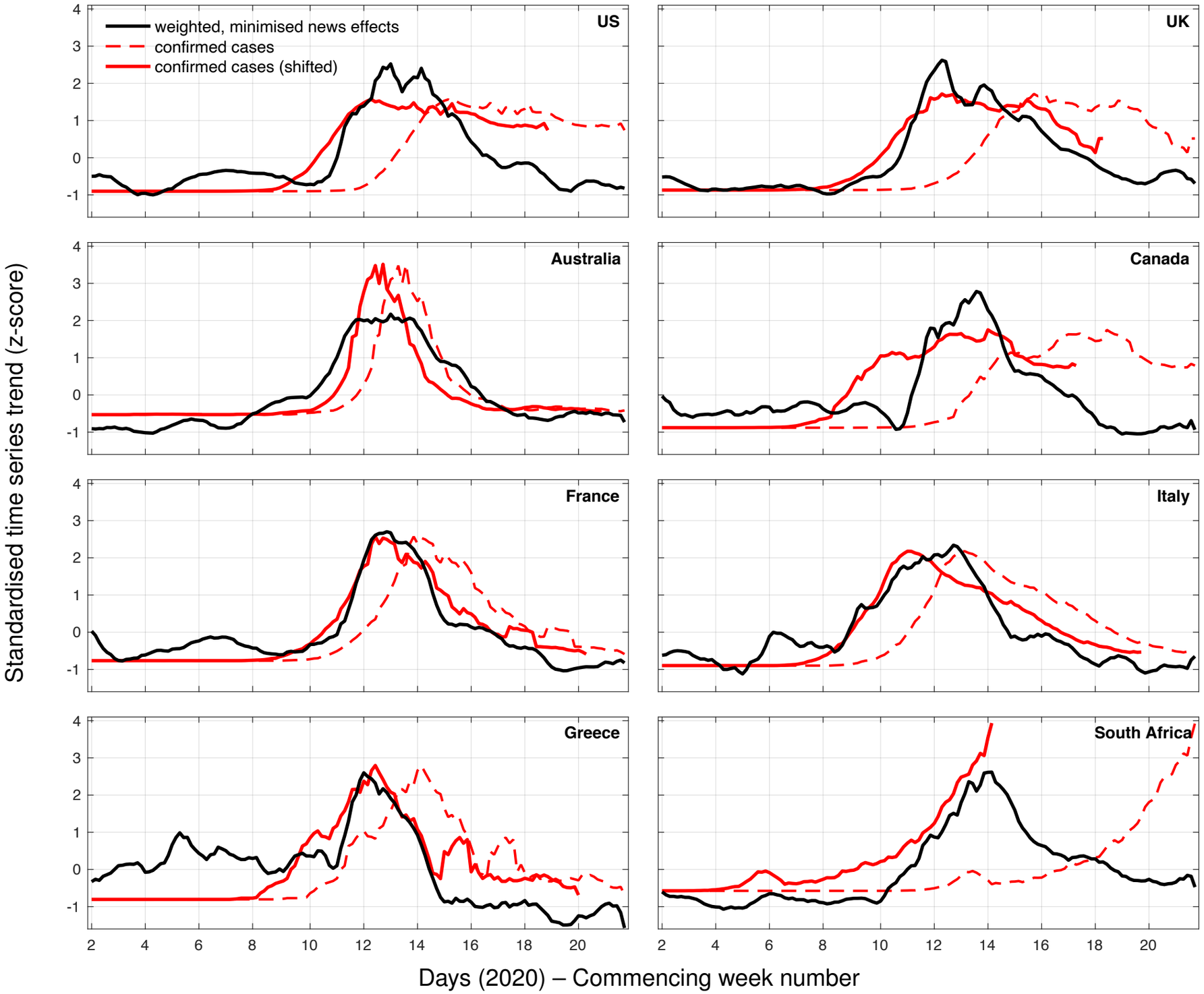
Lampos *et al*. (2021), *npj Digit. Med.*

Comparison with confirmed COVID-19 *cases*



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

Comparison with confirmed COVID-19 *cases*



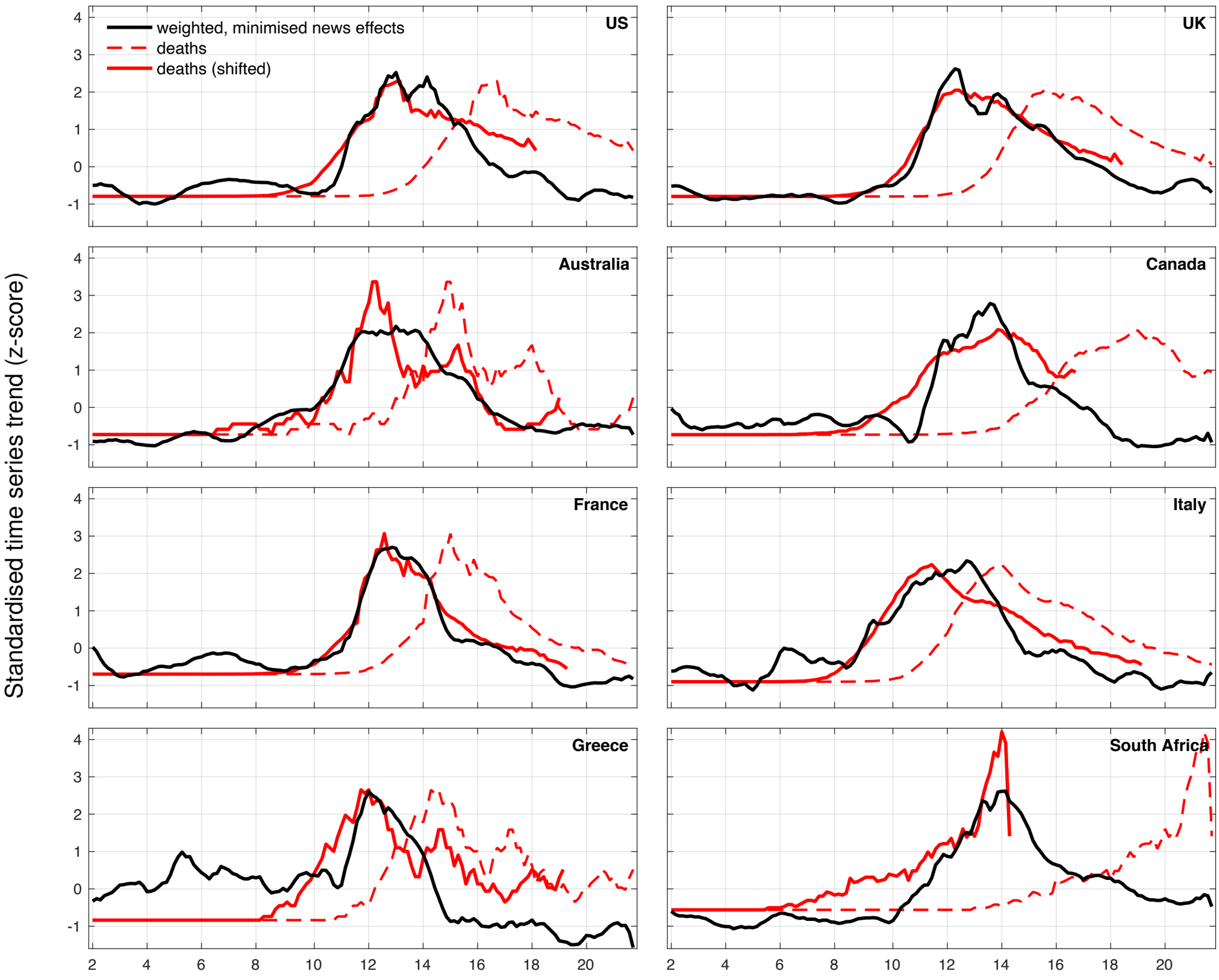
Early warning
— r_{max} = .83 (.74—.92)

