Statistics and Modeling with Novel Data Streams 2025

Case studies characterizing flu, dengue, Zika, Ebola, COVID-19 epidemics.

Mauricio Santillana, PhD and Raul Garrido, M.S.

Northeastern University

Northeastern University
Network Science Institute

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Statistics and Modeling with Novel Data Streams

Mauricio Santillana, Ph.D in Computational and Applied Mathematics

- Professor of Physics and Electrical and Computer Engineering, Northeastern University
- Adjunct Professor of Epidemiology, Harvard T.H. Chan School of Public Health
- Director, Machine Intelligence Group for the betterment of Health and the Environment,
- Core Faculty Member, Network Science Institute, Northeastern University
- Professor of Computer Science, Northeastern University (by courtesy)
- Professor of Health Sciences, Northeastern University (by courtesy)

Raul Garrido, M.S. in Physics

- PhD Student in Physics, Northeastern University
- Team Member, Machine Intelligence Group for the betterment of Health and the Environment,
- Team Member, Network Science Institute, Northeastern University

Mauricio Santillana
Principal Investigator

Fred Lu

Leo Cazares

Nicole Kogan

Gemma Llano

Tamanna Urmi

Leo Clemente

Shihao Yang

Binod Pant

Matt Levine

Xiyu Yang

Raul Garrido

Yash Bhora

Jinho Park

George Dewey

Ang Barrett

Skyler Wu

Anjalika Nande

Austin Meyer

Iris Lang

MIGHTE Team

Machine Intelligence Group
for the betterment of Health
and the Environment

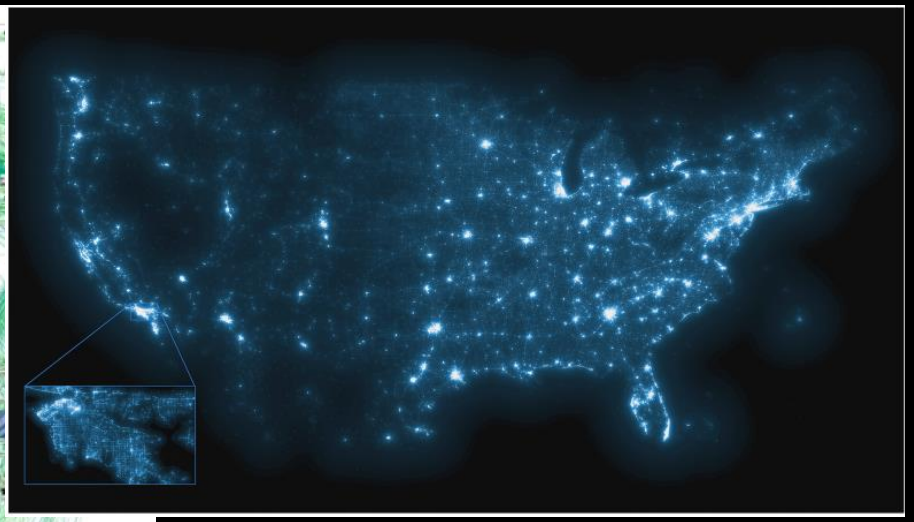
lachine Intelligence Group
r the betterment of Health
nd the Environment

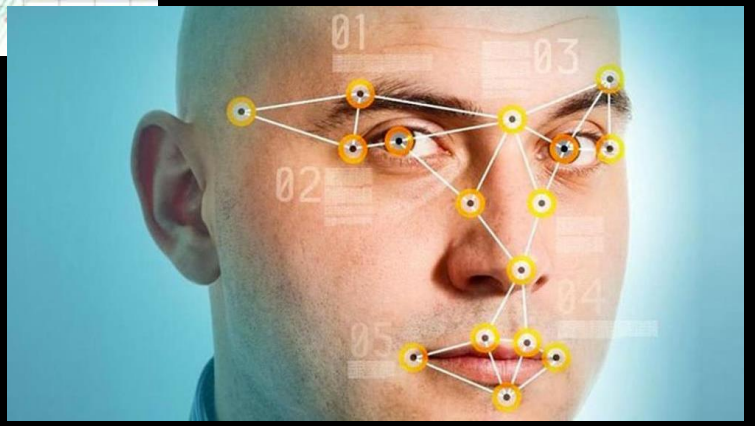
EPISTORM

Areas of **Applied Mathematics** and **Mathematical Physics** used in my research:

- Linear Algebra
- Differential Equation
- Perturbation Theory
- Numerical Analysis
- Optimization
- Data Assimilation
- (Applied) Statistics
- Bayesian Statistics
- Signal Processing
- Time Series Analysis
- Uncertainty Quantification
- Machine Learning

Biosurveillance is a process of gathering, integrating, interpreting, and communicating essential information that might relate to disease activity and threats to human, animal, or plant health.





Big data

Trillions of sensors are monitoring, tracking, and communicating information from multiple locations in real-time

Newspaper articles, Reports, etc…

Predictive Analytics

30+ petabytes of user-generated data stored, accessed, and analyzed

Google

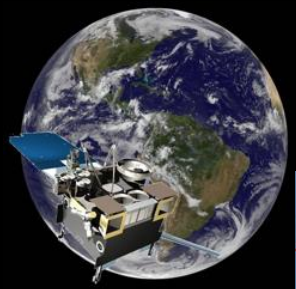
Over 1 **billion** Google searches a day

~2 billion smartphones world wide

230 million tweets every day

Source: IBM

Big data



Weather information in real-time

Newspaper articles, Reports, etc…

30+ petabytes of user-generated data stored, accessed, and analyzed

Medicine and Public Health

Electronic Health Records (**EHR**)

Google

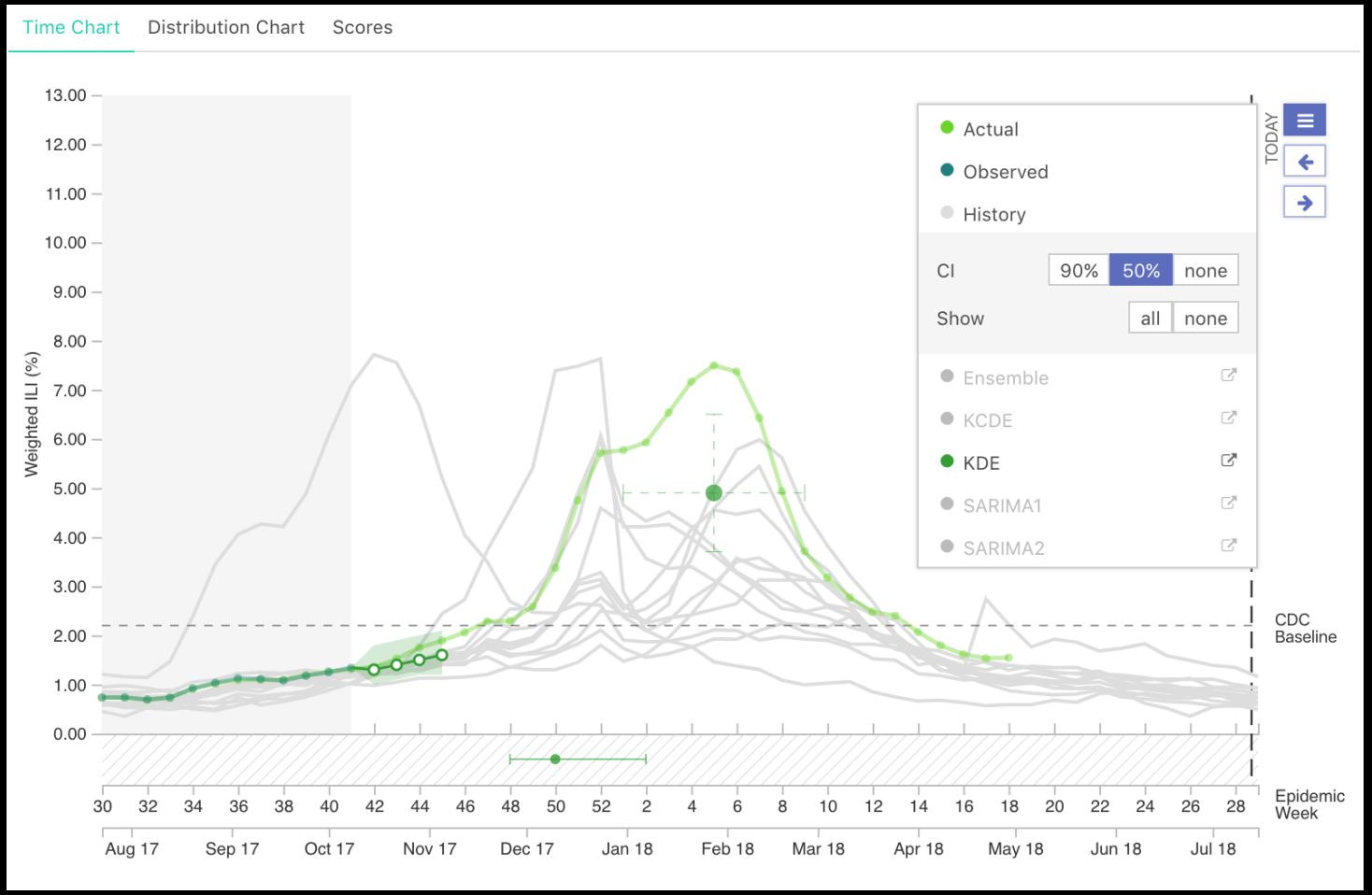
Over 1 **billion** Google searches a day

~2 billion smartphones world wide

230 million tweets every day

The challenge…

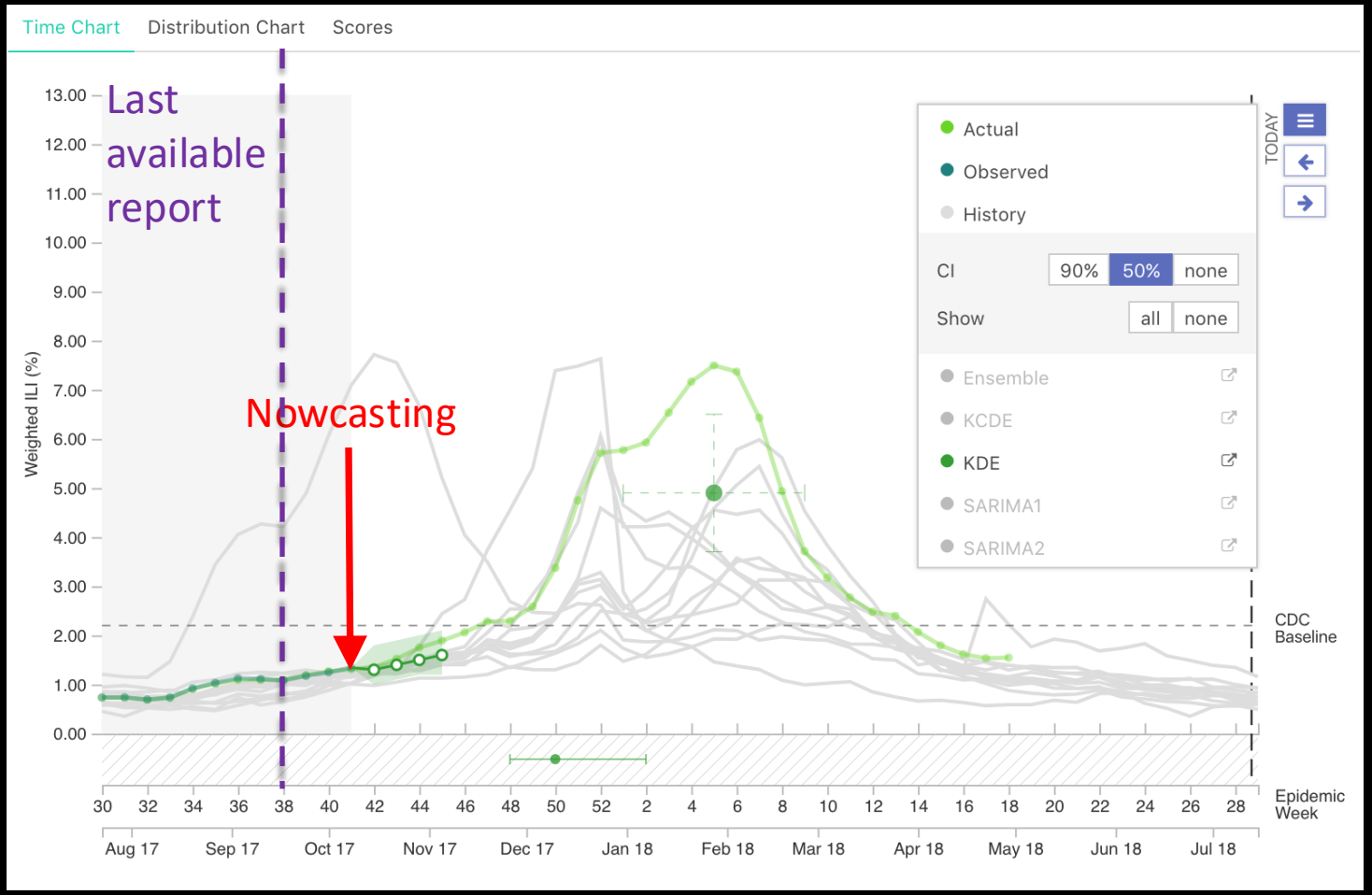
Real-time monitoring of disease activity, **short-term** forecasting (weeks), **long-term** forecasting (months)



Visualization borrowed from: http://reichlab.io/flusight/

The challenge…

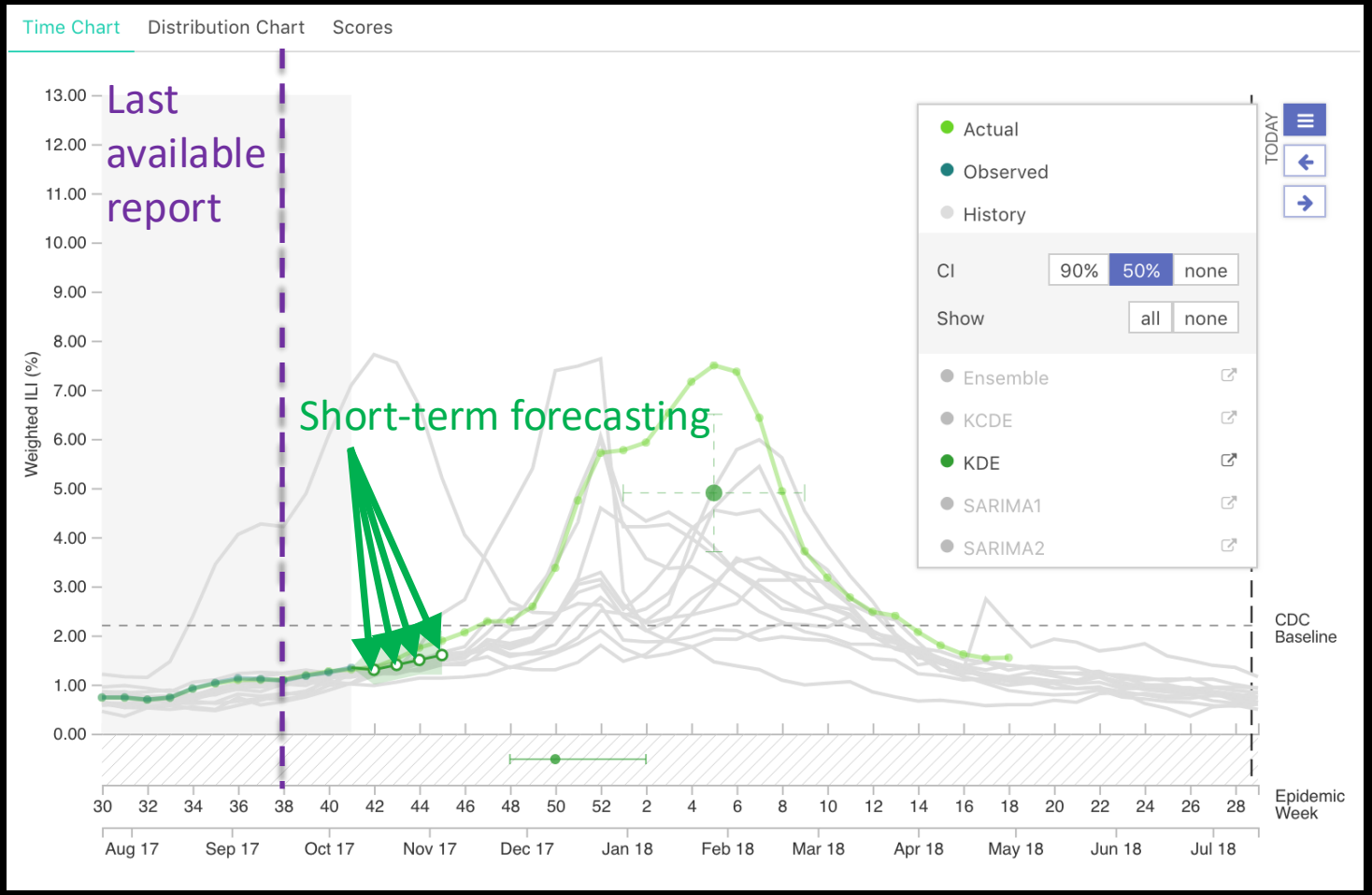
Real-time monitoring of disease activity, **short-term** forecasting (weeks), **long-term** forecasting (months)