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

(*South Africa is excluded*)

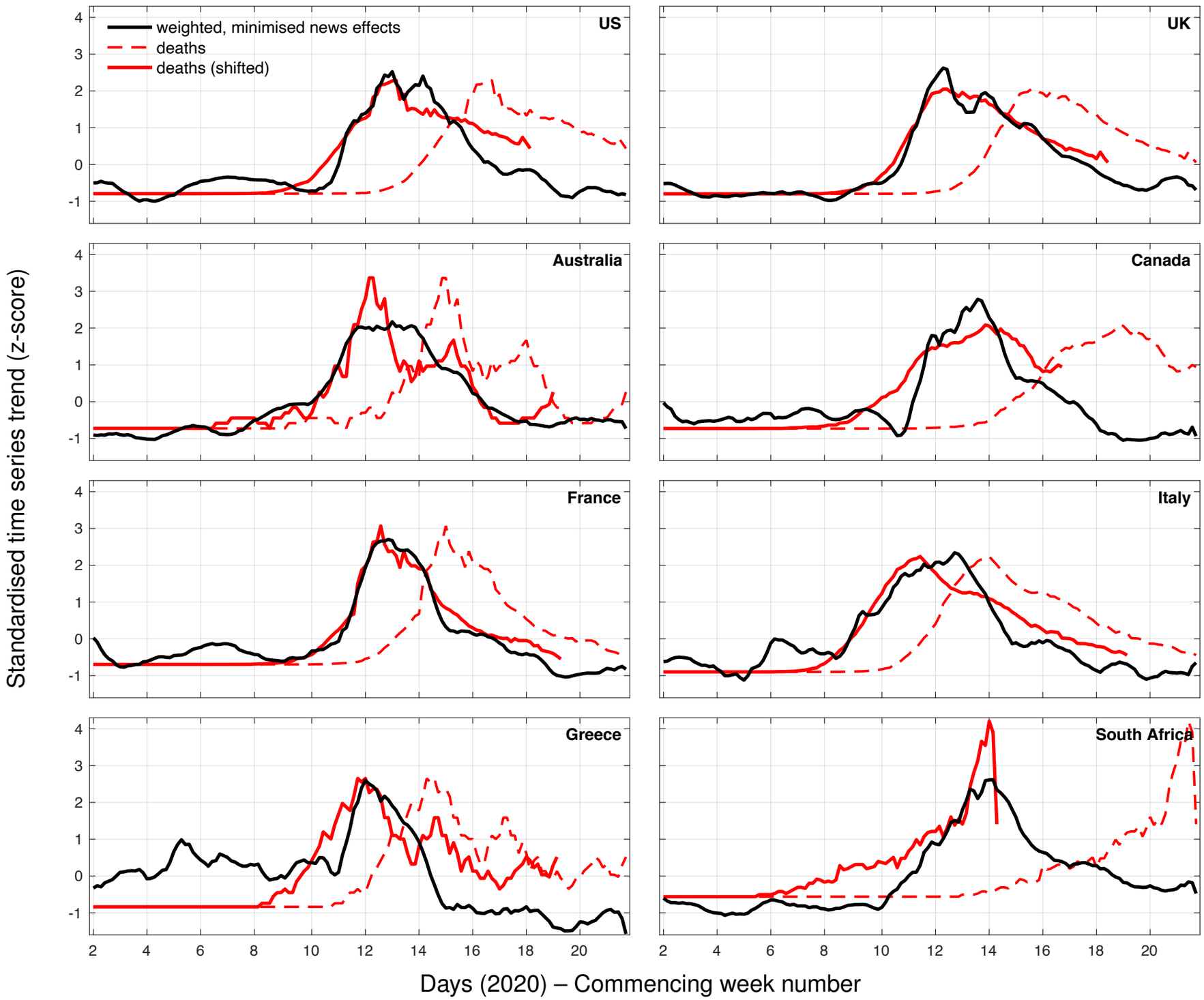
Lampos *et al*. (2021), *npj Digit. Med.*

Comparison with confirmed COVID-19 *deaths*



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

Comparison with confirmed COVID-19 *deaths*



Early warning
— r_{max} = .85 (.70—.99)

when cases are brought forward by
22.1 (17.4—26.9) **days**

(*South Africa is excluded*)

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

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

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

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

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

$\mathbf{S} \in \mathbb{R}^{M \times N}$: M daily frequencies of N search terms
$\mathbf{w} \in \mathbb{R}^N, \beta \in \mathbb{R}$: regression weights and intercept
$\ell_1, \ell_2 \in \mathbb{R}_{\geq 0}$: 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 to quantify **model uncertainty**