Visualization borrowed from: http://reichlab.io/flusight/

The challenge…

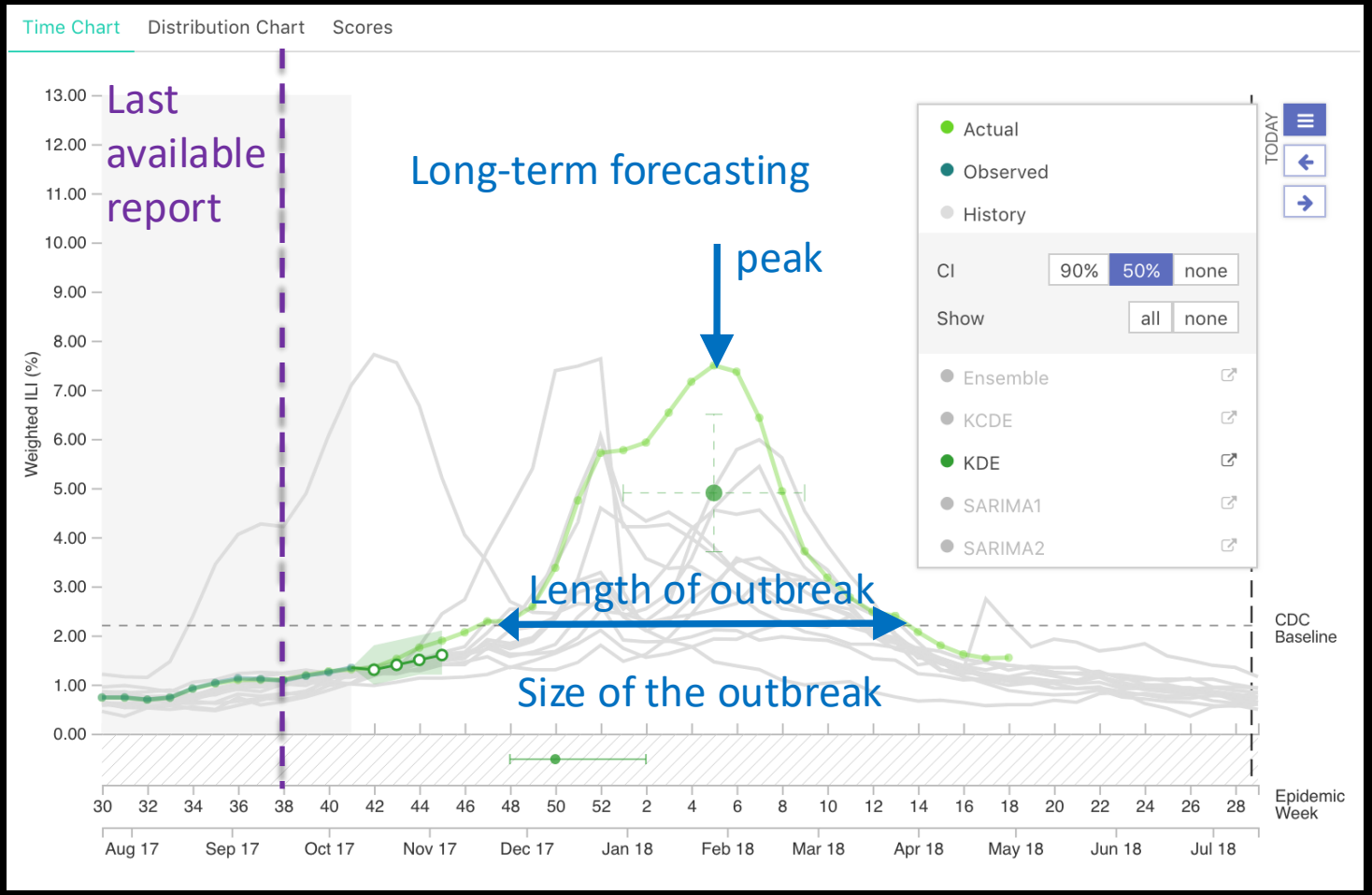
Real-time monitoring of disease activity, **short-term** forecasting (weeks), **long-term** forecasting (months)



Visualization borrowed from: http://reichlab.io/flusight/

The challenge…

Real-time monitoring of disease activity, **short-term** forecasting (weeks), **long-term** forecasting (months)

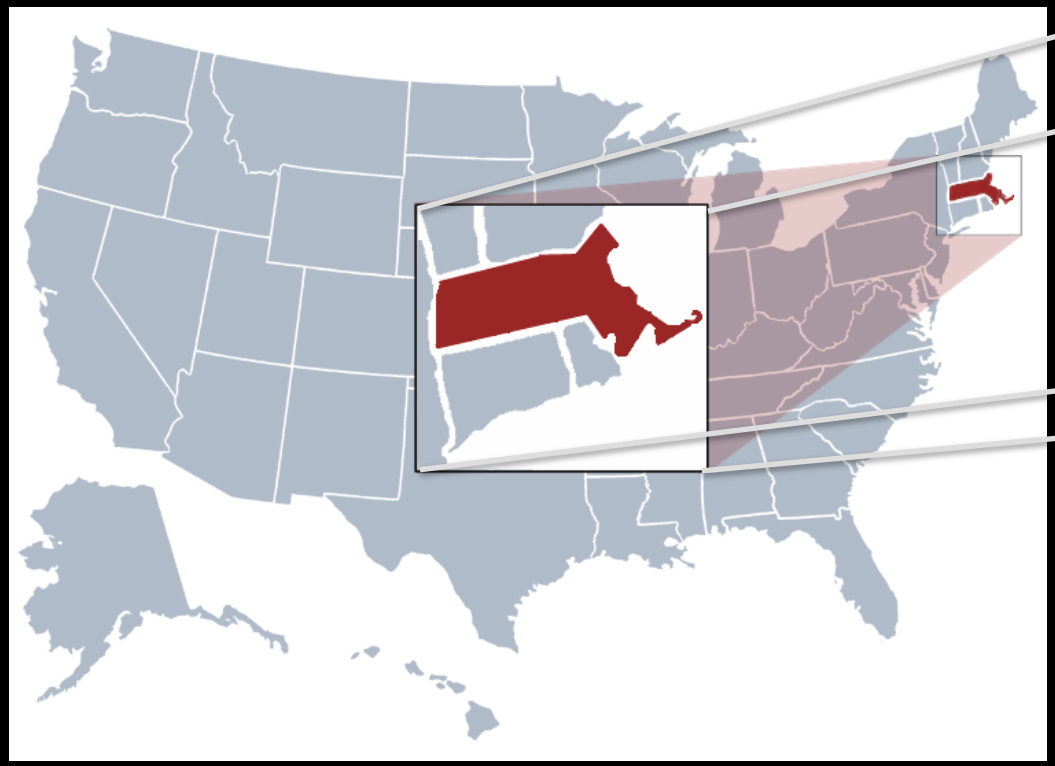


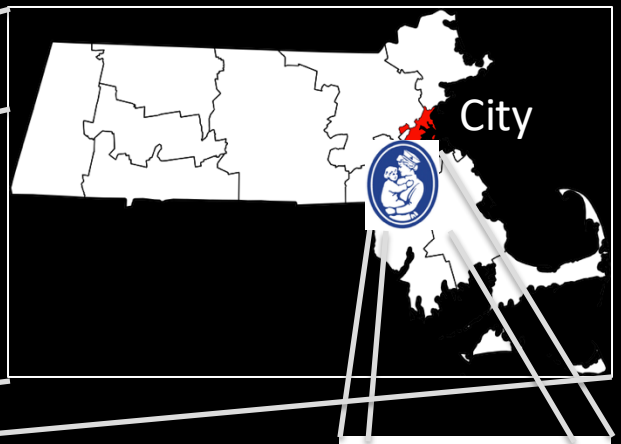
Visualization borrowed from: http://reichlab.io/flusight/

At what spatial resolution and time frequency?

Space: Country-level, state-level, city-level, neighborhood, hospital, patient?
Time: monthly, weekly, daily, hourly?

State



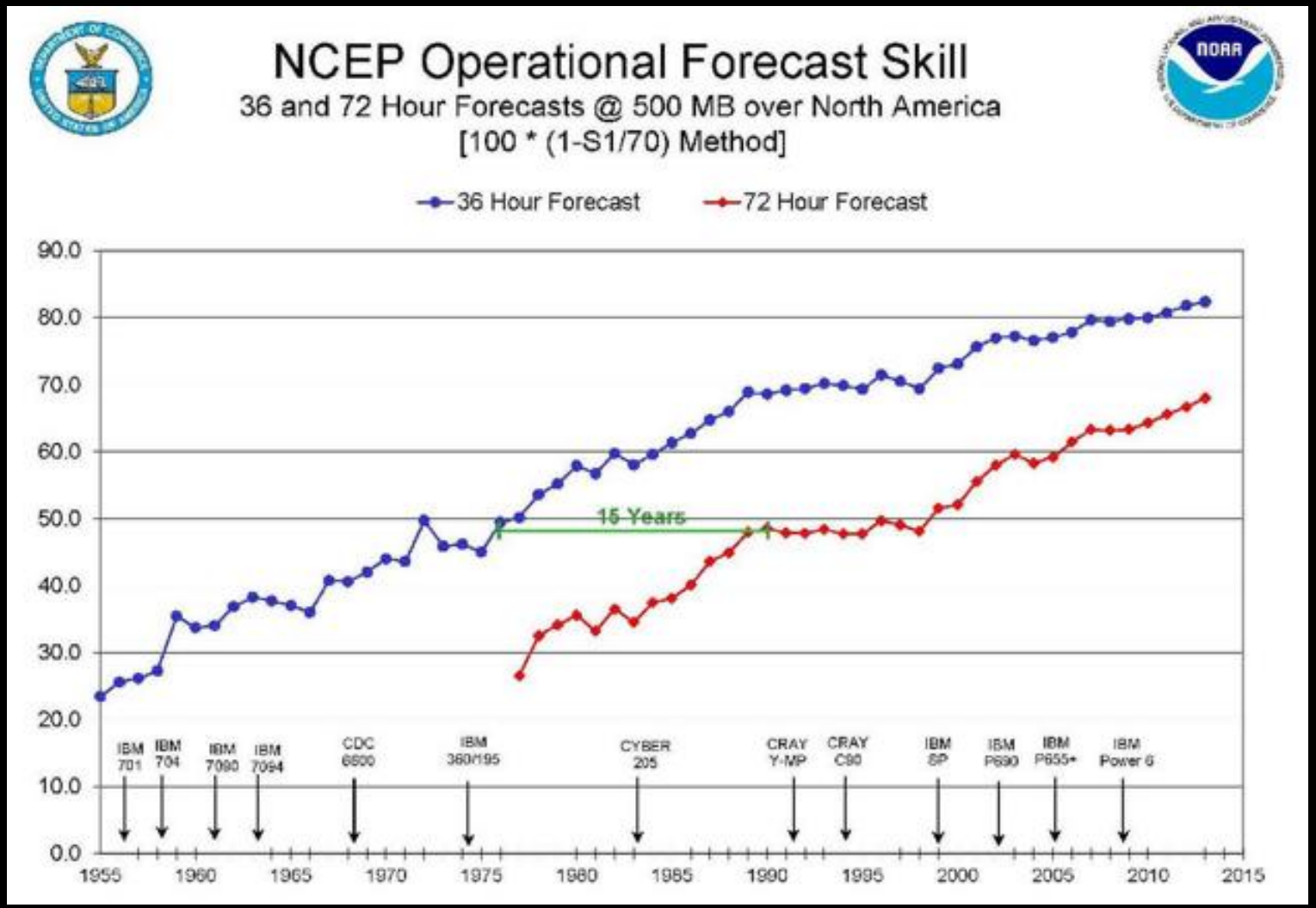


Hospital

Country

Patient?

What if we get it right?



Real-time tracking vs predictions of disease incidence/risk
Similarities and differences with weather prediction

Traditionally, epidemiologist use Ordinary Differential Equations to model epidemic outbreaks

A Contribution to the Mathematical Theory of Epidemics.

By W. O. Kermack and A. G. McKendrick.

(Communicated by Sir Gilbert Walker, F.R.S.—Received May 13, 1927.)

(From the Laboratory of the Royal College of Physicians, Edinburgh.)

Introduction.

(1) One of the most striking features in the study of epidemics is the difficulty of finding a causal factor which appears to be adequate to account for the magnitude of the frequent epidemics of disease which visit almost every population. It was with a view to obtaining more insight regarding the effects of the various factors which govern the spread of contagious epidemics that the present investigation was undertaken. Reference may here be made to the work of Ross and Hudson (1915–17) in which the same problem is attacked. The problem is here carried to a further stage, and it is considered from a point of view which is in one sense more general. The problem may be summarised as follows: One (or more) infected person is introduced into a community of individuals, more or less susceptible to the disease in question. The disease spreads from

Traditionally, epidemiologist use Ordinary Differential Equations to model epidemic outbreaks

In this case the equations are

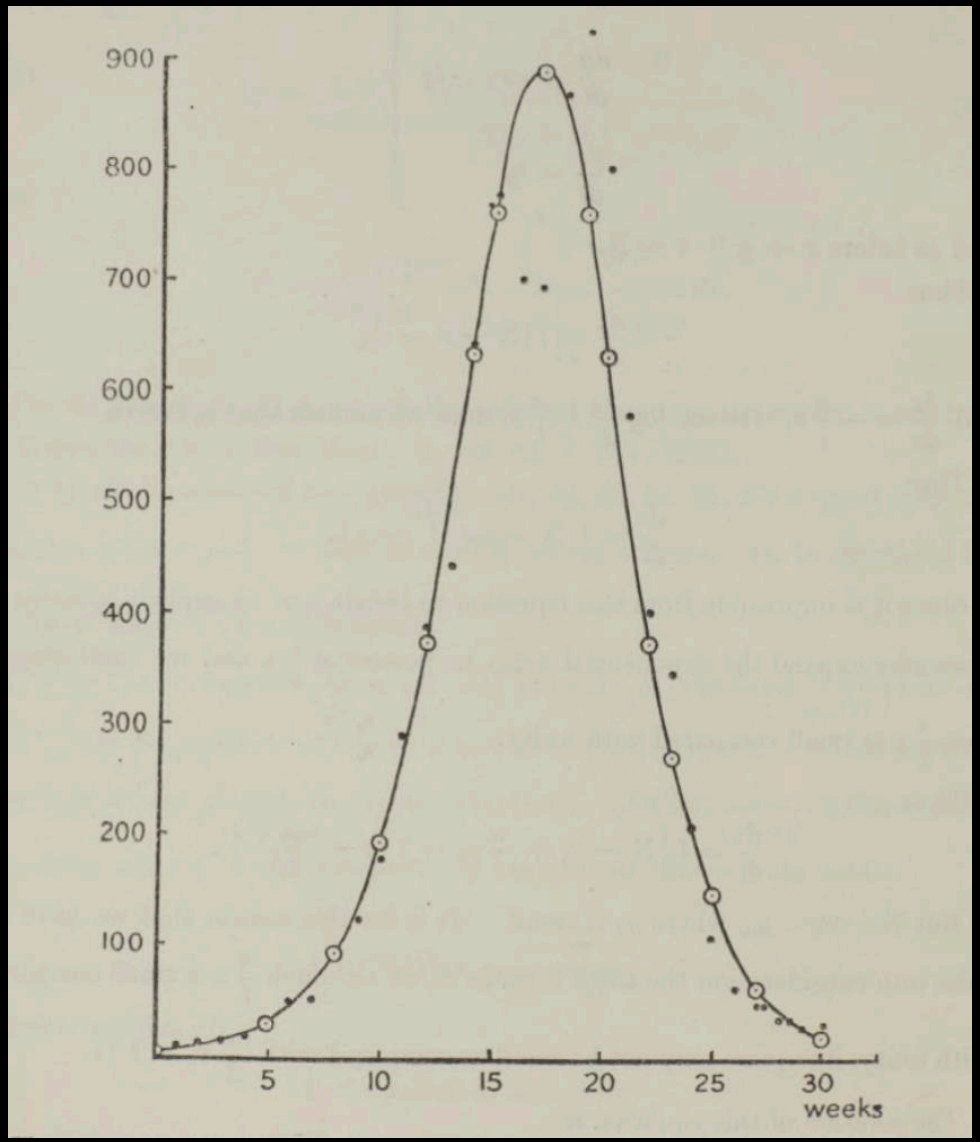
$$\left.\begin{aligned} \frac{dx}{dt} &= -\kappa xy \\ \frac{dy}{dt} &= \kappa xy - ly \\ \frac{dz}{dt} &= ly \end{aligned}\right\}$$

and as before $x + y + z = \mathrm{N}$.

Thus

$$\frac{dz}{dt} = l(\mathrm{N} - x - z),$$

Traditionally, epidemiologist use Ordinary Differential Equations to model epidemic outbreaks

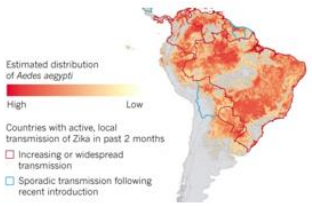


Approach

Data streams

Twitter and Wikipedia activity

Weather variables

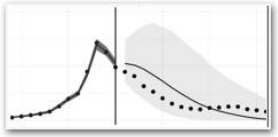


Mosquito prevalence

Human mobility

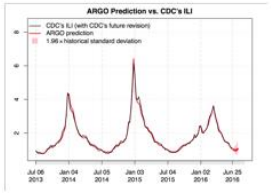
Territory

Modeling approaches



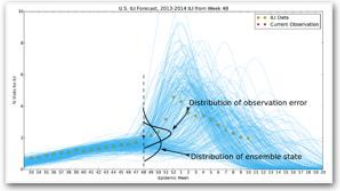
Mechanistic approaches

+



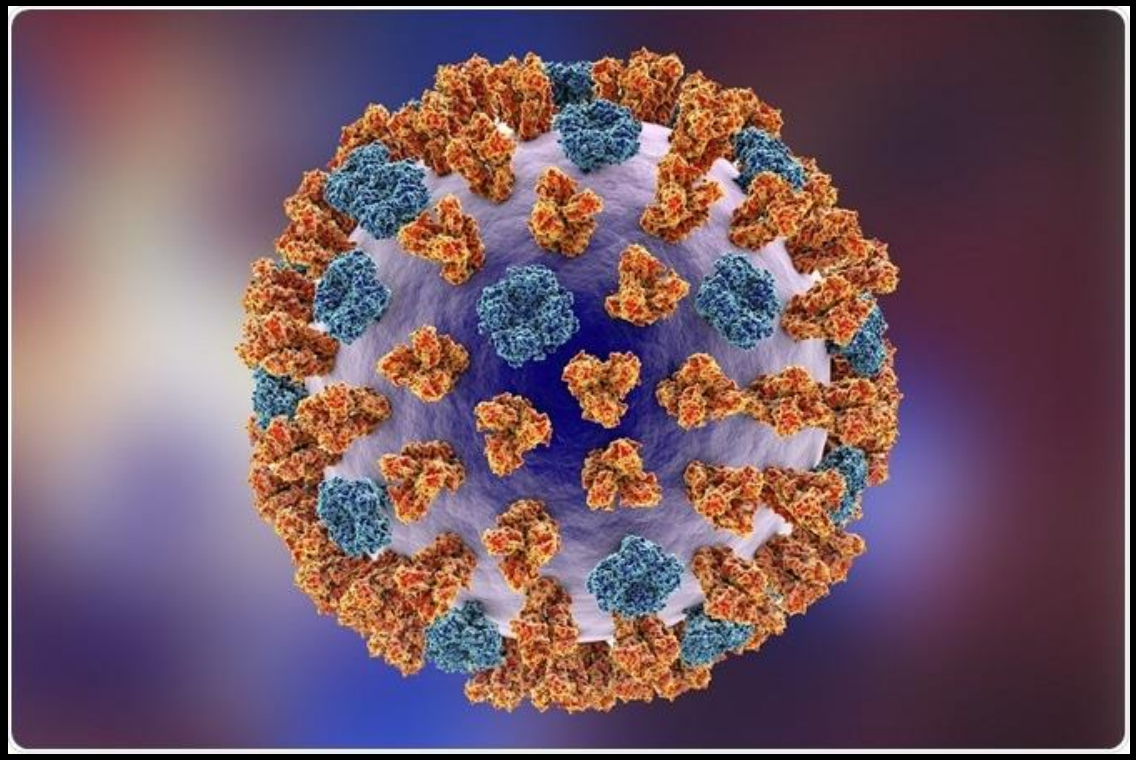
Machine-learning approaches

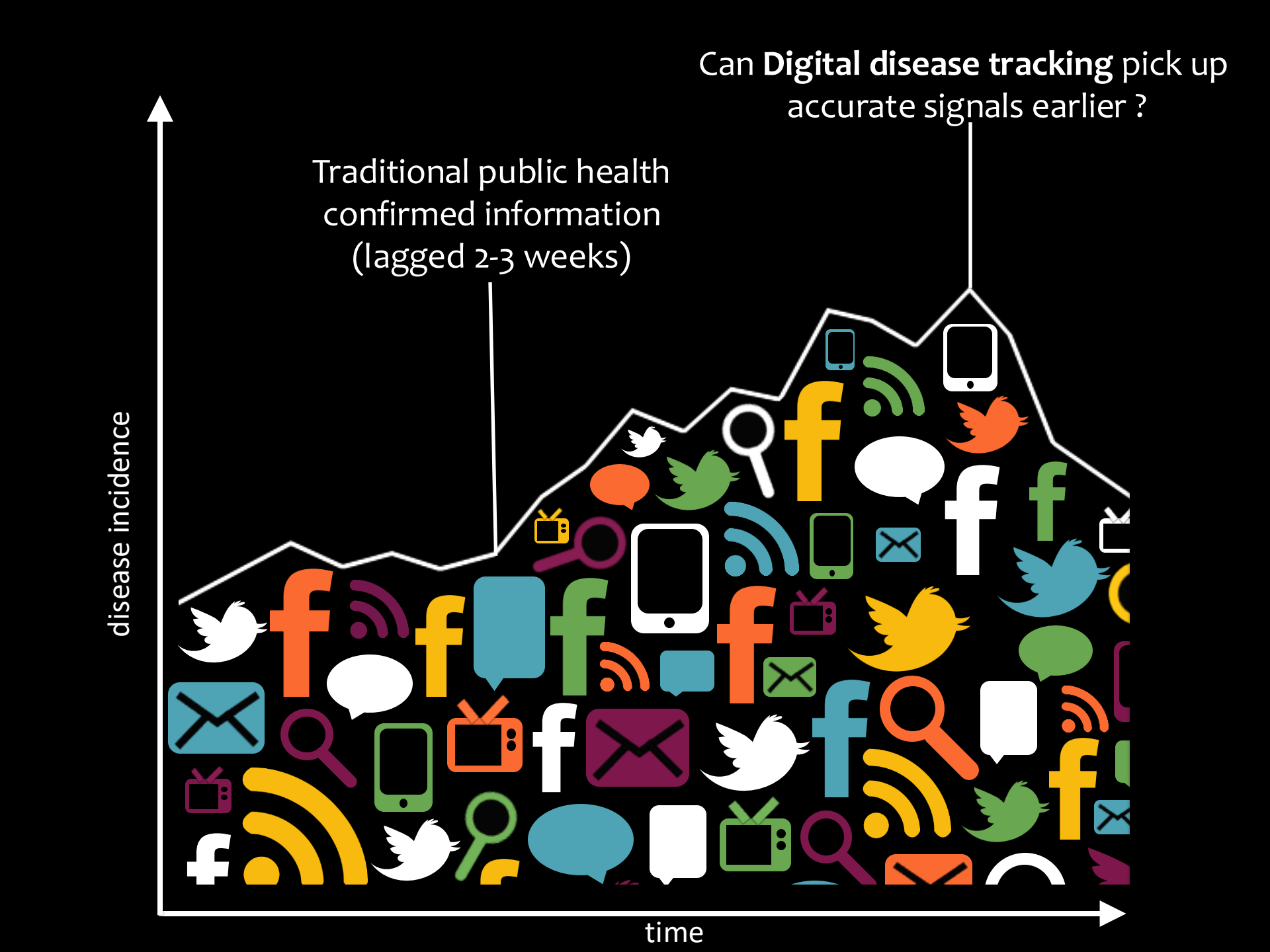
=

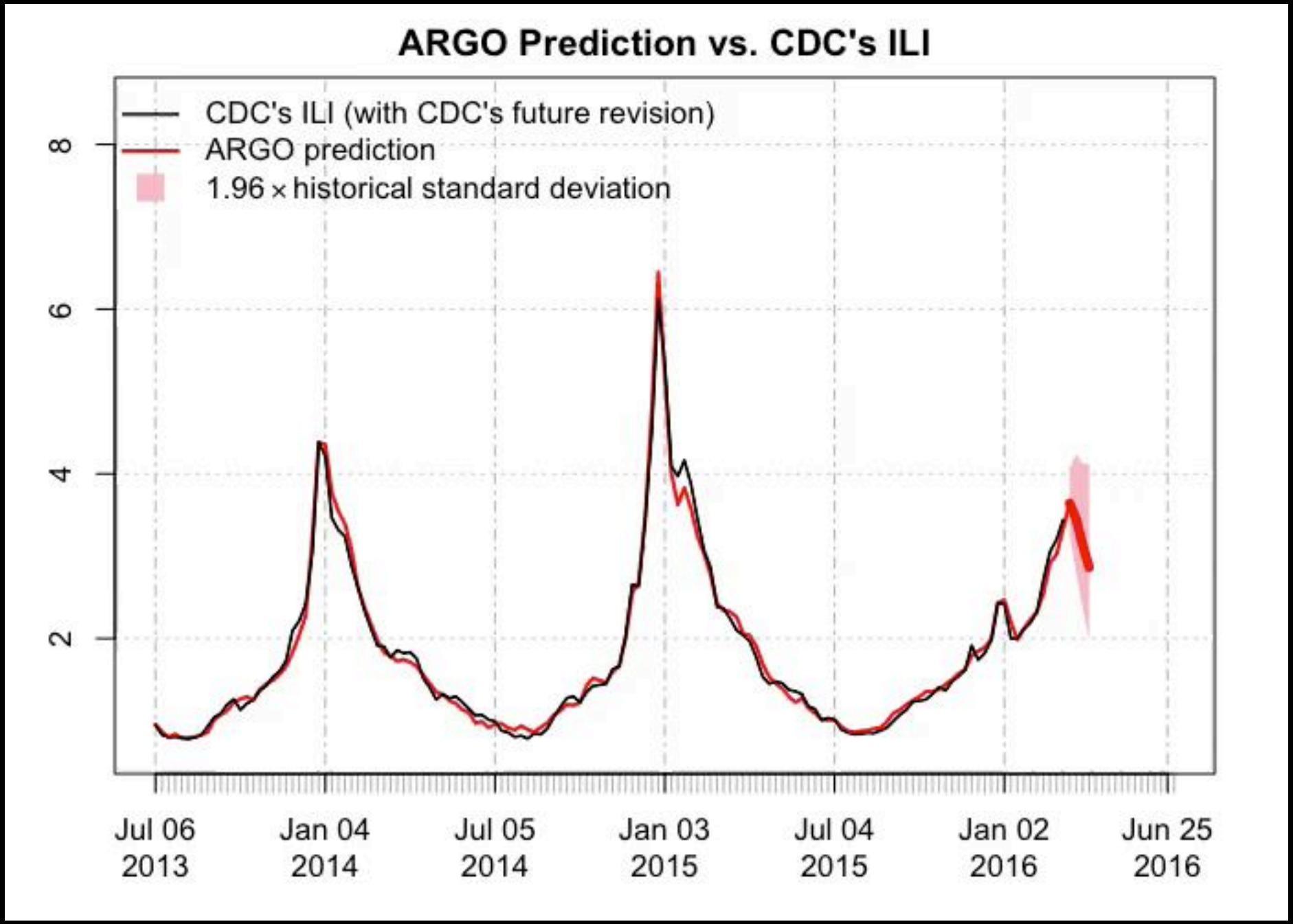


Ensemble forecasting approaches

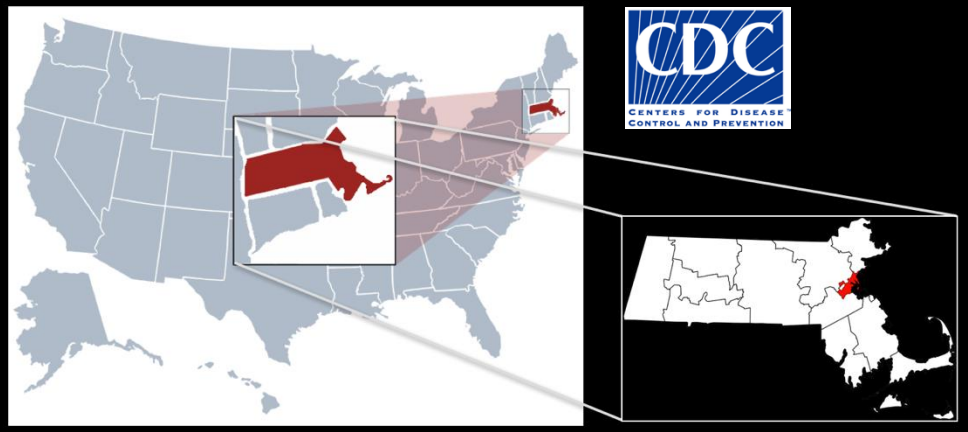
Background: monitoring influenza in rich nations

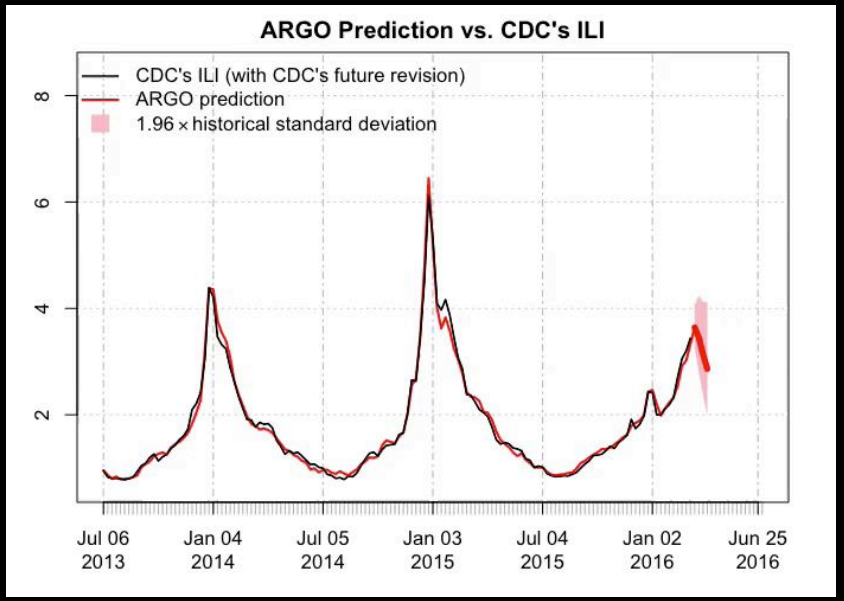



Can **Digital disease tracking** pick up accurate signals earlier ?
Traditional public health confirmed information (lagged 2-3 weeks)
disease incidence
time


ARGO Prediction vs. CDC's ILI
CDC's ILI (with CDC's future revision)
ARGO prediction
1.96 × historical standard deviation
2
4
6
8
Jul 06
2013
Jan 04
2014
Jul 05
2014
Jan 03
2015
Jul 04
2015
Jan 02
2016
Jun 25
2016

Part 1. Previous success stories in tracking and forecasting Influenza in data-rich high-income countries: USA





1. Multiple spatial resolutions: National, multi-state, state, city-level

2. Multiple data sources (hybrid systems): traditional healthcare-based, EHR, Google, Twitter, Crowd-sourced disease surveillance.

Seminal work by Google

The promise of big data in public health

GOOGLE FLU TRENDS

Google Flu Trends

nature International weekly journal of science

Letter

Nature **457**, 1012-1014 (19 February 2009) | doi:10.1038/nature07634; Received 14 August 2008; Accepted 13 November 2008; Published online 19 November 2008; Corrected 19 February 2009

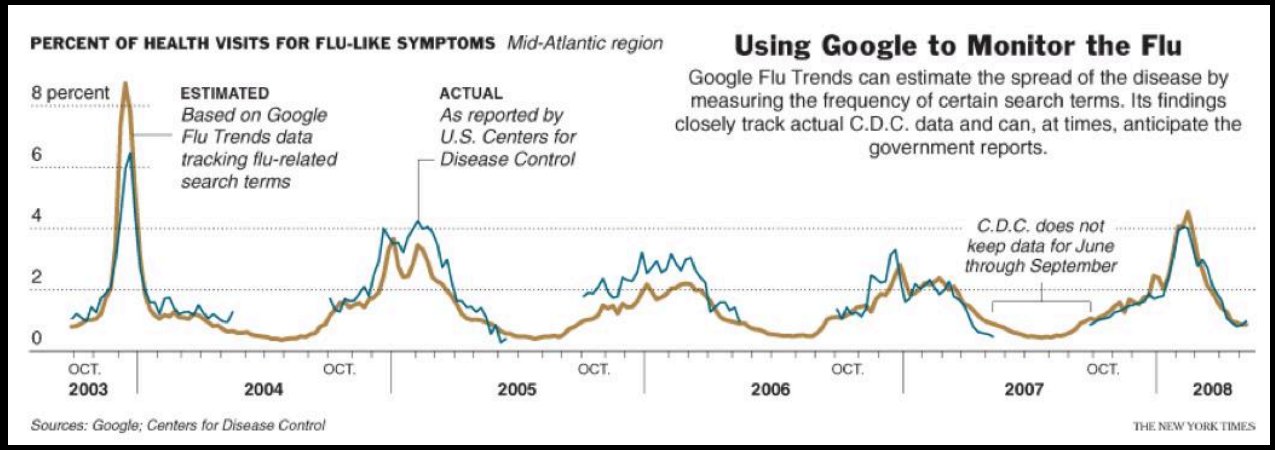
Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

1. Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043, USA
2. Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Atlanta, Georgia 30333, USA

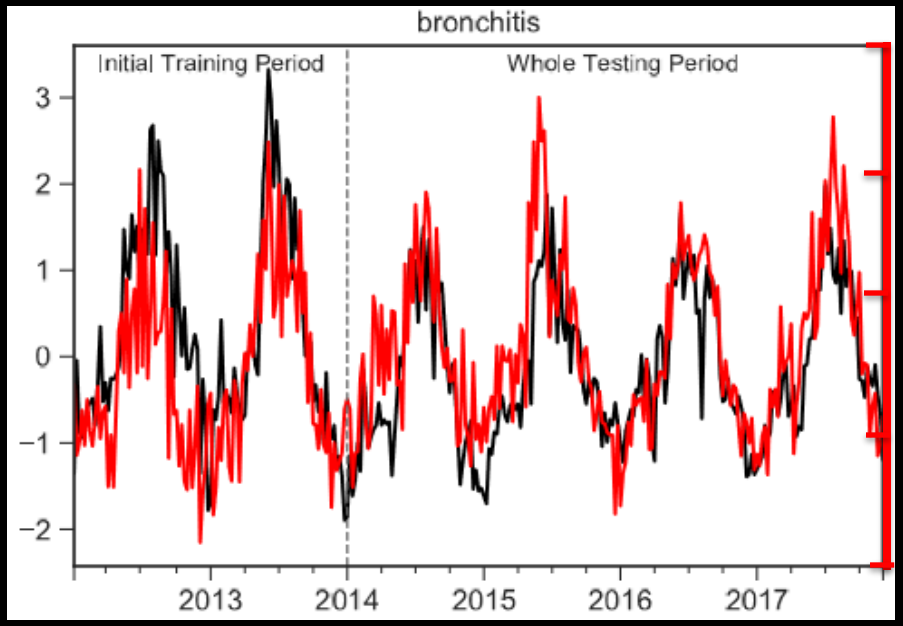
Correspondence to: Matthew H. Mohebbi[1] Correspondence and requests for materials should be addressed to J.G. or M.H.M. (Email: flutrends-support@google.com).

The New York Times

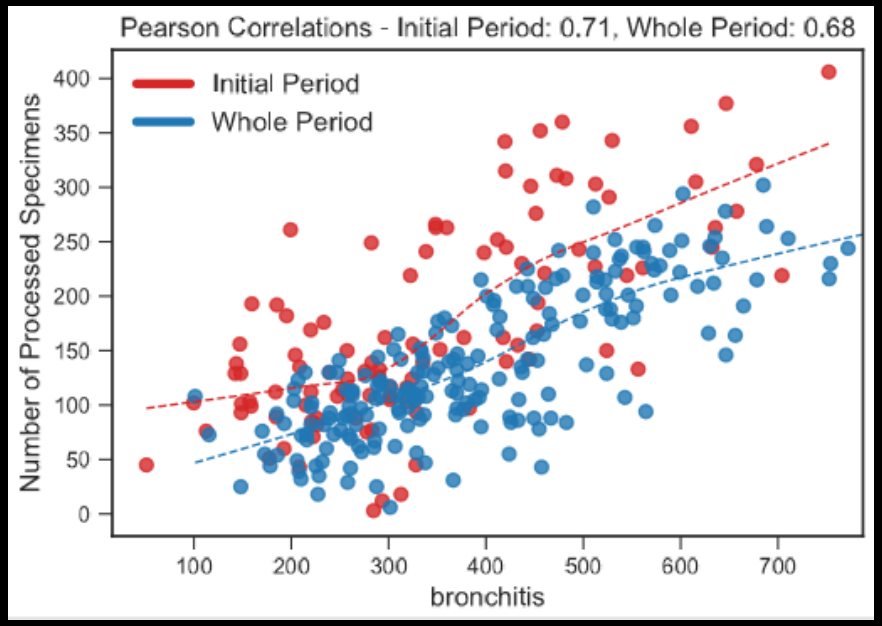


Google Flu Trends

What is the logic behind this approach?