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

Transfer learning for COVID-19 incidence models

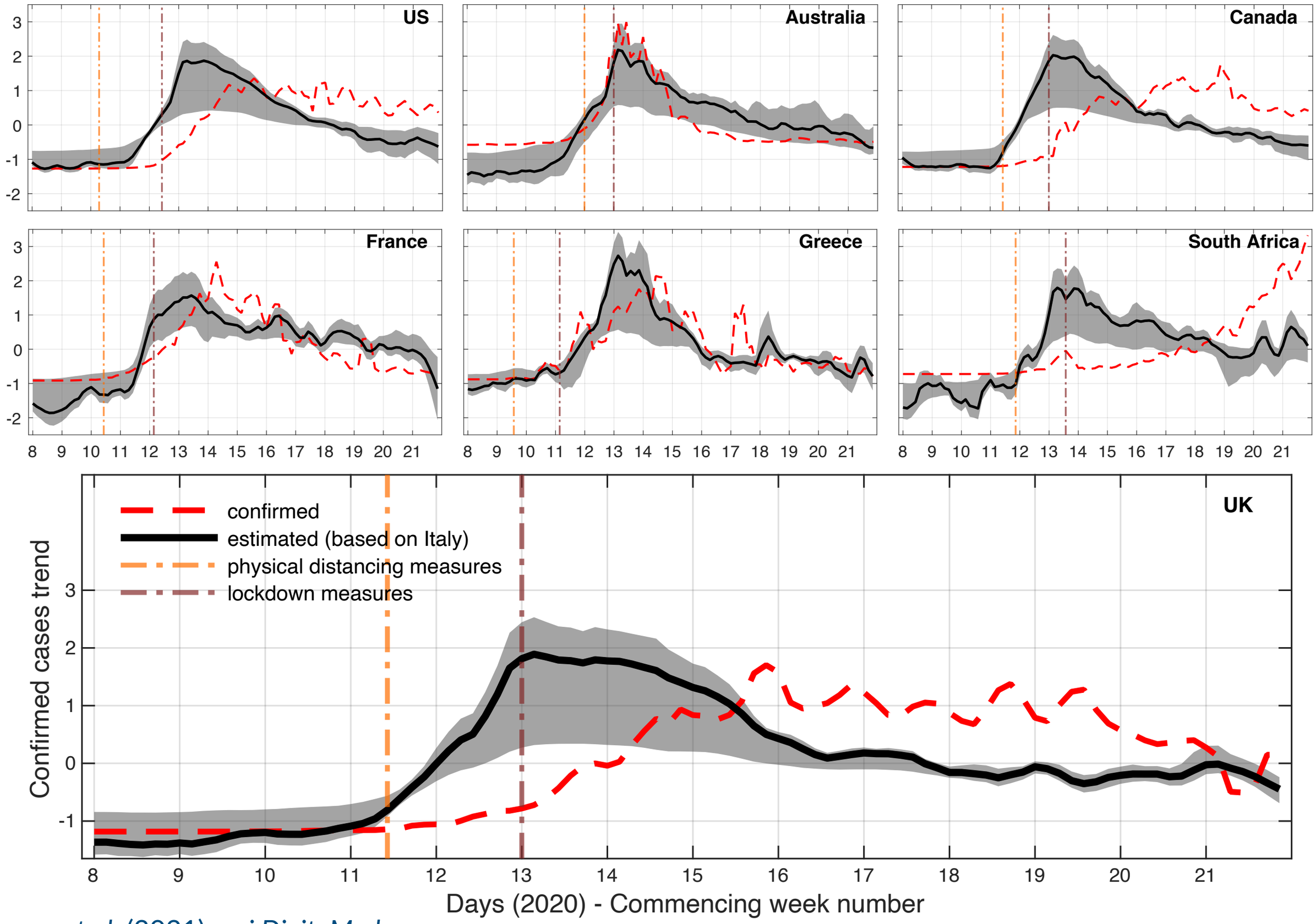
- Establish search query pairs between the source and the target countries
  - lookup for query pairs **within the same symptom category**
  - pair a source query to the target query with the greatest **bivariate correlation**, after identifying an optimal shifting period

- **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 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

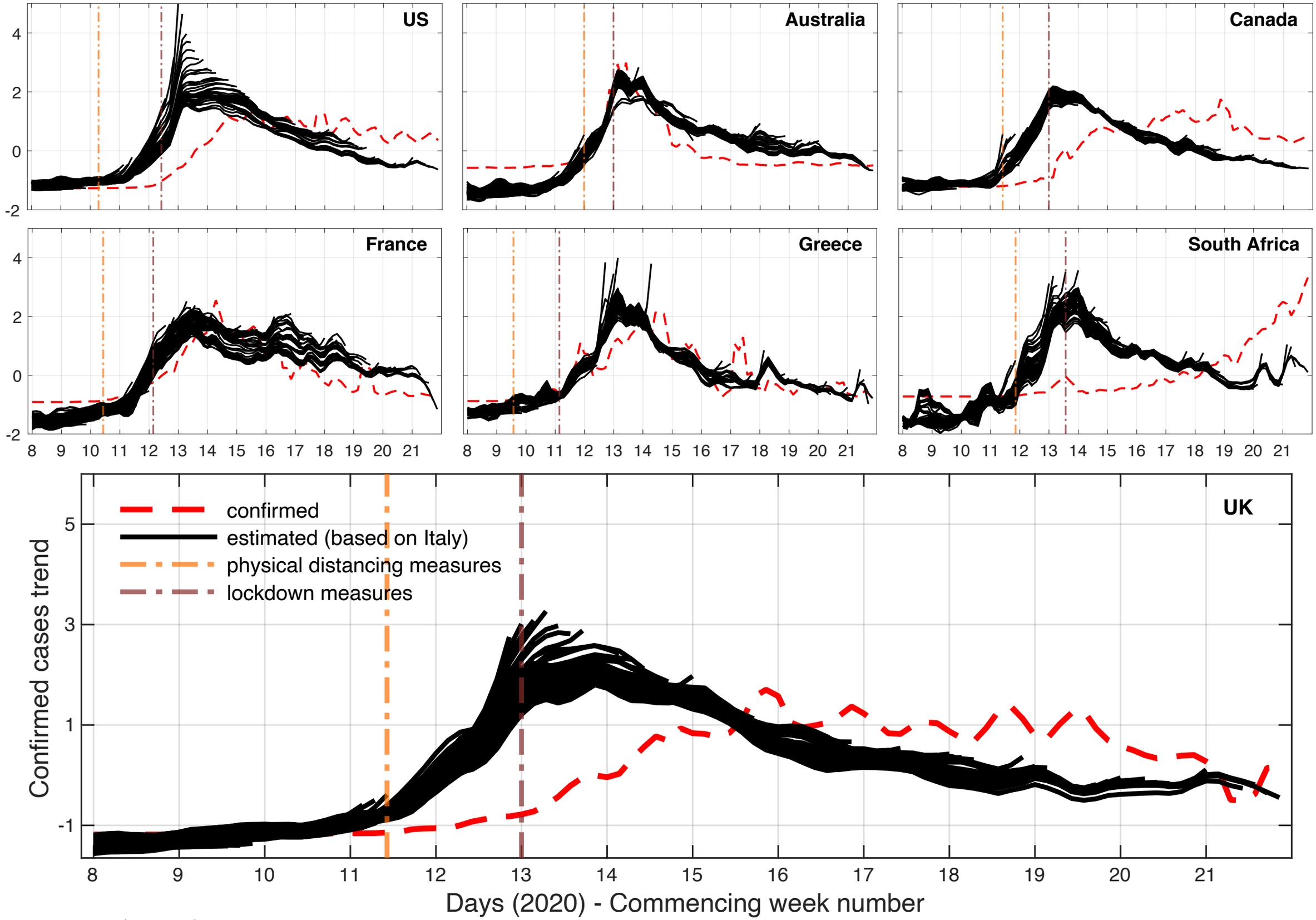
Lampos *et al*. (2021), *npj Digit. Med.*