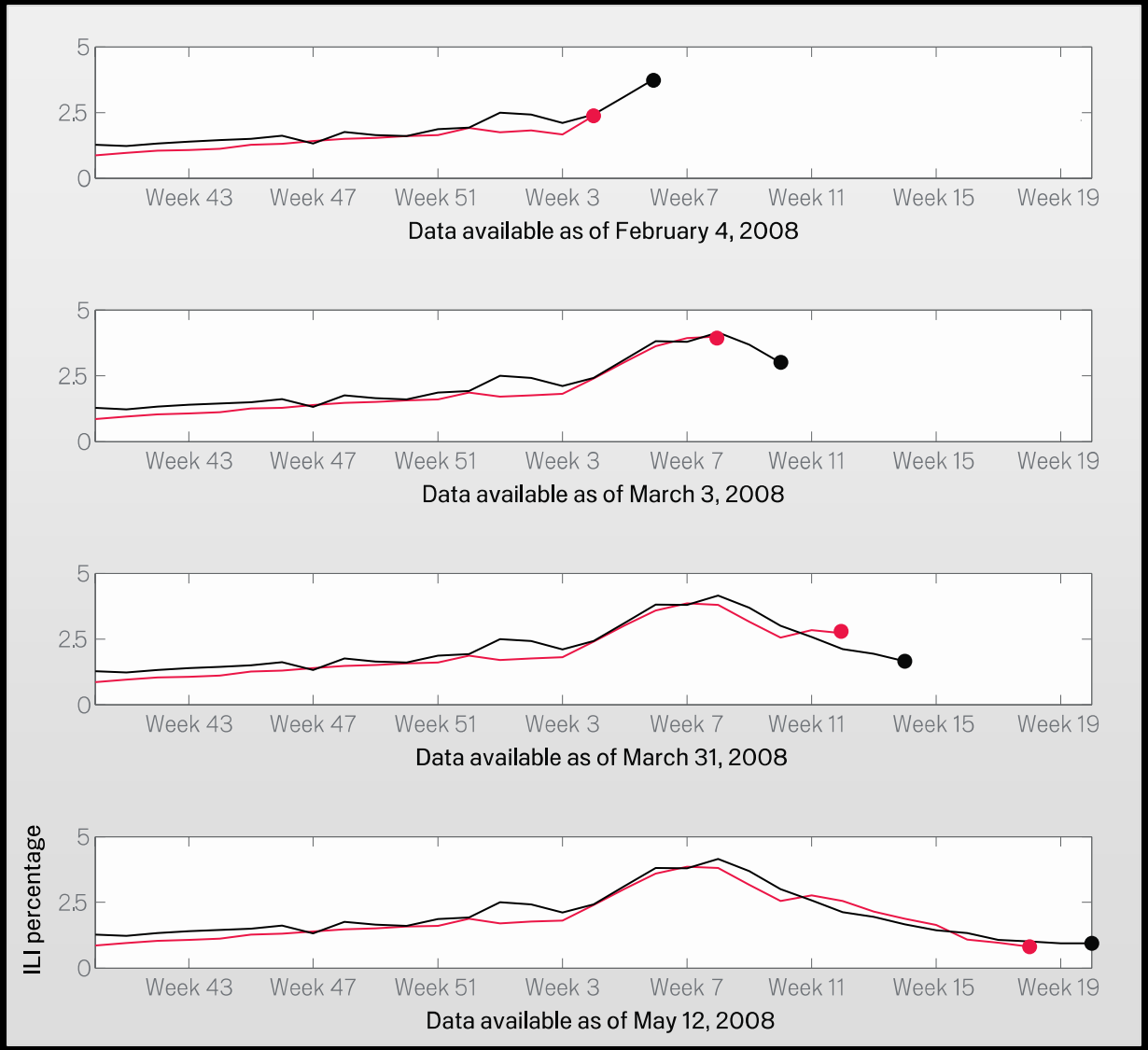
Searches on “bronchitis” vs Flu activity in South Africa



Scatter plot of searches vs flu

google.org Flu Trends

Epidemiological information available 2-3 weeks ahead of traditional clinical tracking systems

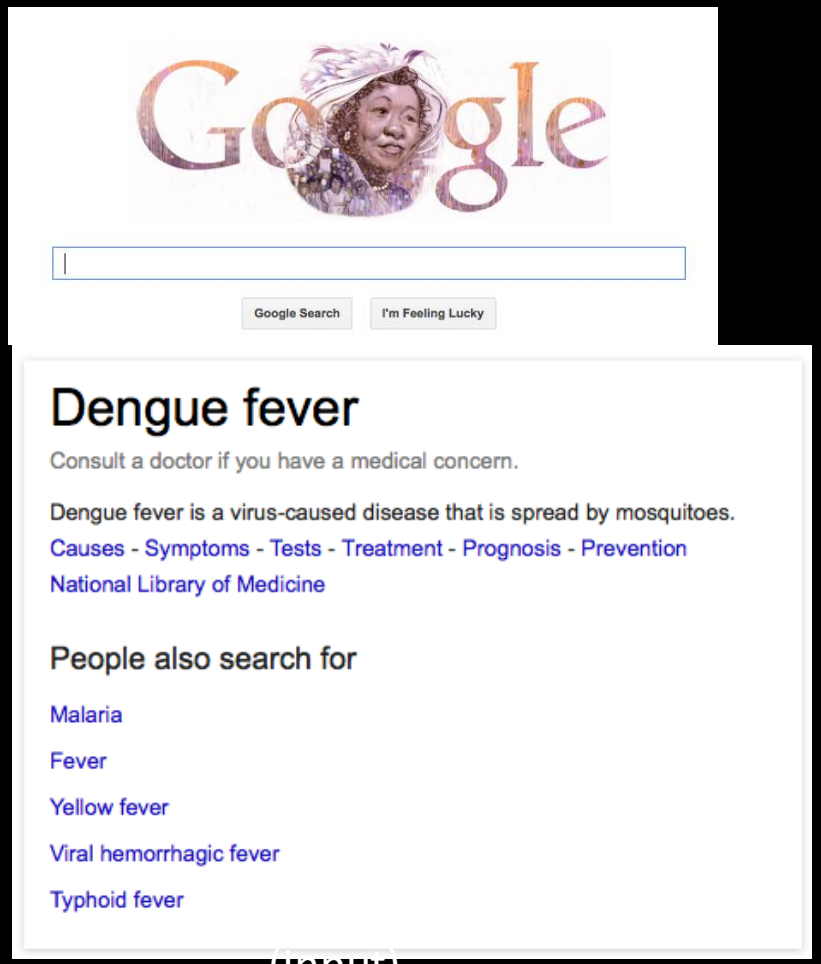


We started working on Exercise 1

(Static training, static including lag 1)

Supervised machine learning examples:

Given the number of Google searches associated to the term “dengue”, and given the number of confirmed cases of dengue in Mexico from 2004 to 2006 (**Training period**), can we estimate how many people will most likely get dengue based on the number of searches during the subsequent years?



(Input)

(Output)

4. Least squares in Public Health. (30 points)
Dengue fever is a virus-caused disease that is spread by mosquitoes that affects millions of people in tropical environments around the Globe. In this problem, you are asked to construct a simple version of the digital disease detection tool: "Google Dengue Trends" for Mexico. For this, you will download the spreadsheet `Dengue trends AM 111.xls` from the course website. The first column in the spreadsheet represents the date (in months, from 2004-2011), the second column represents the number of Google searches of the term "dengue" in Mexico, in a given month. The third column represents the number of cases of Dengue in Mexico, as reported by the Mexican Ministry of Health. You may use Matlab or Excel for this problem.

(a) Plot the number of cases of Dengue as a function of time.

(b) For the **training period** 2004-2006 (36 months), find the best line that explains the number of cases of Dengue as a function of the number of searches of the term "dengue". You should do this by solving the least squares problem, and you should obtain the value of the y-intercept and the slope.

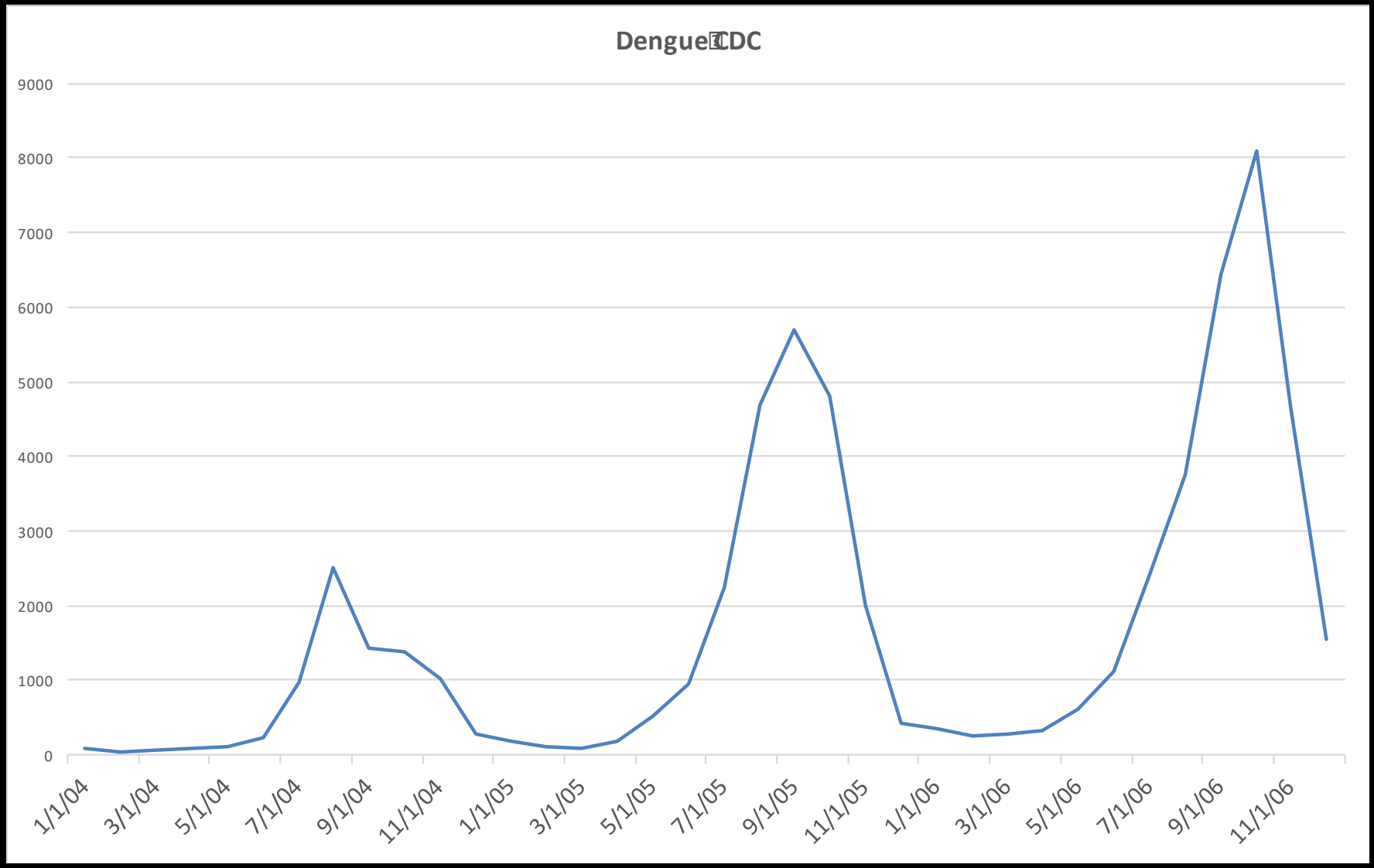
(c) Use the equation of the line you obtained in (b) and plot the number of cases as a function of the number of searches of the term "dengue", predicted by your method during the training period. Compare your results to the plot in (a) for such time period.

(d) For the **prediction or validation period** 2007-2011, use the equation of the line you obtained in (b) to predict the number of the dengue cases as a function of the number of searches of the term "dengue" from 2007-2011. Plot your predictions and compare them to the actual number of cases.

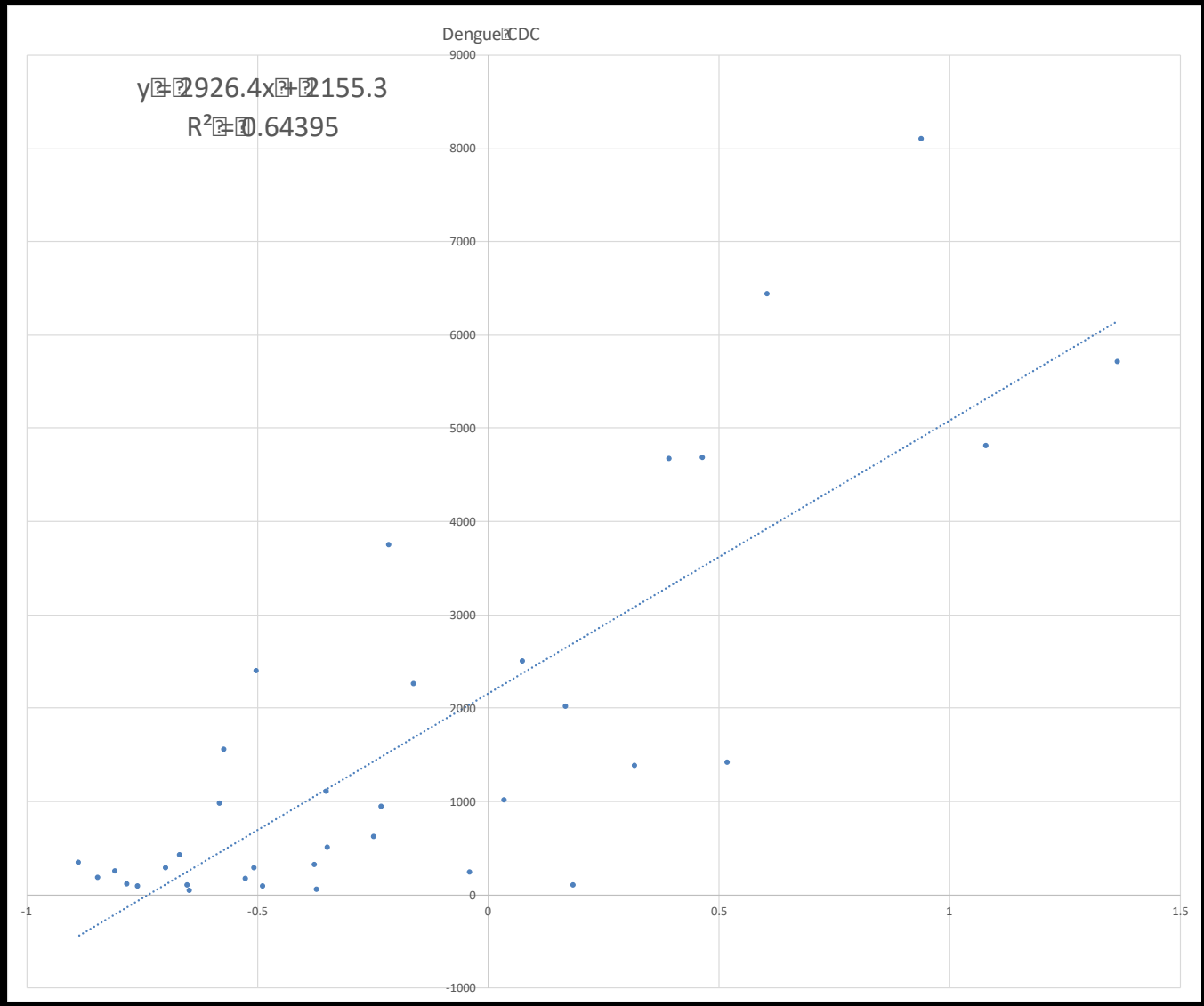
(d) Discuss your results. Could you improve this modeling approach? If so, how?

Did you get it to work?

(a)


Dengue CDC
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1/1/04
3/1/04
5/1/04
7/1/04
9/1/04
11/1/04
1/1/05
3/1/05
5/1/05
7/1/05
9/1/05
11/1/05
1/1/06
3/1/06
5/1/06
7/1/06
9/1/06
11/1/06

(b)



Google search activity

Using Google searches to track diseases statically

begin
%% Load data %%
CDC=load(CDC ILI Data) (One column of values)
X=load(Google search Data) (Multiple columns of values)

%% initialize output array %%
Y=zeros(1: end.of.predictions) (Initialize array to store predictions)

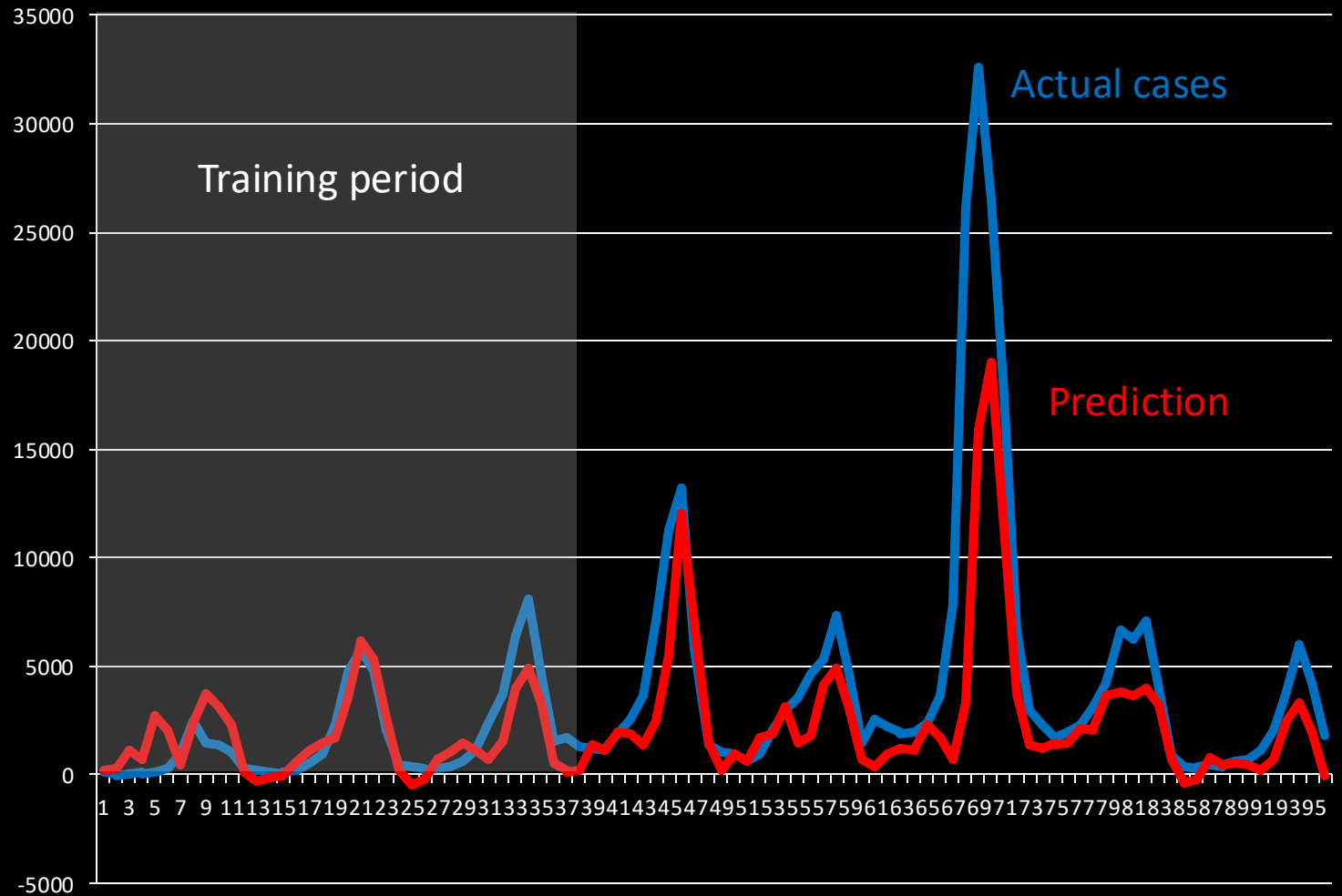
%% train model and produce predictions %%
CDC $\leftarrow$ standardize(CDC) (Perhaps use a transform:z-score, logit)
X $\leftarrow$ standardize(X) (Perhaps use a transform:z-score, logit)
model=LASSOroutine.fit(CDC[1 : *training*] $\sim$ X[1 : *training*]) (Training: in-sample model)

Y[1 : *training*]=
LASSOroutine.predict(model, X[1 : *training*])
(in-sample predictions)
Y[*training* + 1 : *end.of.predictions*]=
LASSOroutine.predict(model, X[*training* + 1 : *end.of.predictions*])
(produce out-of-sample predictions)

end

Supervised machine learning examples:

Static approach, fixed training set



1st lecture ended here

Session 2 started here

How could the previous approach be improved with the given information?