Transfer learning for COVID-19 incidence models



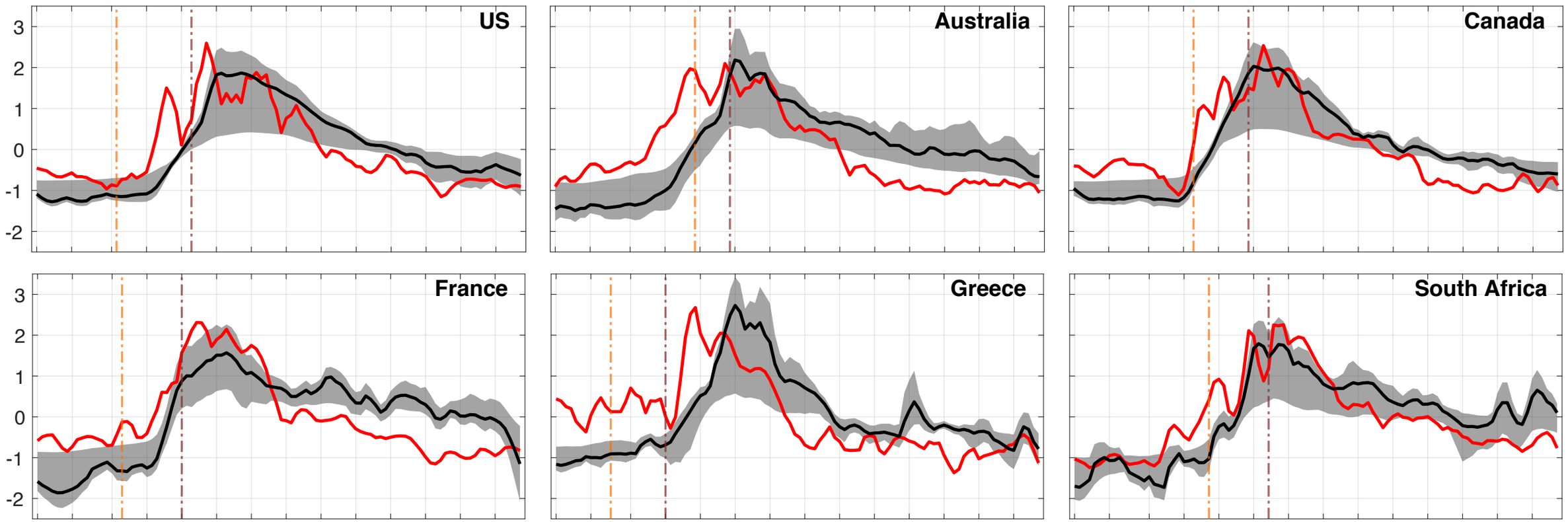
Lampos *et al*. (2021), *npj Digit. Med.*