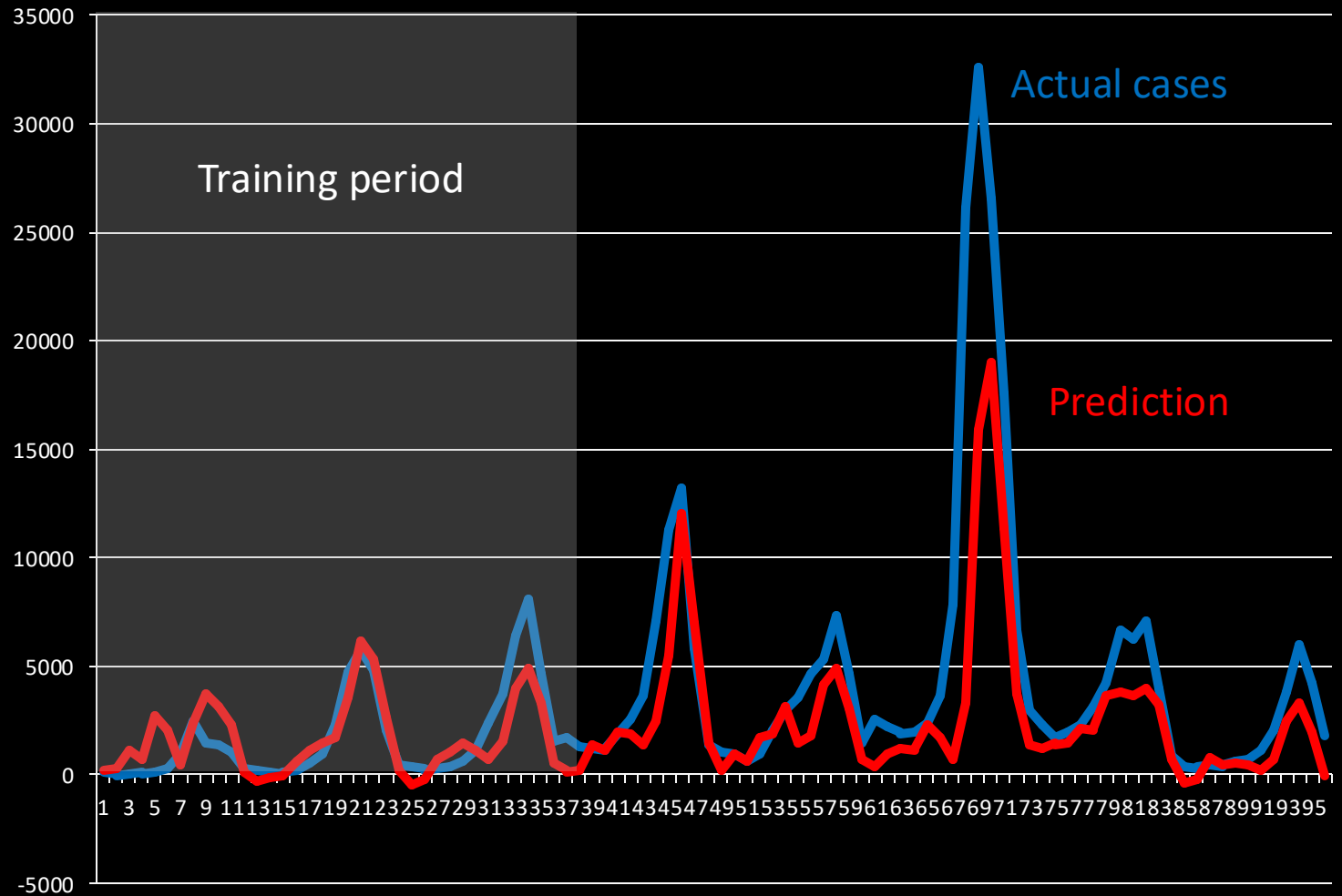
```
begin
%% Load data %%
CDC=load(CDC ILI Data)                (One column of values)
X=load(Google search Data)            (Multiple columns of values)

%% initialize output arrays %%
Y=zeros(1:end.of.predictions)  (Initialize array to store predictions)
coefficients=zeros(1:end.of.predictions)  (Initialize array to store coeffs)

%% train models and produce out-of-sample predictions %%
for i = training : end.of.predictions
    CDC ← standardize(CDC)        (Perhaps use a transform:z-score, logit)
    X ← standardize(X)            (Perhaps use a transform:z-score, logit)
    model=LASSOroutine.fit(CDC[1 : i] ∼ X[1 : i]) (Training: in-sample model )
    coefficients(i) ← model(coefficients)
    Y(i + 1)=LASSOroutine.predict(model, X(i + 1)) (produce out-of-sample
                                                    predictions)
    if(i == training)
            Y[1:i]=LASSOroutine.predict(model, X[1:i]) in -sample predictions
    end
end
end
```

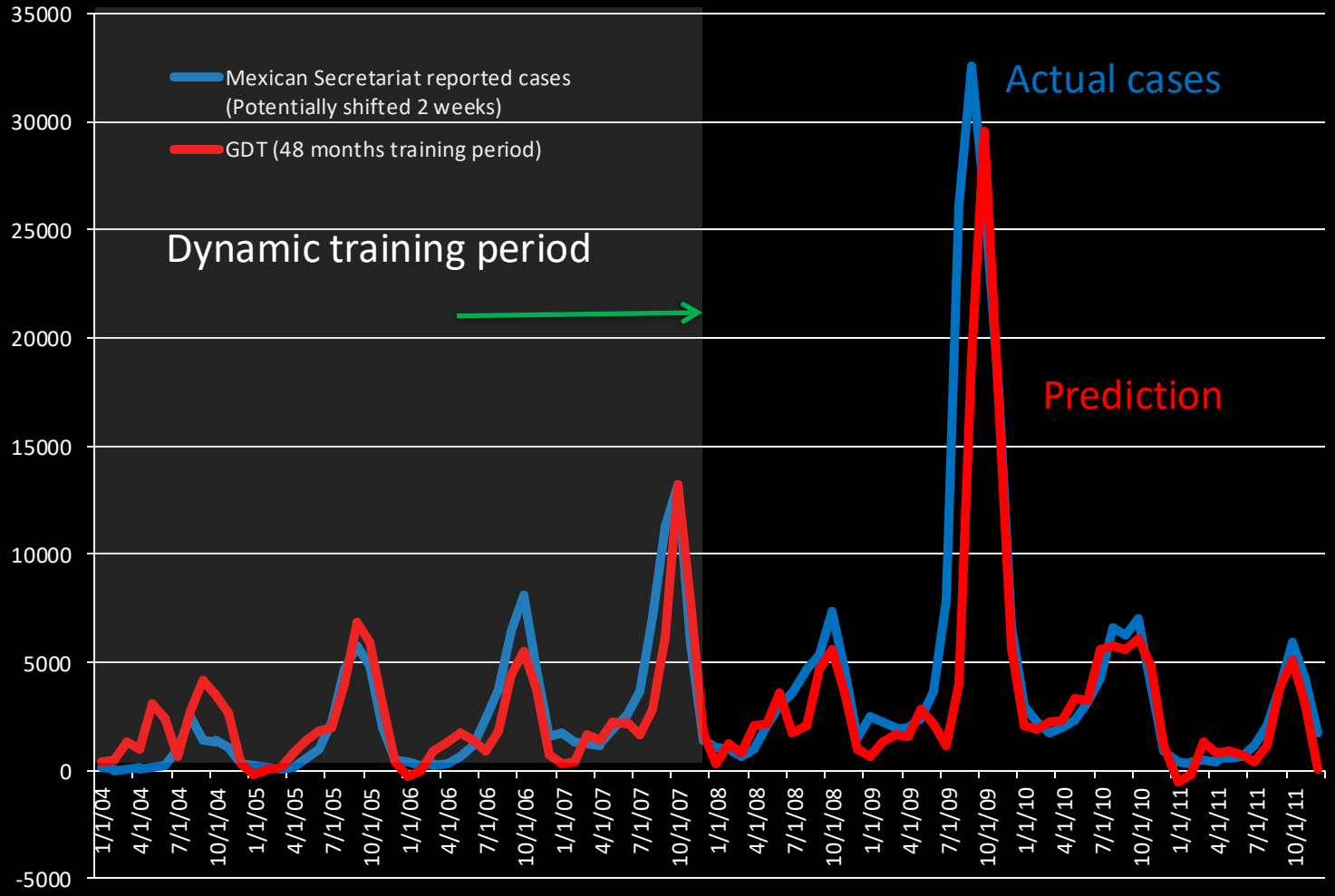
Supervised machine learning examples:

Static approach, fixed training set



Supervised machine learning examples:

Dynamic approach, letting the training set expand as more information becomes available



Seminal work by Google

The promise of big data in public health

GOOGLE FLU TRENDS

Google Flu Trends

nature International weekly journal of science

Letter

Nature **457**, 1012-1014 (19 February 2009) | doi:10.1038/nature07634; Received 14 August 2008; Accepted 13 November 2008; Published online 19 November 2008; Corrected 19 February 2009

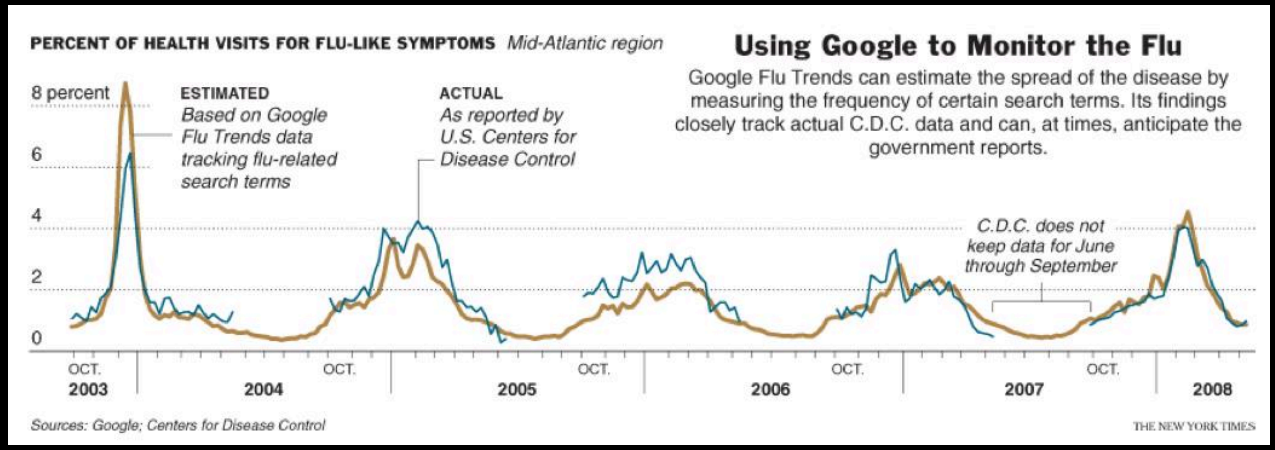
Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

1. Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043, USA
2. Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Atlanta, Georgia 30333, USA

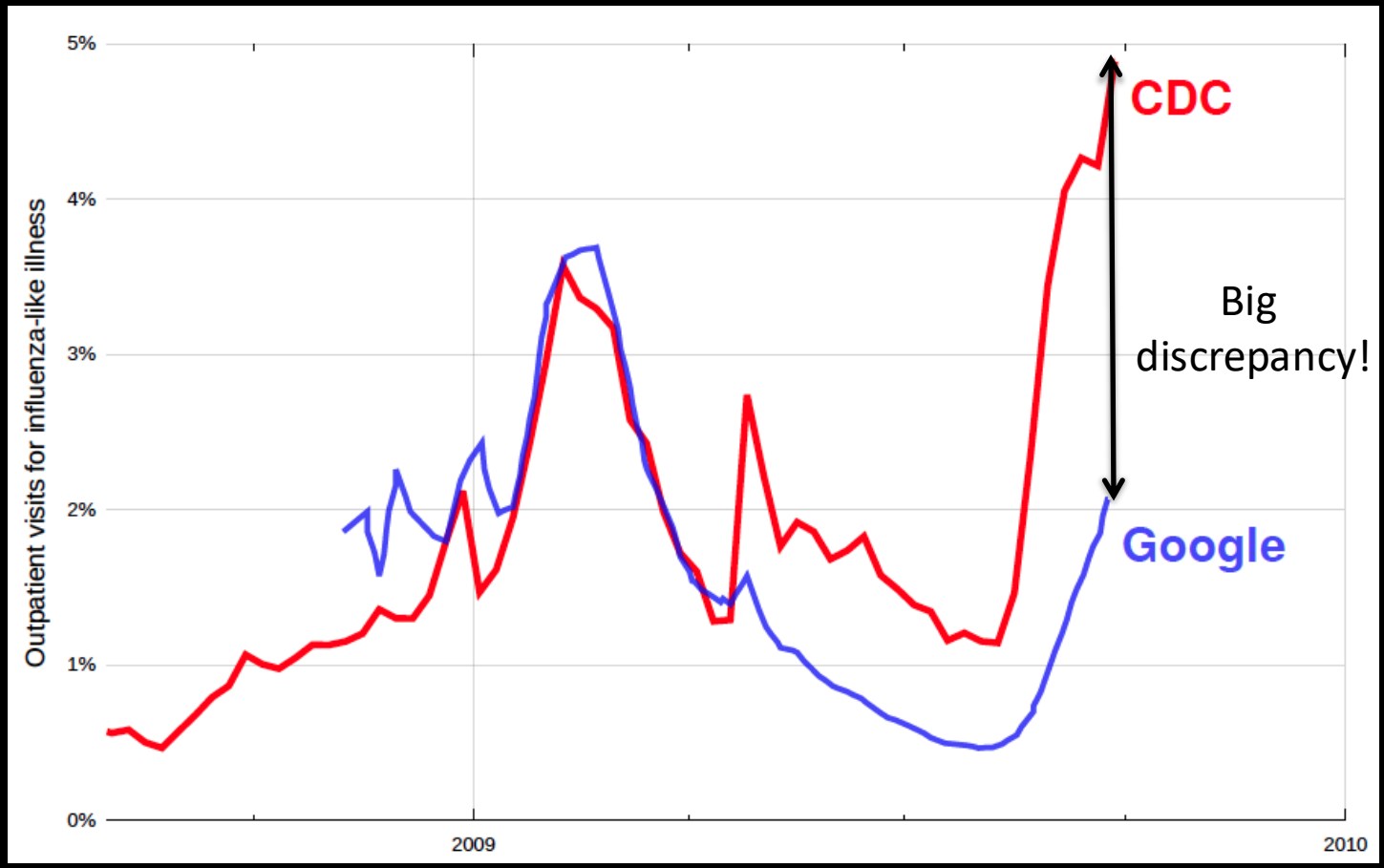
Correspondence to: Matthew H. Mohebbi[1] Correspondence and requests for materials should be addressed to J.G. or M.H.M. (Email: flutrends-support@google.com).

The New York Times



Real-time performance, first year…

Big errors seen during H1N1 pandemic (off-season)

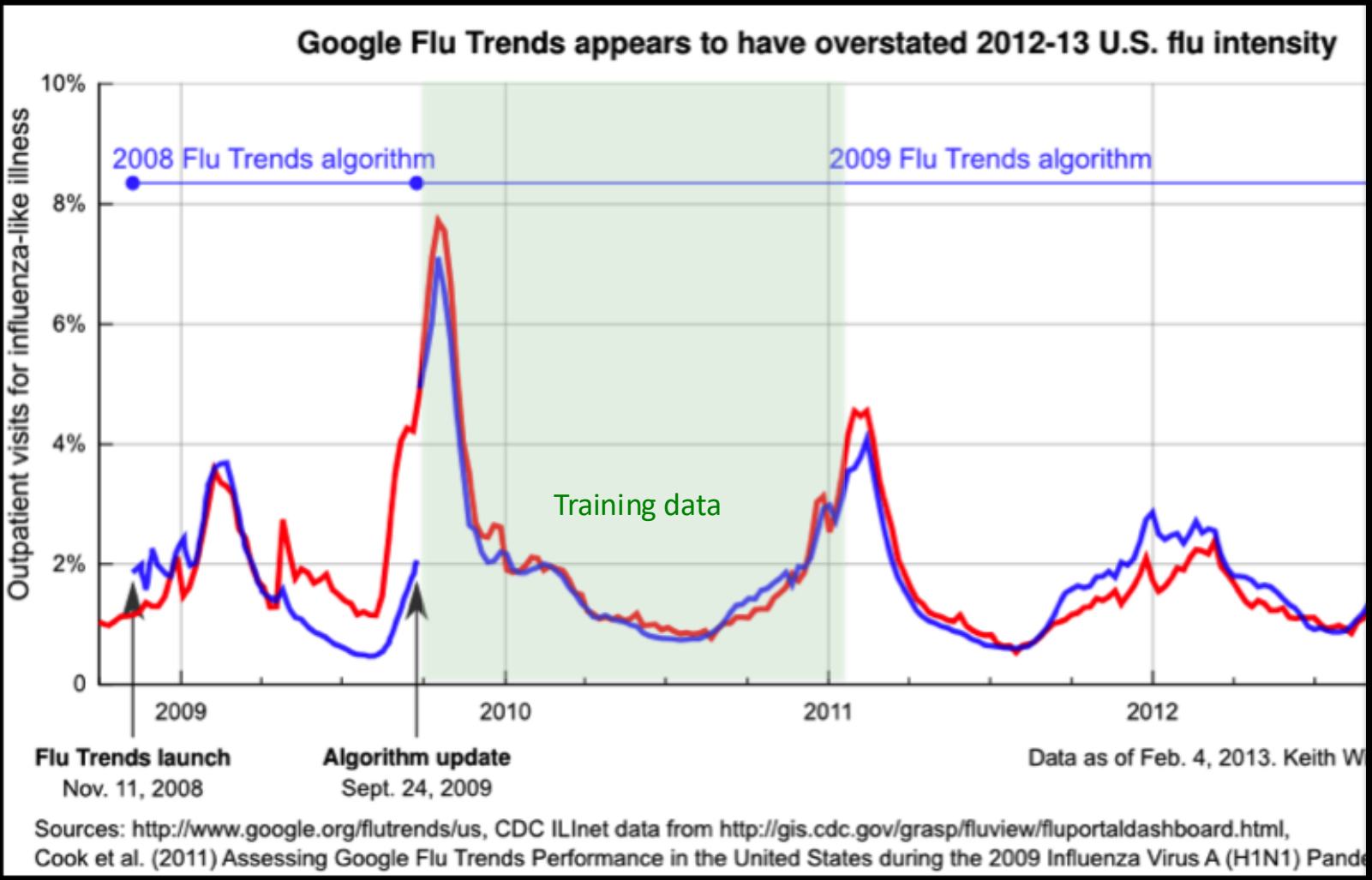


To some extent GFT was good at predicting seasons: fall-winter, not flu!

Plot obtained from:
http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html

What next?

need to remove (not useful) search terms

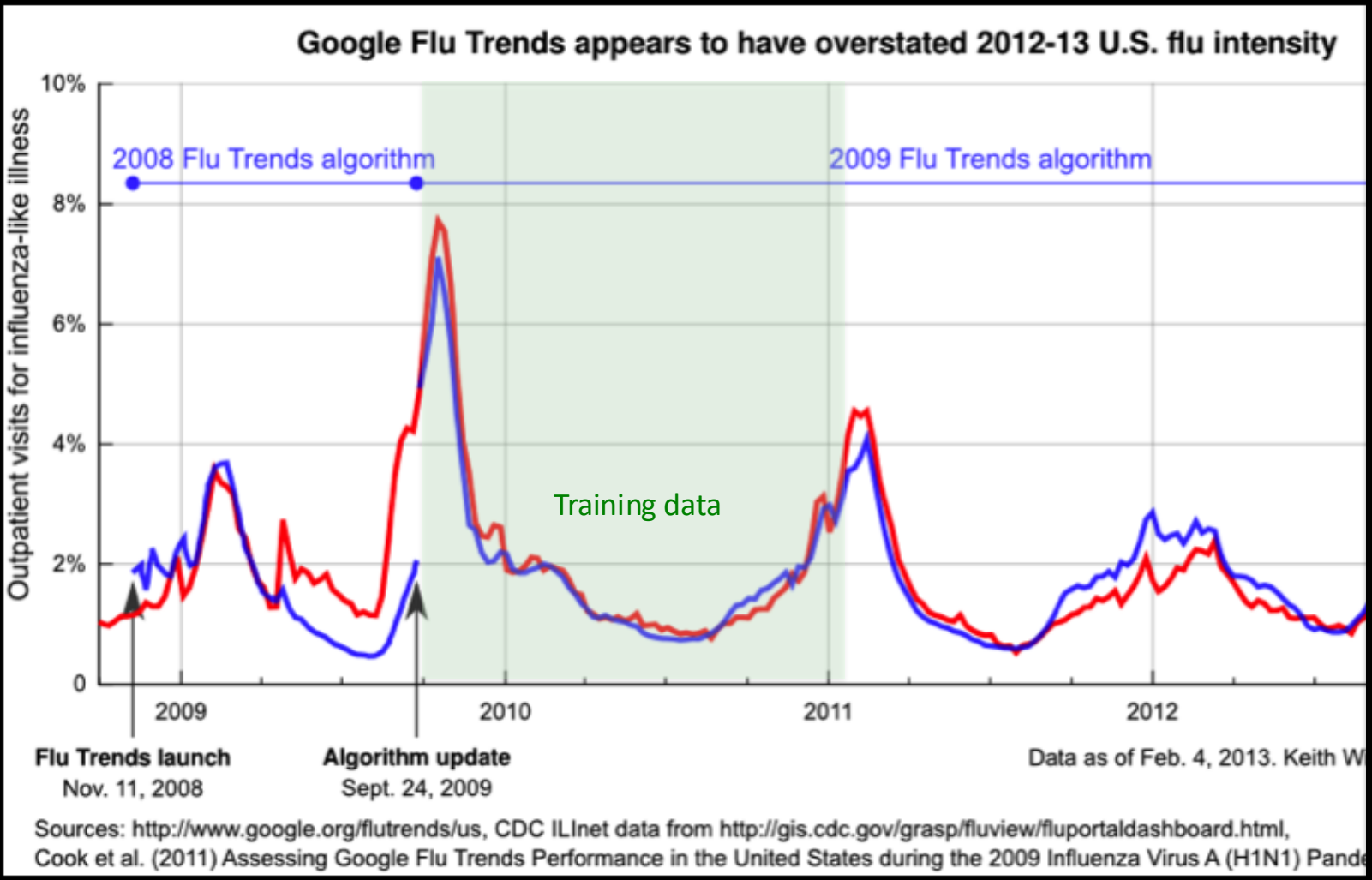


Fixes were reported in: Cook et al. (2011) Assessing Google flu trends performance in the U.S. during the 2009 influenza virus A (H1N1) pandemic. PLoS One

Plot obtained from: http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html

What next?

need to remove (not useful) search terms

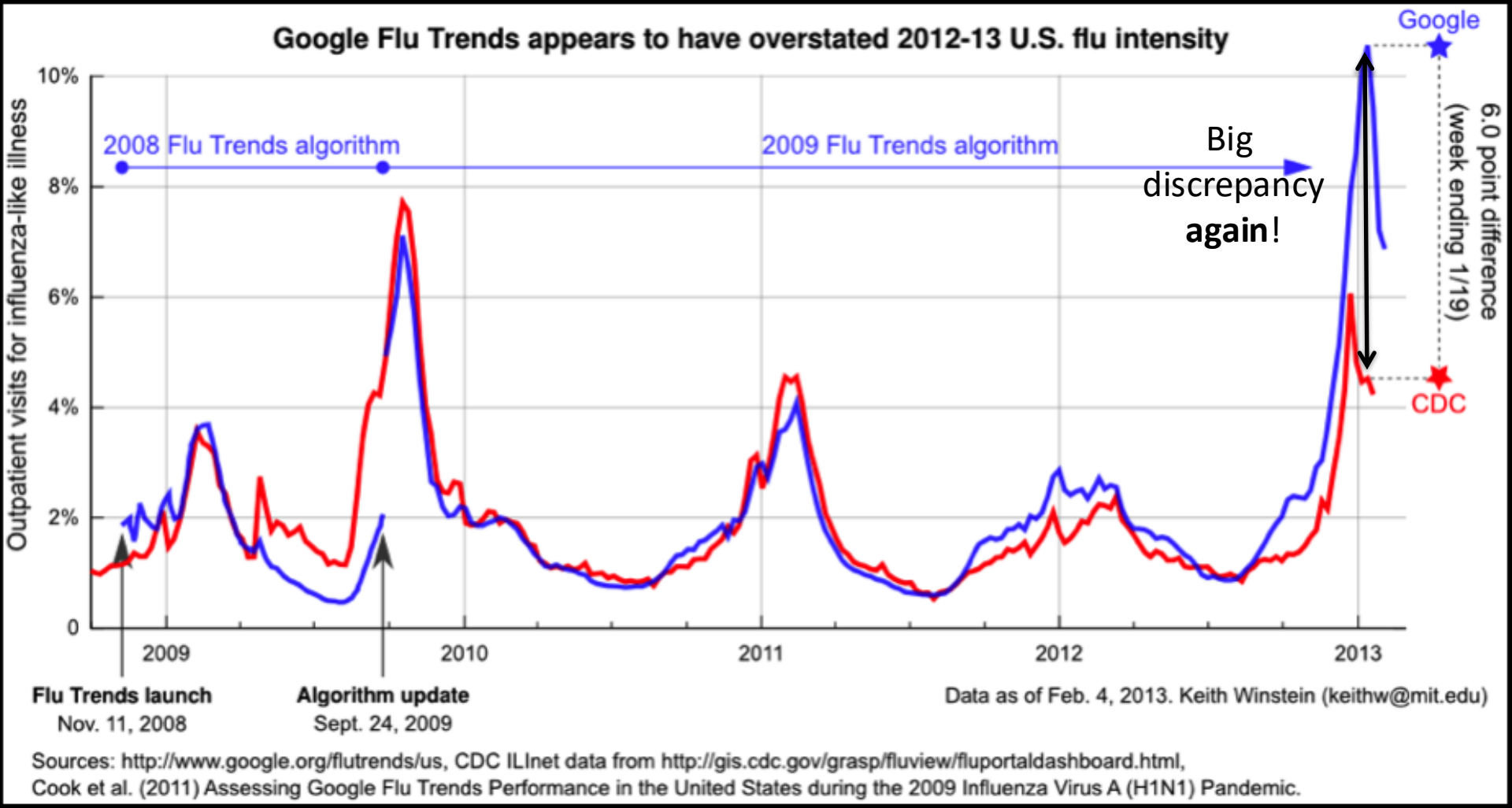


promising performance

Fixes were reported in: Cook et al. (2011) Assessing Google flu trends performance in the U.S. during the 2009 influenza virus A (H1N1) pandemic. PLoS One

Plot obtained from: http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html

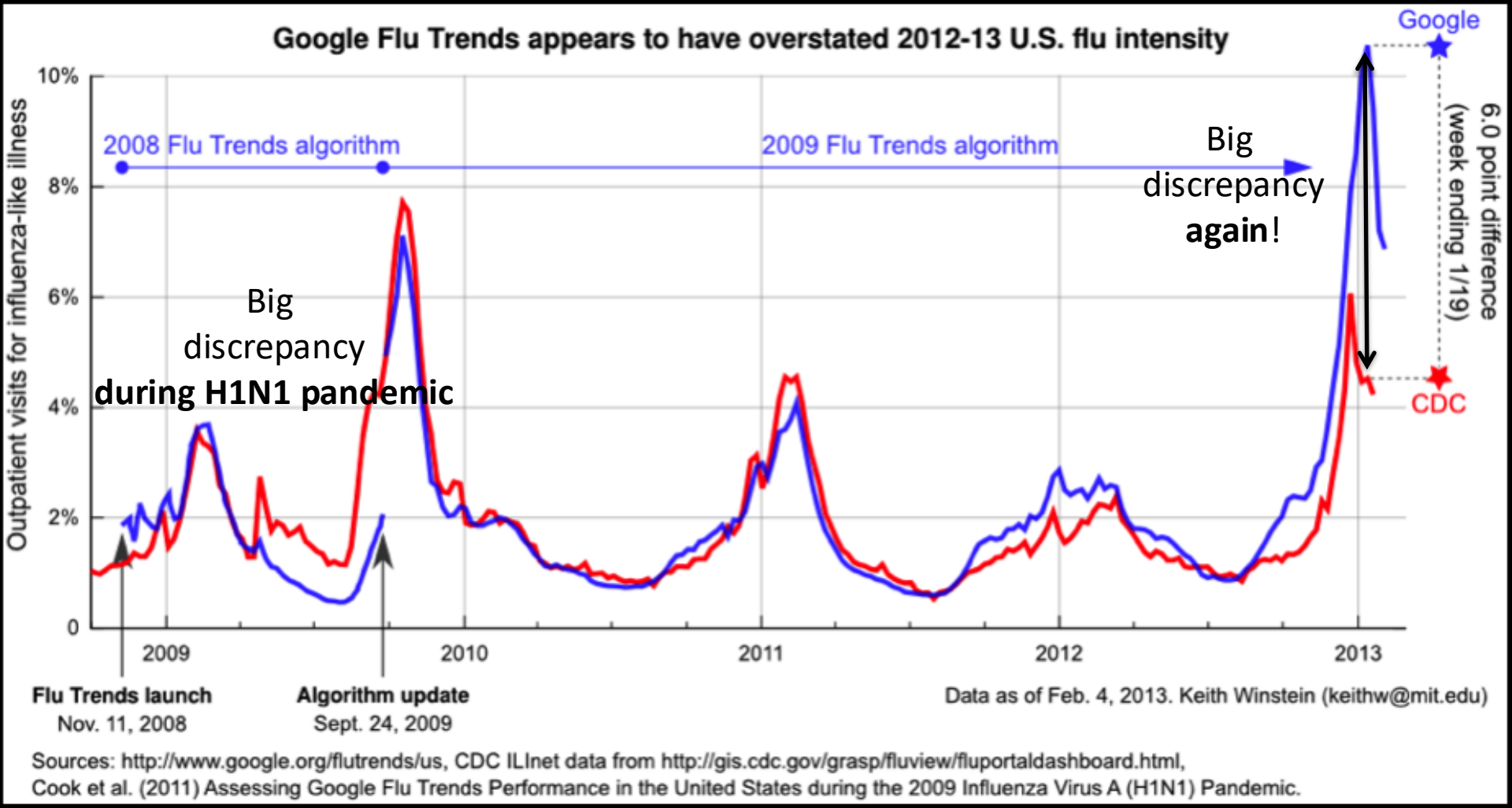
What next? need to remove (not useful) terms. **Big discrepancies again!**



Fixes were reported in: Cook et al. (2011) Assessing Google flu trends performance in the U.S. during the 2009 influenza virus A (H1N1) pandemic. PLoS One

Plot obtained from: http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html

What next? need to remove (not useful) terms. **Big discrepancies again!**



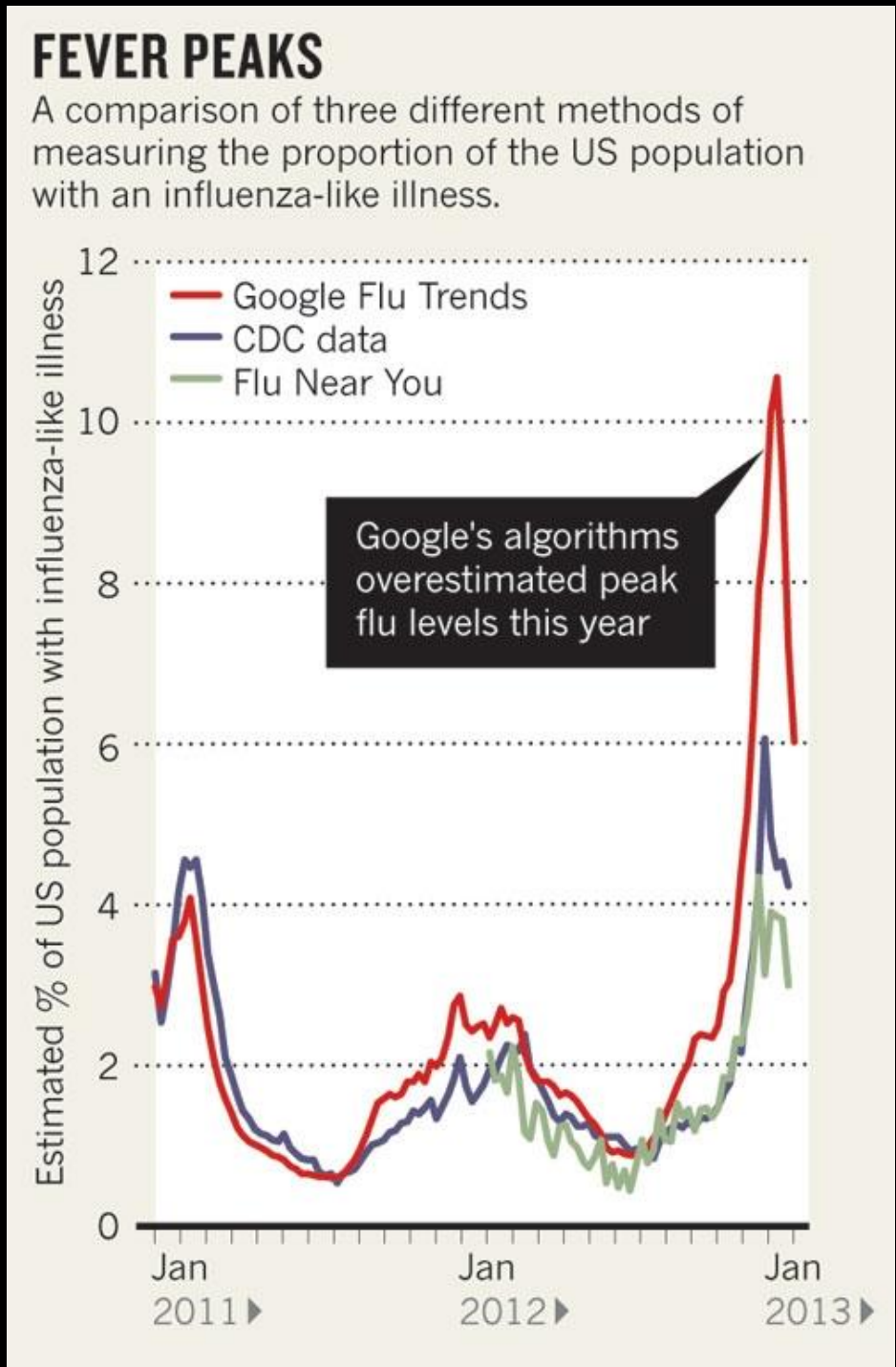
Fixes were reported in: Cook et al. (2011) Assessing Google flu trends performance in the U.S. during the 2009 influenza virus A (H1N1) pandemic. PLoS One

Plot obtained from: http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html

nature International weekly journal of science

When Google got flu wrong.

nature.com/news/when-google-got-flu-wrong.



Google Flu Trends heavily criticized in a paper published by Alex's research team

BIG DATA

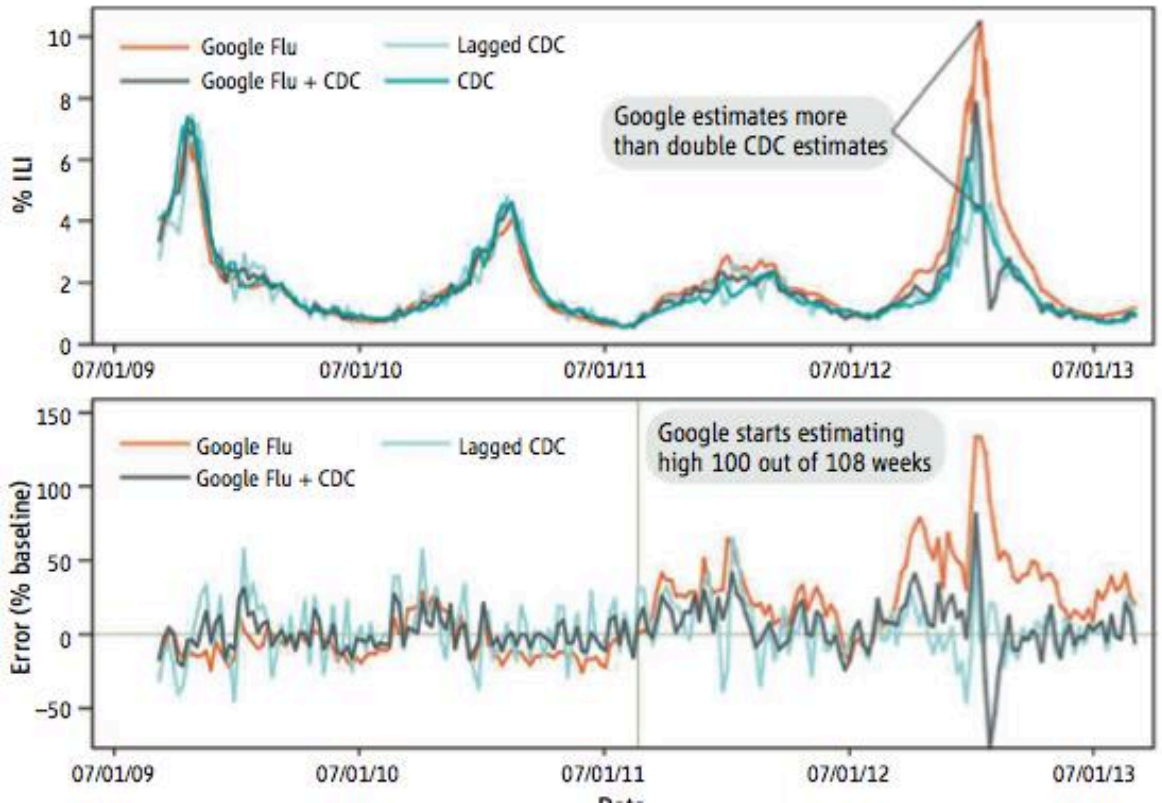
The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

www.sciencemag.org SCIENCE VOL 343 14 MARCH 2014

Published by AAAS



1. Lagged (CDC-based) models capable of outperforming GFT.

2. GFT + lagged CDC can outperform GFT (recalibrating importance of GFT)

3. Google search engine itself changed 86 times in June and July 2012 potentially leading to changes in Google search results (independent variable)

4. Feedbacks (recommended search terms depend on previous searches)

Forbes

New Posts
+5 posts this hour

Most Popular
Forbes 400 Newcomers

Lists
The Forbes 400

63
Share

106
Tweet

61

Snowden And The Challenge Of Intelligence: The Practical Case Against The NSA's Big Data

12 comments, 7 called-out + Comment Now + Follow Comments

We should soon be able to keep track of most activities on the surface of the earth, day or night, in good weather or bad.

siliconANGLE

where computer science meets social science

{SILICON ANGLE}

CLOUD

MOBILE

SOCIAL

SERVICES

DEVOPS

RESEARCH

SiliconANGLE » Can Nate Silver's Data Culture Lead Us Out Of The NSA + Public Data Scare?