Transfer learning — *how it actually works*



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

Transfer vs. *unsupervised* learning

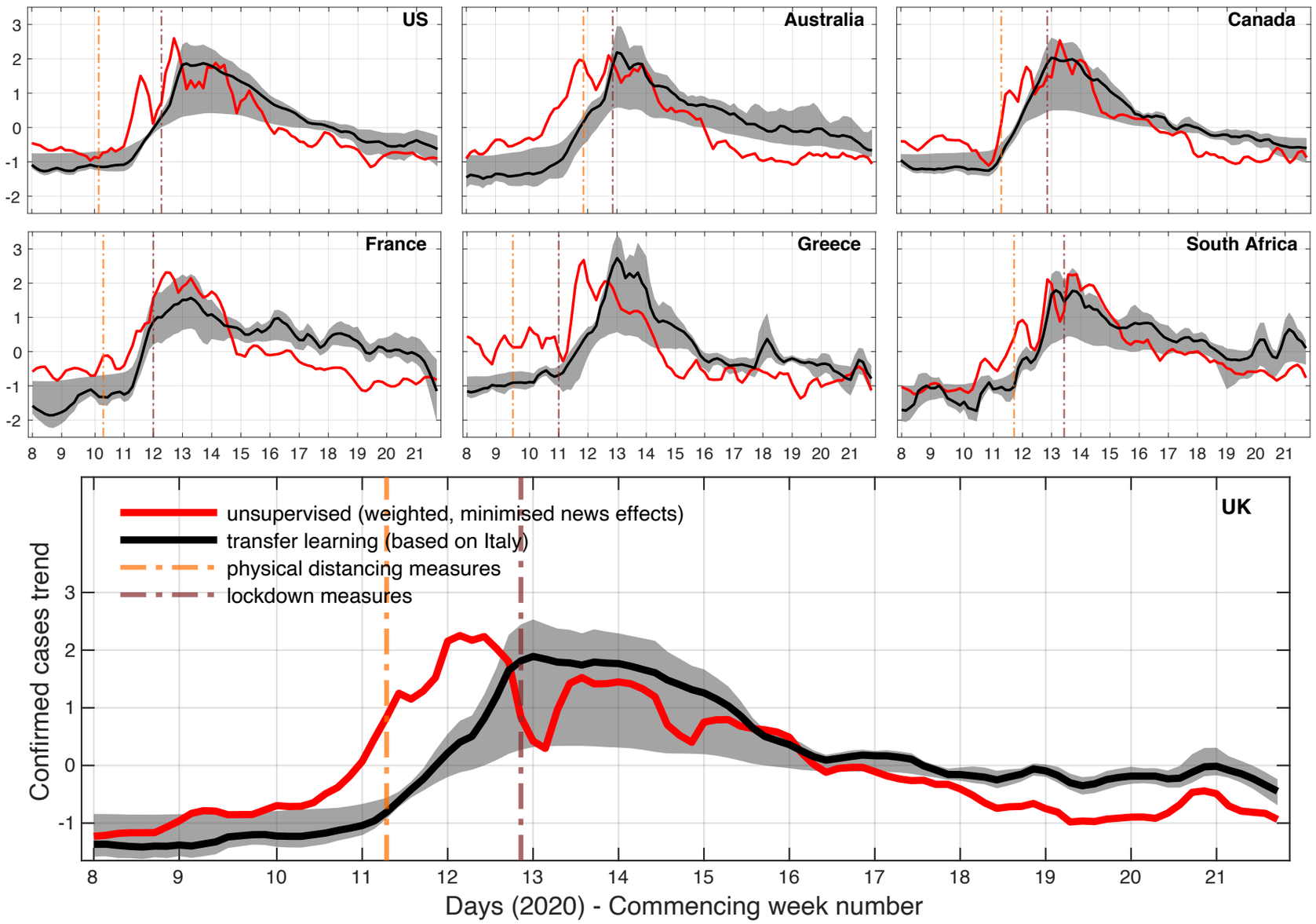


Lampos *et al*. (2021), *npj Digit. Med.*

Transfer vs. *unsupervised* learning

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

Transfer 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**

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

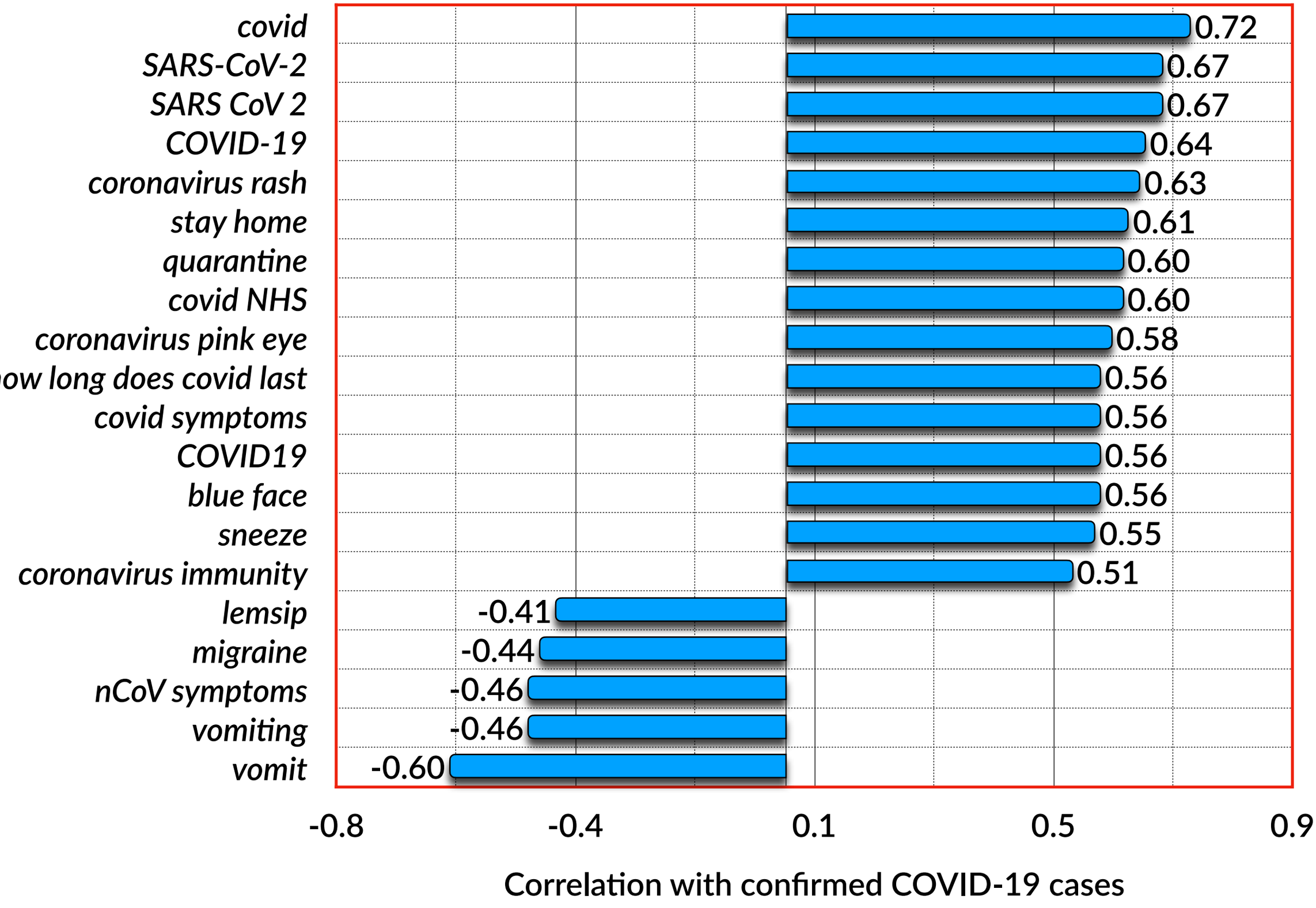
Correlation analysis

- 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

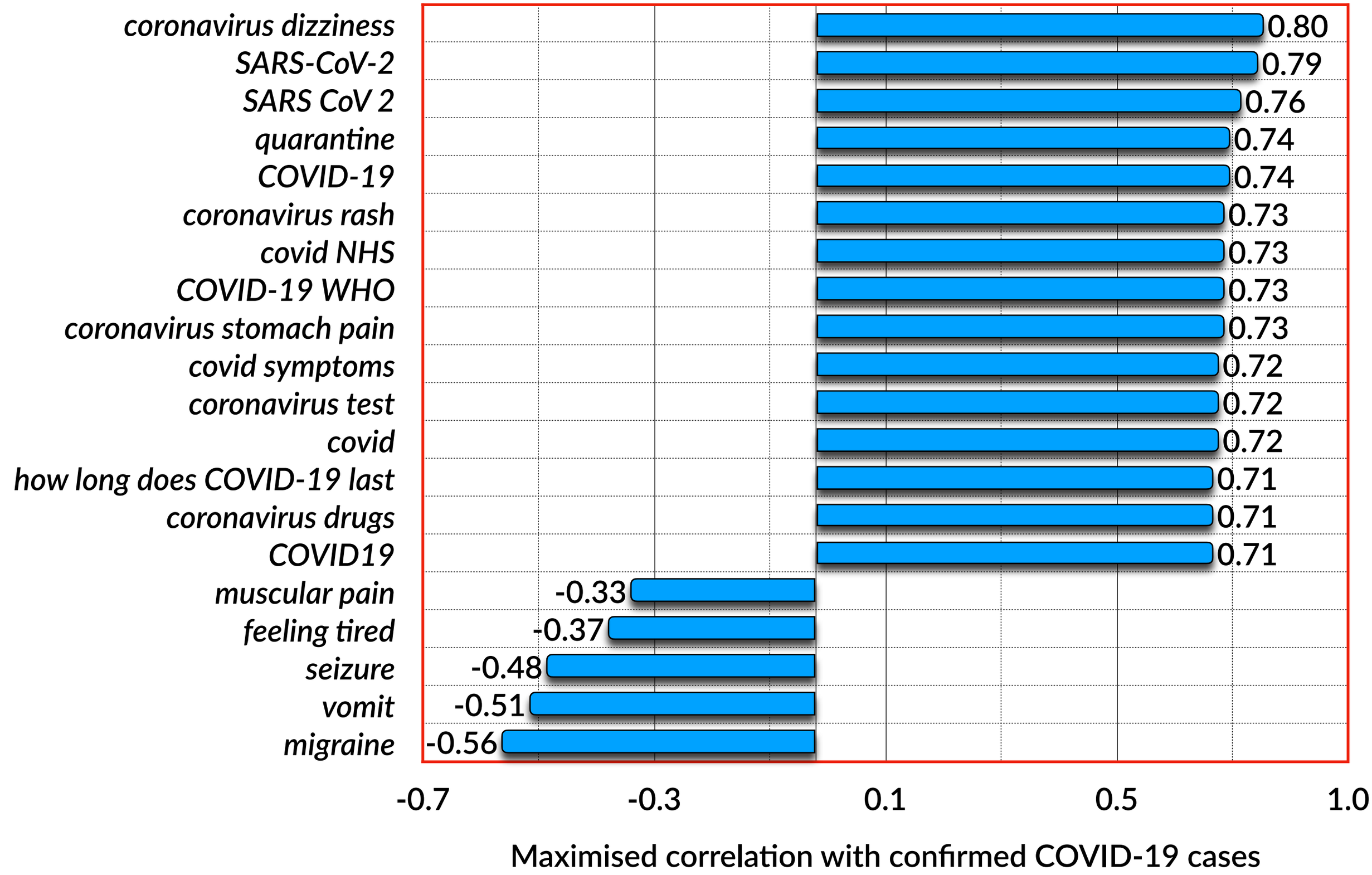
Lampos *et al*. (2021), *npj Digit. Med.*