Can Nate Silver's Data Culture Lead Us Out of the NSA + Public Data Scare?

RYAN COX | SEPTEMBER 18TH

READ MORE

PLOS COMPUTATIONAL BIOLOGY

Browse Publish About

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Advances in using Internet searches to track dengue

Shihao Yang, Samuel C. Kou, Fred Lu, John S. Brownstein, Nicholas Brooke, Mauricio Santillana

Published: July 20, 2017 • https://doi.org/10.1371/journal.pcbi.1005607

Abstract

Dengue is a mosquito-borne disease that threatens over half of the world's population. Despite being endemic to more than 100 countries, government-led efforts and tools for timely identification and tracking of new infections are still lacking in many affected areas. Multiple methodologies that leverage the use of Internet-based data sources have been proposed as a

Lessons learned

Assumptions in Google Flu Trends:

1. Number of (influenza-like) ill people proportional to number of **total** searches of (Influenza-like illnesses) related terms

$$logit(P) = \beta_0 + \beta_1 \times logit(Q) + \varepsilon$$

where P is the percentage of ILI physician visits, Q is the ILI-related query fraction, β_0 is the intercept,

Assumptions in Google Flu Trends:

1. Number of (influenza-like) ill people proportional to number of **total** searches of (Influenza-like illnesses) related terms

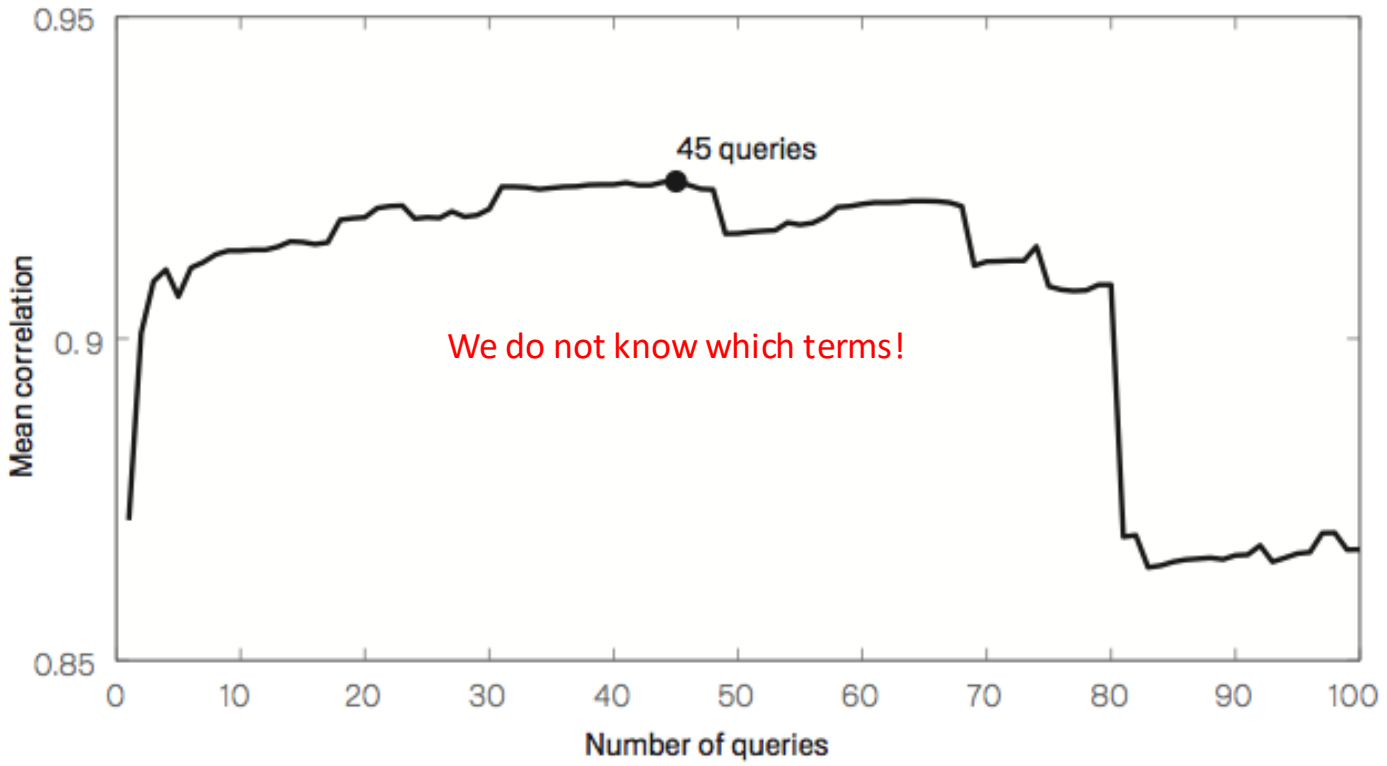


Figure 1: An evaluation of how many top-scoring queries to include in the ILI-related query fraction. Maximal performance at estimating out-of-sample points during cross-validation was obtained by summing the top 45 search queries. A steep drop in model performance occurs after adding query 81, which is "oscar nominations".

Assumptions in Google Flu Trends:

2. Relationship between search volume and proportion of (influenza like) ill people is **static** (during a given year).

Assumptions in Google Flu Trends:

2. Relationship between search volume and proportion of (influenza like) ill people is **static** (during a given year).

Consequences: Model needed constant supervision by human experts

a. **Human experts** needed to **assess** relevance of individual search terms,

b. **Human Experts** needed to **recalculate** relationship between total number of searches and ill people, and

c. It is bound to **deliver poor predictions** at some point in the near future!

We proposed an alternative method and tested it using low quality input from Google Correlate in January 2013.

(with D. Wendong Zhang)

New model:

1. Each search term may contribute to prediction of ILI rate separately **(multi-variate approach)**

2. Relationship between search volume for each individual term and proportion of ill people is **dynamic** and should be found using supervised machine learning optimization techniques.

$$\boldsymbol{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{M} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{M} |\beta_j| \right\}$$

Every week the multiplicative coefficients (β's) would be automatically updated by expanding the training set (labeled data) as new information from the CDC became available.

Top correlated terms to CDC-reported data from 1/2004- 3/2009 (using Google Correlate)

1	influenza type a	35	is the flu contagious	68	fever in adults
2	bronchitis	36	flu in children	69	decongestant
3	influenza a	37	fever flu	70	normal body
4	symptoms of pneumonia	38	take action tour	71	low body temperature
5	flu incubation	39	flu remedies	72	a fever
6	influenza incubation	40	flu report	73	influenza a symptoms
7	flu contagious	41	nasal congestion	74	dangerous fever
8	influenza contagious	42	fever reducer	75	is flu contagious
9	flu incubation period	43	sinus infections	76	lauderdale florida
10	tussionex	44	rhode island wrestling	77	hotel fort lauderdale
11	benzonatate	45	symptoms of influenza	78	webmail shaw ca
12	influenza symptoms	46	castaway bay	79	high fever
13	a influenza	47	coral by the sea	80	robitussin ac
14	sinus	48	cold or flu	81	bronchitis contagious
15	pneumonia	49	respiratory infection	82	indoor driving
16	flu fever	50	take action	83	tussionex pennkinetic
17	flu duration	51	respiratory flu	84	wrestling report
18	taste of chaos	52	soweto gospel	85	walking pneumonia
19	bronchitis symptoms	53	soweto gospel choir	86	days inn miami
20	symptoms of bronchitis	54	illinois wrestling	87	body temperature
21	how long does the flu last	55	how long is the flu contagious	88	phlegm
22	symptoms of the flu	56	cold symptoms	89	flu relief
23	taste of chaos tour	57	the taste of chaos	90	mt sunapee
24	influenza incubation period	58	is bronchitis	91	harlem globe
25	sinus infection	59	upper respiratory	92	levaquin
26	flu recovery	60	afrin	93	strep throat
27	chaos tour	61	painful cough	94	coughing
28	type a influenza	62	laprepsoccer	95	whistler snow
29	flu symptoms	63	upper respiratory infection	96	fever temperature
30	tessalon	64	amoxicillin	97	sales tax credit
31	type a flu	65	ski harness	98	glitches
32	treat the flu	66	robitussin dm	99	pennkinetic
33	treating the flu	67	treating flu	100	histinex
34	how to treat the flu				

AMERICAN JOURNAL OF
Preventive Medicine
A Journal of the American College of Preventive Medicine and Association for Prevention Teaching and Research

What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?

Mauricio Santillana, PhD, MS, D. Wendong Zhang, MA, Benjamin M. Althouse, PhD, ScM, John W. Ayers, PhD, MA

First week after being published online, it became the second most read paper in journal's history! (After a paper published in 1998)

AMERICAN JOURNAL OF
Preventive Medicine
A Journal of the American College of Preventive Medicine and Association for Prevention Teaching and Research

Articles in Press | **Most Read** | Most Cited

Relationship of Childhood Abuse and Household Dysfunction to Many of the Leading Causes of Death in Adults: The Adverse Childhood Experiences (ACE) Study

Vincent J Felitti, Robert F Anda, Dale Nordenberg, David F Williamson, Alison M Spitz, Valerie Edwards, Mary P Koss, James S Marks

Vol. 14, Issue 4

Published in issue: May, 1998

Abstract | Full-Text HTML | PDF

What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?

Mauricio Santillana, D. Wendong Zhang, Benjamin M. Althouse, John W. Ayers

Vol. 47, Issue 3

Published online: July 1, 2014

Abstract | Full-Text HTML | PDF

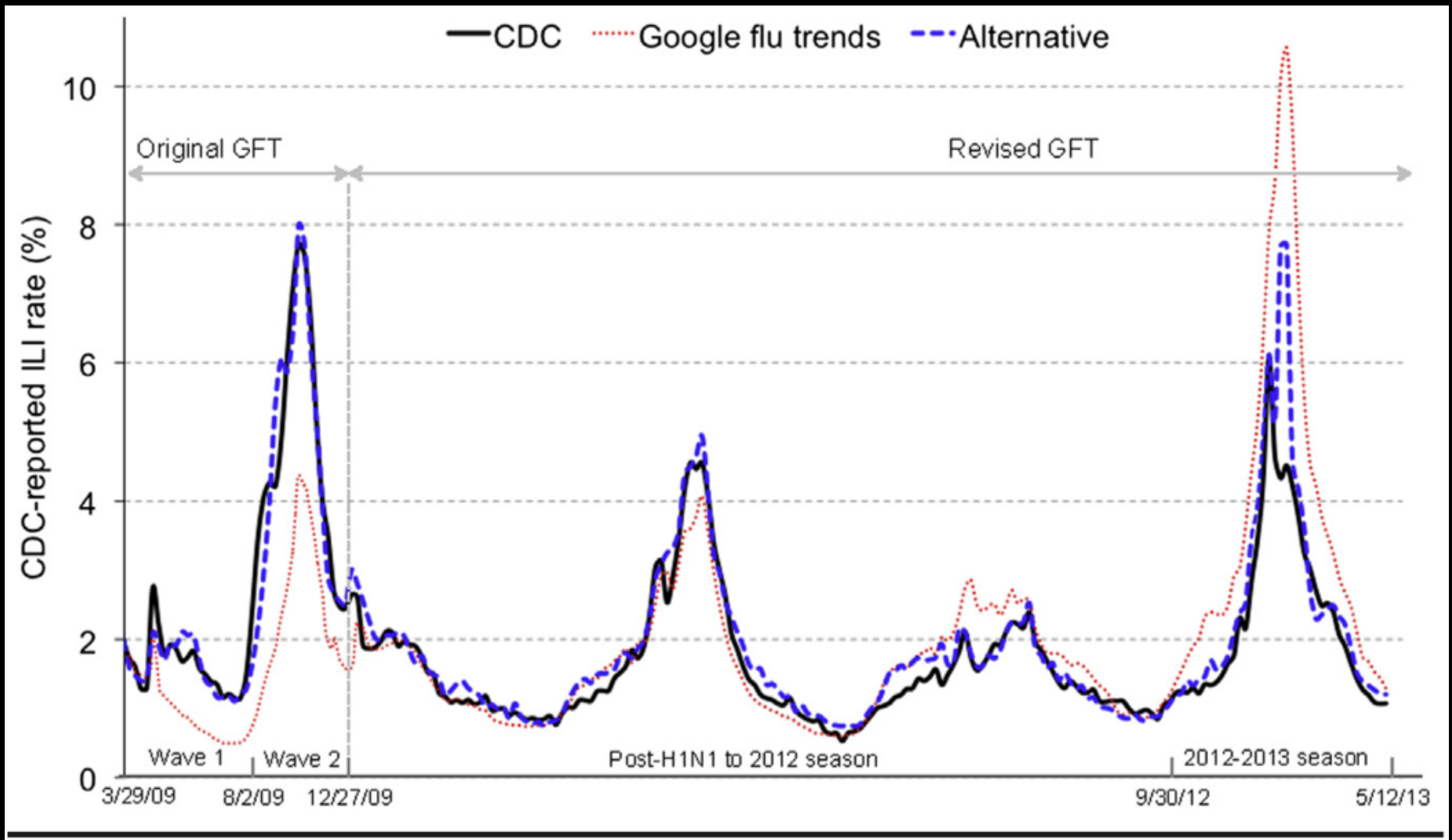
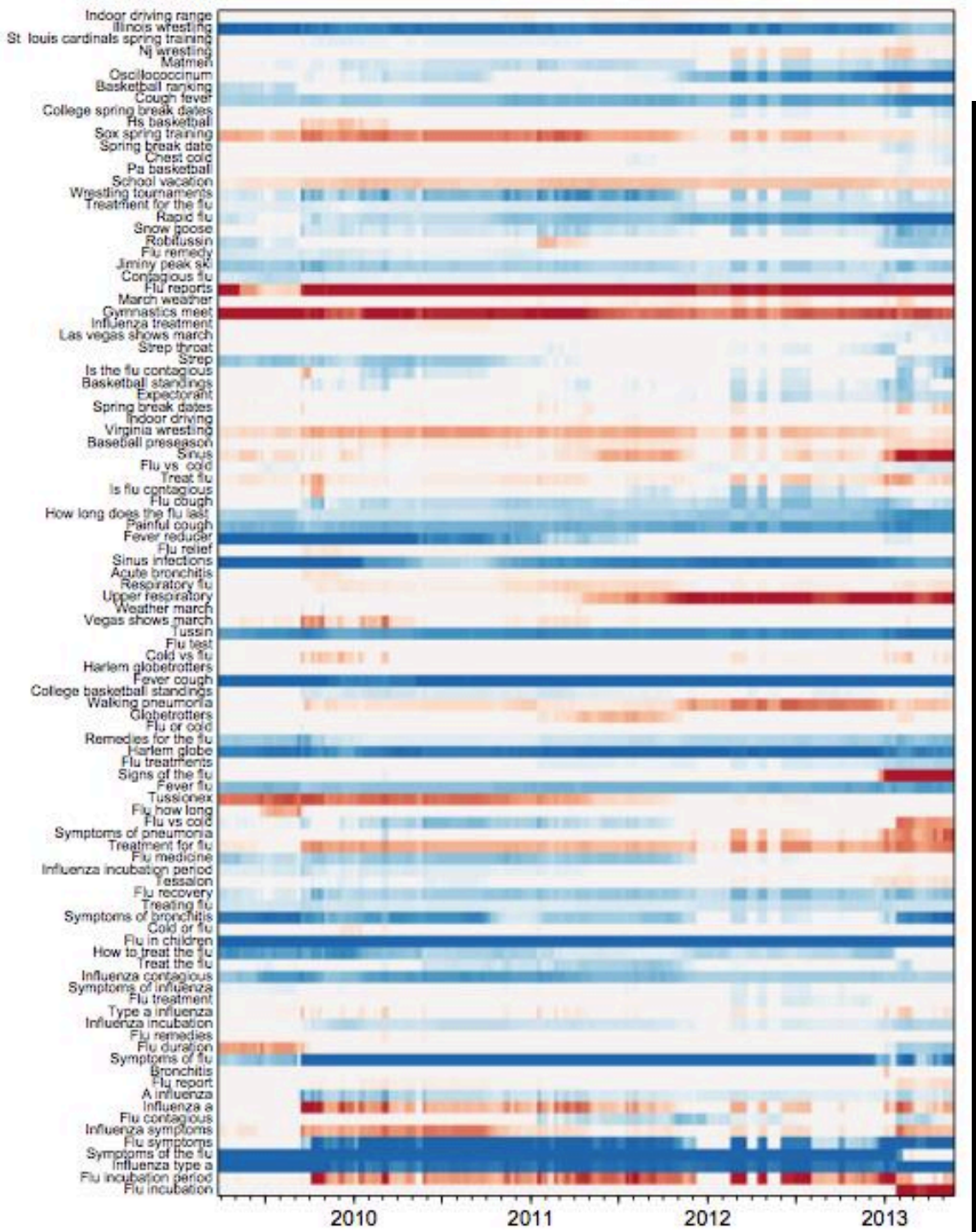
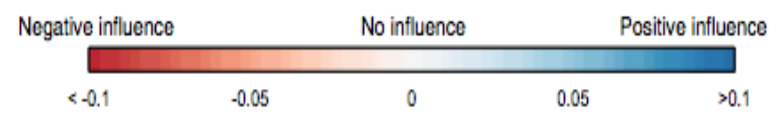