Correlation with confirmed cases



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

Correlation with confirmed cases (*maximised*)



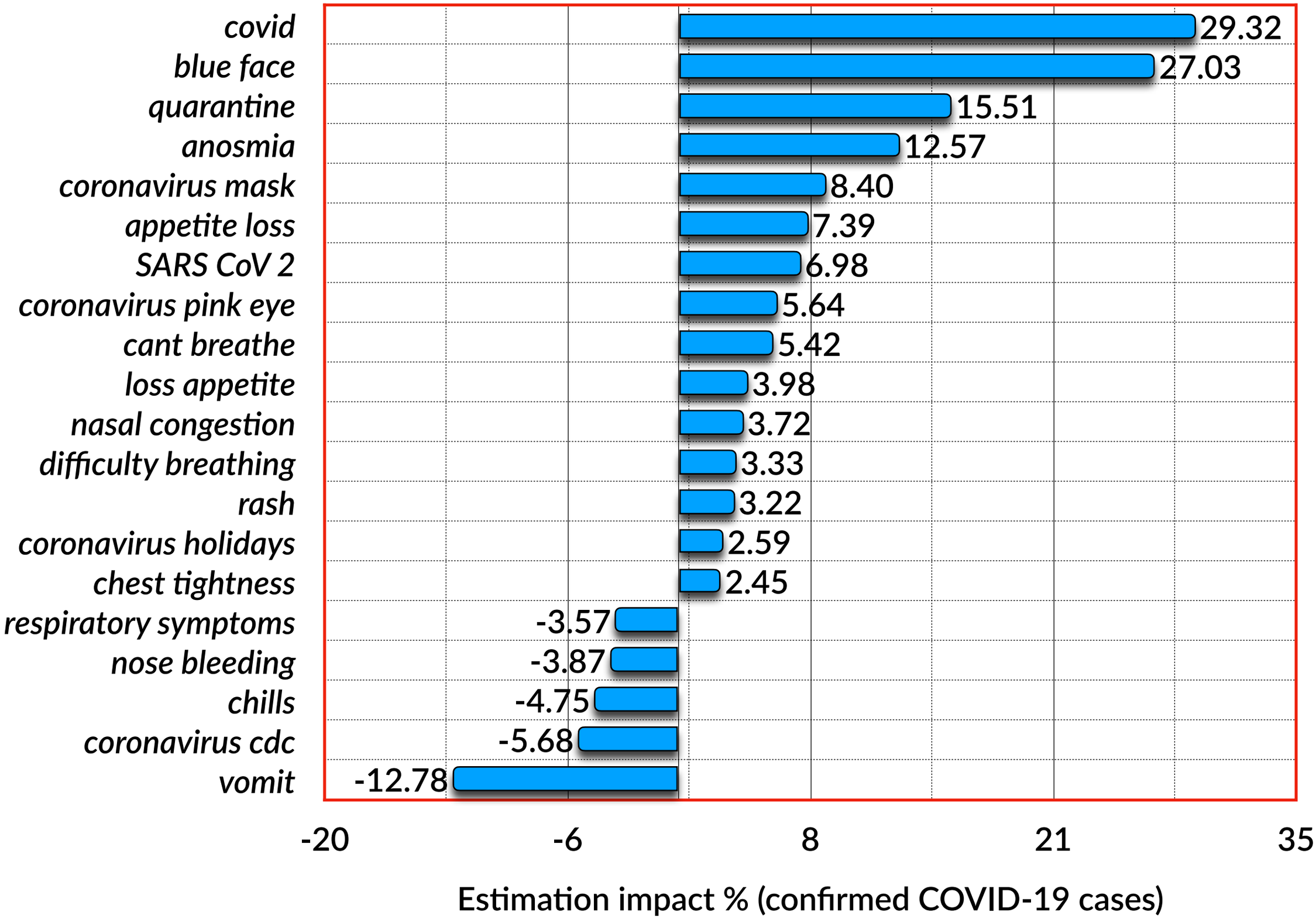
Lampos *et al*. (2021), *npj Digit. Med.*

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

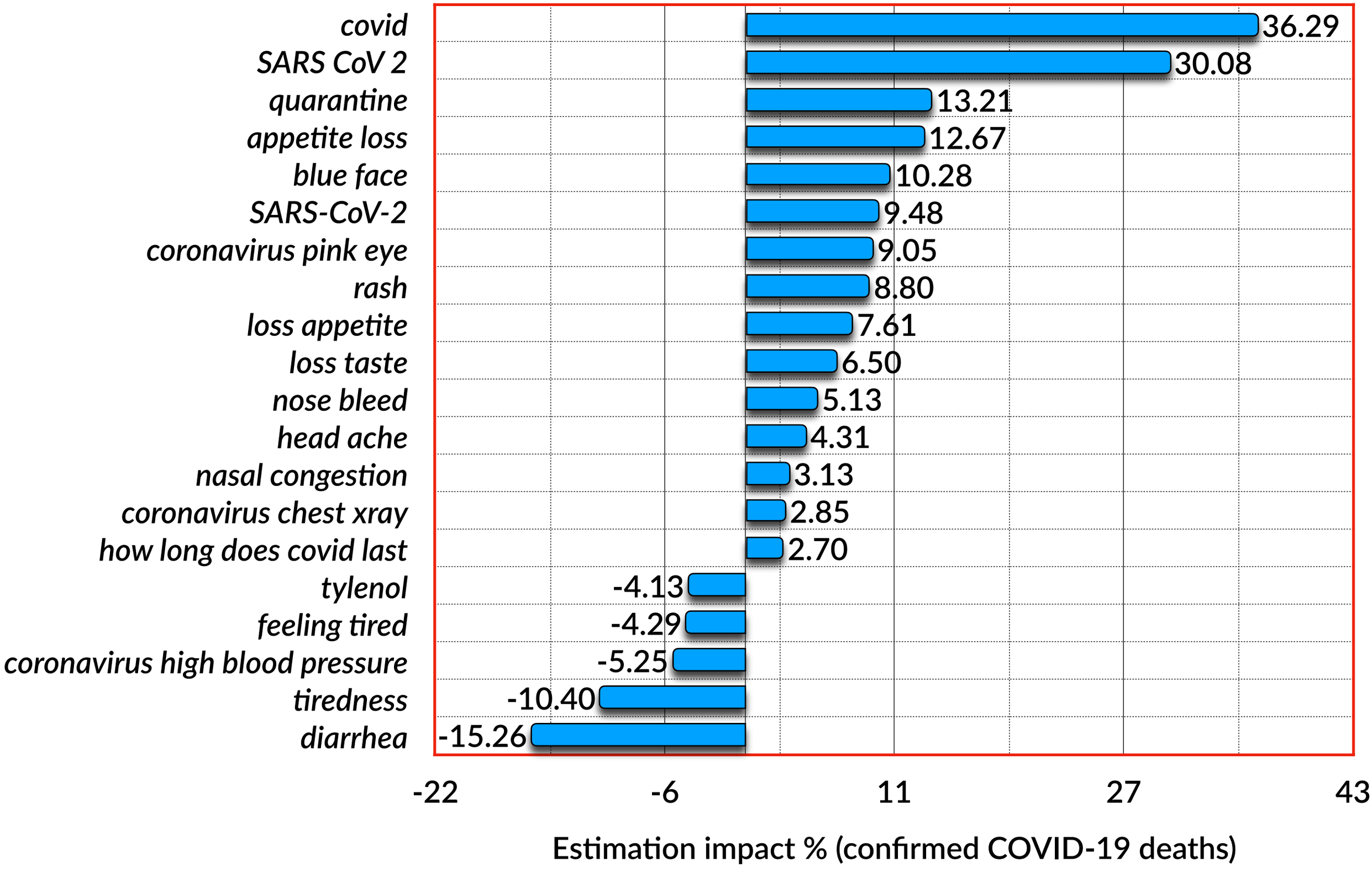
Lampos *et al*. (2021), *npj Digit. Med.*

Regression analysis — *confirmed COVID-19 cases*



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

Regression analysis — *COVID-19 deaths*



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

Outbreak in Italy and web searches elsewhere

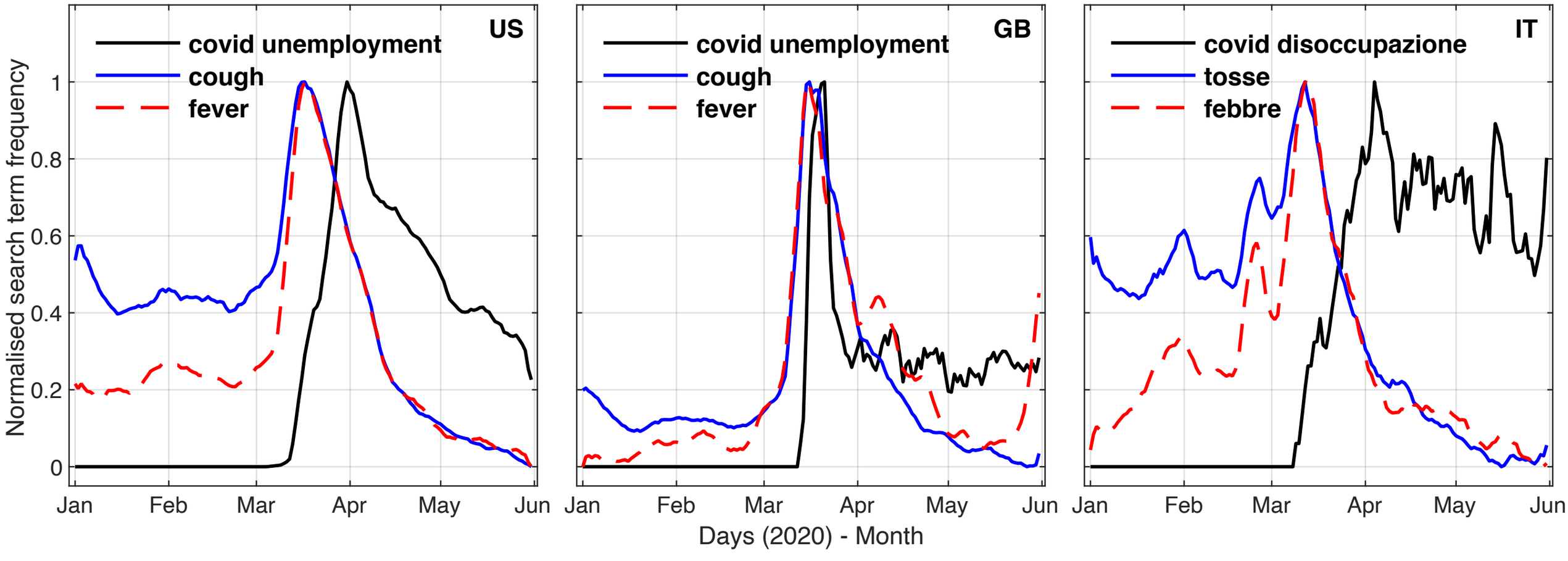
Did the outbreak in Italy **cause** an increase in the frequency of the web searches (*the ones used in our analysis*) elsewhere?

- 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 also attempt 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

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

Symptom-related vs. *general interest* search terms

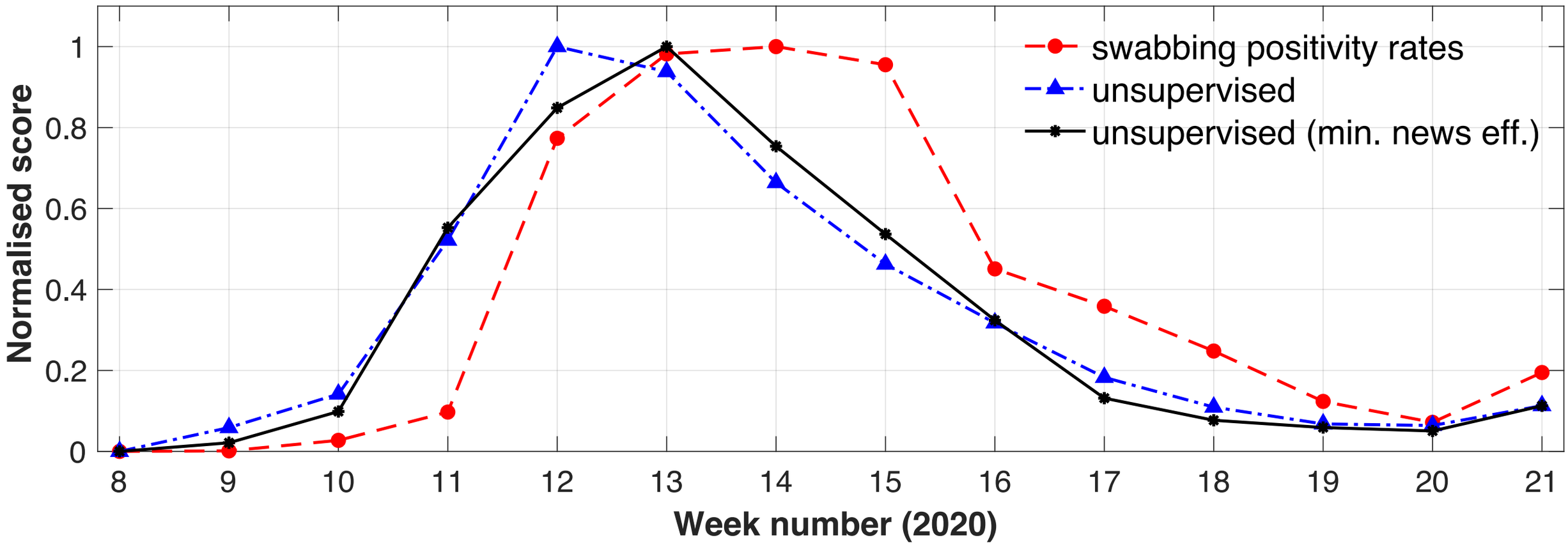
Search terms that are less likely to represent infection (“COVID unemployment”) follow the corresponding trends of search terms about COVID-19-related symptoms (“cough”, “fever”)



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

RCGP swabbing scheme

The **RCGP swabbing scheme** included people with **no** COVID-19-related symptoms → better capturing community-level spread



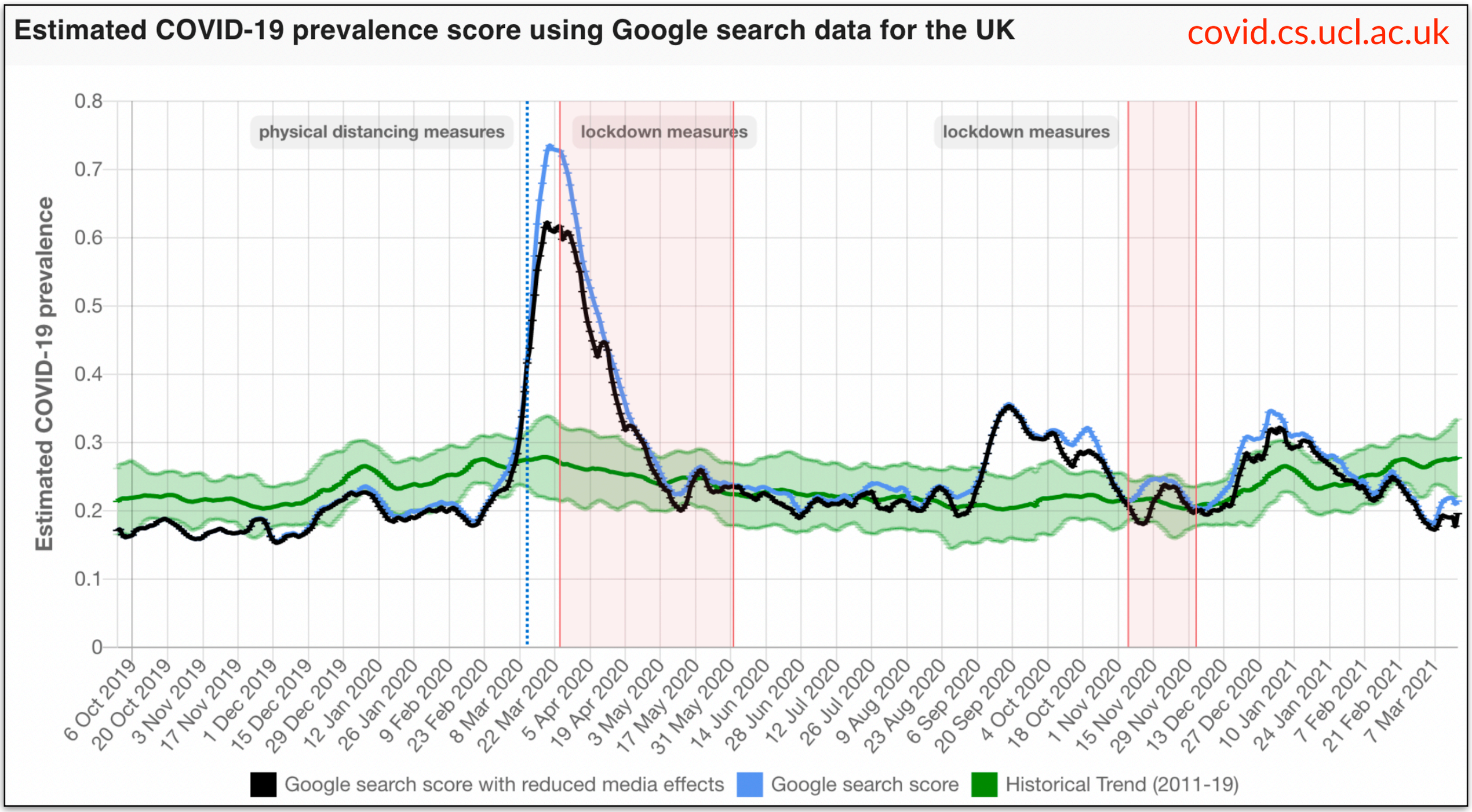
Lampos *et al*. (2021), *npj Digit. Med.*

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**

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

Translation and impact



Public Health England

gov.uk/government/statistics/national-flu-and-covid-19-surveillance-reports

Take-aways

- **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?

Acknowledgements

Collaborators

Ingemar J. Cox (*UCL*), Elad Yom-Tov (*Microsoft Research*), Richard Pebody (*WHO*), Bin Zou (*UCL*), Andrew Miller (*Apple*), Michael Edelstein (*Bar Ilan*), Maimuna Majumder (*Harvard*), Lele Rangaka (*UCL*), Rachel McKendry (*UCL*), Michael Morris (*UCL*), Moritz Wagner (*LSHTM*), and many more

Contributing Organisations

Microsoft Research, Google, Royal College of General Practitioners (RCGP), Public Health England (PHE)

Funding

EPSRC (*i-sense*), Google, MRC (*VirusWatch*)

@lampos

lampos.net

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