Figure 1. The alternative model outperforms Google Flu Trends

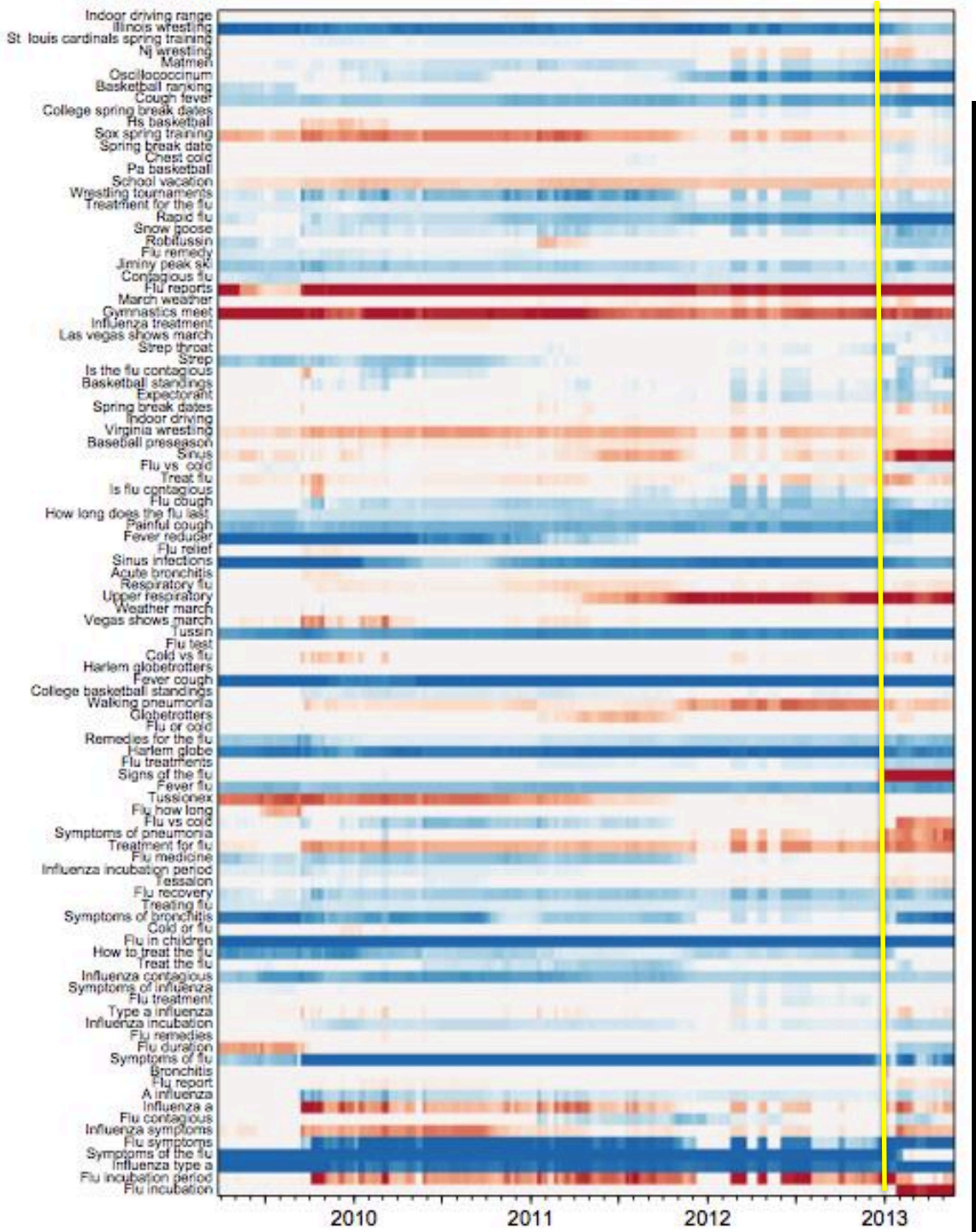
$$\text{logit}[I(t)] = \sum_{i=1}^{n} a_i(t)\text{logit}[Q_i(t))] + \text{e},$$

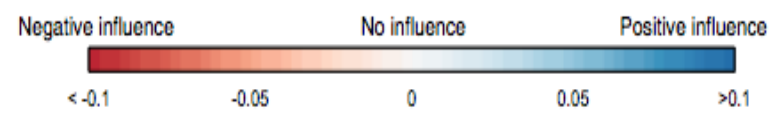
Santillana et al. American Journal of Preventive Medicine, 2014; 47 (3) pp 341-347





Santillana et al. American Journal of Preventive Medicine, 2014; 47 (3) pp 341-347





Santillana et al. American Journal of Preventive Medicine, 2014; 47 (3) pp 341-347

theguardian

Google Flu Trends promises are overstated, researchers say

New study finds way to improve Google Flu Trends accuracy threefold - but says systems must be more open

Charles Arthur

Follow @charlesarthur Follow @guardiantech

theguardian.com, Friday 4 July 2014 11.44 EDT


theguardian

News | US | World | Sports | Comment | Culture | Business | Money

News › Technology › Google

Google Flu Trends promises are overstated, researchers say

New study finds way to improve Google Flu Trends accuracy threefold - but says systems must be more open


HealthData Management

NE

POLICY & REGULATION | EHR | HEALTH INFO EXCHANGE | REVENUE CYCLE & PAYMENTS | CHRONIC CARE

Researchers Suggest Fixes to Google Flu Trends Analytics

A new study concludes that "revising the inner plumbing" of the Google Flu Trends disease surveillance system can improve the accuracy of forecasts for the severity of a flu season.

theguardian

News | US | World | Sports | Con

News › Technology › Google

Google Flu Trends p
overstated, research

New study finds way to improve
threefold - but says systems mu

HealthData Management

POLICY & REGULATION | EHR | HEA EXC

Researchers Sug
Flu Trends Analyt

A new study concludes that "revising the in
surveillance system can improve the accur

ScienceDaily®

Your source for the latest research news

Featured Research *from universities, journals, and other organizatio*

Finding real value in big data for public health

Date: July 2, 2014

Source: San Diego State University

Summary: Media reports of public health breakthroughs from big data have been largely oversold, according to a new study. But don't throw away that data just yet. The authors maintain that the promise of big data can be fulfilled by tweaking existing methodological and reporting standards. In the study, the research team demonstrate this by revising the inner plumbing of the Google Flu Trends (GFT) digital disease surveillance system, which was heavily criticized last year (see here and here) after producing erroneous forecasts.

Share This

- Email to a friend
- Facebook
- Twitter
- LinkedIn
- Google+
- Print this page

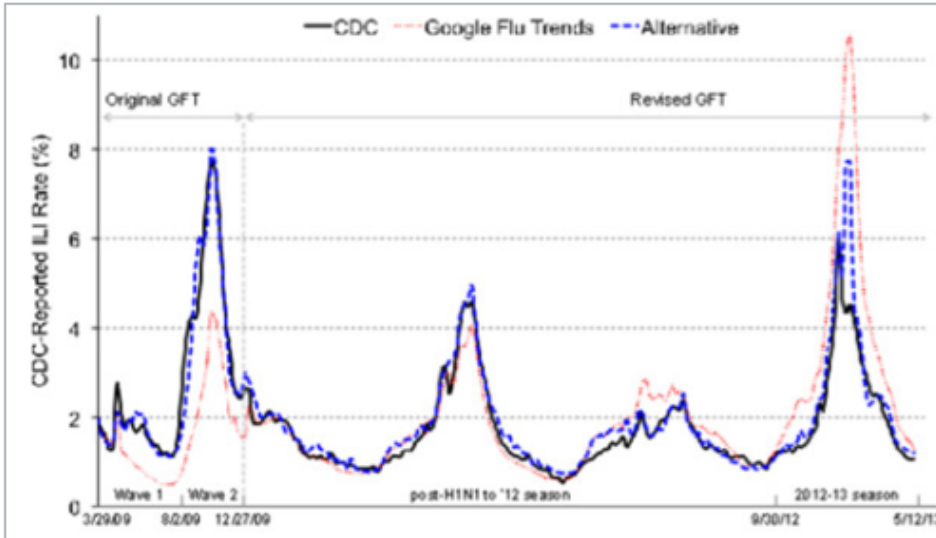
Related Topics

Health & Medicine
- Health Policy
- Public Health Education

Computers & Math
- Computers and Internet
- Computer Modeling

Science & Society
- Public Health
- Surveillance

Related Articles
- Public health
- Data mining



A graph depicting Google Flu Trends.

theguardian

News | US | World | Sports | Con

News › Technology › Google

Google
oversta
New study
threefold - b

Hea
Ma

POLICY & REGULATION

Resea
Flu Tr

A new study concludes that "revising the in
surveillance system can improve the accur

ScienceDaily®

Your source for the latest research news

Featured Research

from universities, journals, and other organizatio

iHealthBeat

Reporting Technology's Impact on Health Care

HOME | INSIGHT | PERSPECTIVES | PICTURE OF HEALTH

NEWS ARCHIVE

SHARE EMAIL PRINT REPUBLISH

Study: Methodology Changes Improve Google Flu Trend Accuracy

Monday, July 7, 2014

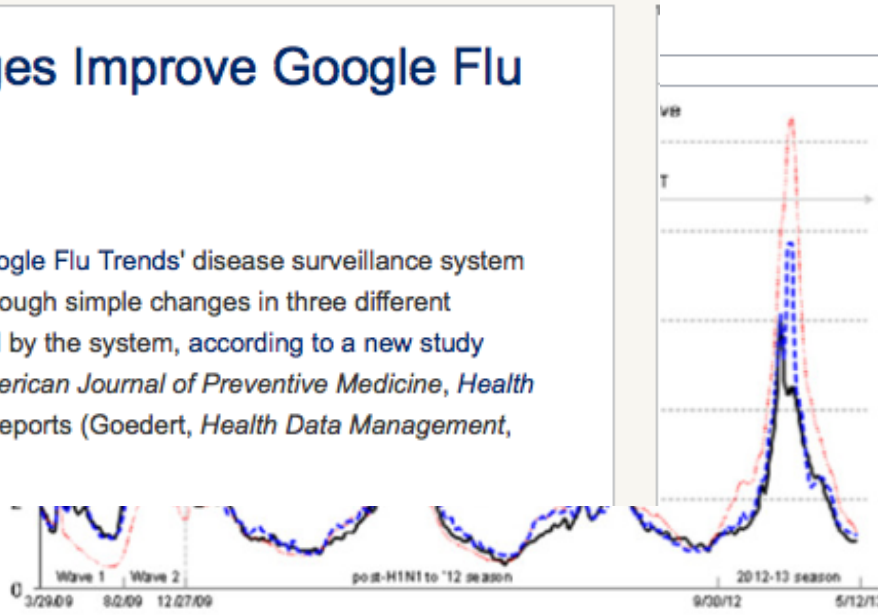
RELATED TOPICS:

- Public Health

The accuracy of Google Flu Trends' disease surveillance system can be improved through simple changes in three different methodologies used by the system, according to a new study published in the *American Journal of Preventive Medicine*, *Health Data Management* reports (Goedert, *Health Data Management*, 7/7).

This

nail to a friend
cebook
vitter
nkedIn
oogle+
int this page



A graph depicting Google Flu Trends.

Related Articles

› Public health
› Data mining

theguardian

ScienceDaily®

datacconomy NEWS EVENTS OPINION START UPS INDUSTRY RESOURCES ABOUT JOBS

BI & ANALYTICS MACHINE LEARNING DATA SCIENCE DATABASE TECHNOLOGY DATA SECURITY UNDERSTANDING BIG DATA

Search

Big Data's Potential in Public Health: Revisiting Google Flu Trends

July 7, 2014 Written by: Dan Gray 1 Reply

Resea
Flu Tr

A new study concludes that "revising the in
surveillance system can improve the accur

RELATED TOPICS:

- Public Health

can be improved through simple changes in three different methodologies used by the system, according to a new study published in the *American Journal of Preventive Medicine*, *Health Data Management* reports (Goedert, *Health Data Management*, 7/7).

Related Articles

- Public health
- Data mining

A graph depicting Google Flu Trends.

This

nail to a friend
icebook
witter
nkedIn
oogle+
int this page

nd other organizatio

Google Research Blog

The latest news from Research at Google

Google Flu Trends gets a brand new engine

Posted: Friday, October 31, 2014

Posted by Christian Stefansen, Senior Software Engineer

Each year the flu kills thousands of people and affects millions around the world. So it's important that public health officials and health professionals learn about outbreaks as quickly as possible. In 2008 we launched Google Flu Trends in the U.S., using aggregate web searches to indicate when and where influenza was striking in real time. These models nicely complement other survey systems—they're more fine-grained geographically, and they're typically more immediate, up to 1-2 weeks ahead of traditional methods such as the CDC's official reports. They can also be incredibly helpful for countries that don't have official flu tracking. Since launching, we've expanded Flu Trends to cover 29 countries, and launched Dengue Trends in 10 countries.

The original model performed surprisingly well despite its simplicity. It was retrained just once per year, and typically used only the 50 to 300 queries that produced the best estimates for prior seasons. We then left it to perform through the new season and evaluated it at the end. It didn't use the official CDC data for estimation during the season—only in the initial training.

AMERICAN JOURNAL OF
Preventive Medicine
A Journal of the American College of Preventive Medicine and Association for Prevention Teaching and Research

What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?

Mauricio Santillana, PhD, MS, D. Wendong Zhang, MA, Benjamin M. Althouse, PhD, ScM, John W. Ayers, PhD, MA

 Am J Prev Med 2014;47(3):341–347

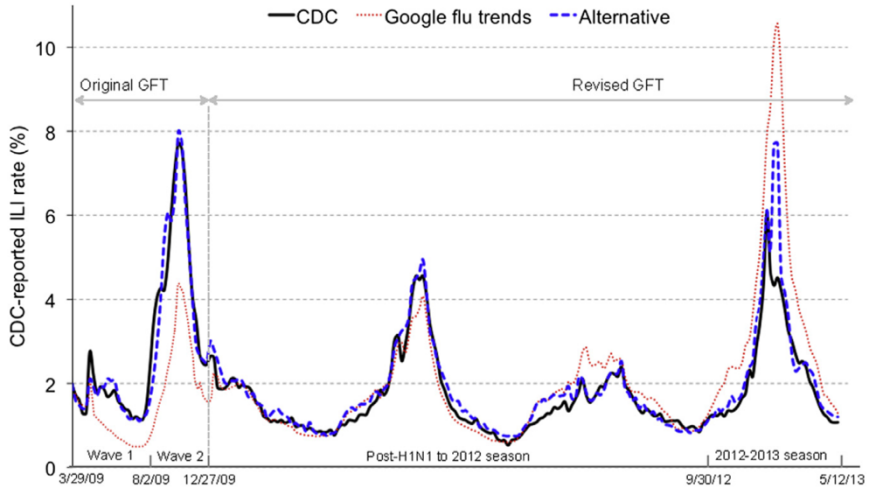


Figure 1. The alternative model outperforms Google Flu Trends

We published a paper proposing changes to GFT's engine (2014)

Google incorporated our proposed changes to GFT's engine in Oct 2014

Google Research Blog

The latest news from Research at Google

Google Flu Trends gets a brand new engine

Posted: Friday, October 31, 2014

Posted by Christian Stefansen, Senior Software Engineer

Each year the flu kills thousands of people and affects millions around the world. So it's important that public health officials and health professionals learn about outbreaks as quickly as possible. In 2008 we launched Google Flu Trends in the U.S., using aggregate web searches to indicate when and where influenza was striking in real time. These models nicely complement other survey systems—they're more fine-grained geographically, and they're typically more immediate, up to 1-2 weeks ahead of traditional methods such as the CDC's official reports. They can also be incredibly helpful for countries that don't have official flu tracking. Since launching, we've expanded Flu Trends to cover 29 countries, and launched Dengue Trends in 10 countries.

The original model performed surprisingly well despite its simplicity. It was retrained just once per year, and typically used only the 50 to 300 queries that produced the best estimates for prior seasons. We then left it to perform through the new season and evaluated it at the end. It didn't use the official CDC data for estimation during the season—only in the initial training.

SCIENTIFIC REPORTS

PDF

Article | OPEN

Advances in nowcasting influenza-like illness rates using search query logs

Vasileios Lampos ✉, Andrew C. Miller, Steve Crossan & Christian Stefansen

Scientific Reports **5**,
Article number: 12760 (2015)
doi:10.1038/srep12760
Download Citation

Applied mathematics
Computer science | Epidemiology
Influenza virus

Received: 07 May 2015
Accepted: 06 July 2015
Published online: 03 August 2015

Google and collaborators published a paper improving our AJPM 2014 methodology in August 2015

We improved last effort by Google team and published our results in PNAS in September 2015

PNAS

CrossMark click for updates

Accurate estimation of influenza epidemics using Google search data via ARGO

Shihao Yang[a], Mauricio Santillana[b,c,1], and S. C. Kou[a,1]

[a]Department of Statistics, Harvard University, Cambridge, MA 02138; [b]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; and [c]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA 02115

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved September 30, 2015 (received for review August 6, 2015)

Accurate real-time tracking of influenza outbreaks helps public health officials make timely and meaningful decisions that could save lives. We propose an influenza tracking model, ARGO (AutoRegression with GOogle search data), that uses publicly available online search data. In addition to having a rigorous statistical foundation, ARGO outperforms all previously available Google-search–based tracking models, including the latest version of Google Flu Trends, even though it uses only low-quality search data as input from publicly available Google Trends and Google Correlate websites. ARGO not only incorporates the seasonality in influenza epidemics but also captures changes in people's online search behavior over time. ARGO is also flexible, self-correcting, robust, and scalable, making it a potentially powerful tool that can be used for real-time tracking of other social events at multiple temporal and spatial resolutions.

CDC's ILI reports have a delay of 1–3wk due to the time for processing and aggregating clinical information. This time lag is far from optimal for decision-making purposes. To alleviate this information gap, multiple methods combining climate, demographic, and epidemiological data with mathematical models have been proposed for real-time estimation of flu activity (18, 21–25). In recent years, methods that harness Internet-based information have also been proposed, such as Google (1), Yahoo (2), and Baidu (3) Internet searches, Twitter posts (4), Wikipedia article views (5), clinicians' queries (6), and crowdsourced self-reporting mobile apps such as Influenzanet (Europe) (26), Flutracking (Australia) (27), and Flu Near You (United States) (28). Among them, GFT has received the most attention and has inspired subsequent digital disease detection systems (3, 8,

PPLIED HEMATICS

SCIENTIFIC REPORTS

PDF

Article | OPEN

Advances in nowcasting influenza-like illness rates using search query logs

Vasileios Lampos ✉, Andrew C. Miller, Steve Crossan & Christian Stefansen

Scientific Reports **5**,
Article number: 12760 (2015)
doi:10.1038/srep12760
Download Citation

Applied mathematics
Computer science | Epidemiology
Influenza virus

Received: 07 May 2015
Accepted: 06 July 2015
Published online: 03 August 2015

Google and collaborators published a paper improving our AJPM 2014 methodology in August 2015

We improved last effort by Google team and published our results in PNAS in September 2015

PNAS

CrossMark click for updates

Accurate estimation of influenza epidemics using Google search data via ARGO

Shihao Yang[a], Mauricio Santillana[b,c,1], and S. C. Kou[a,1]

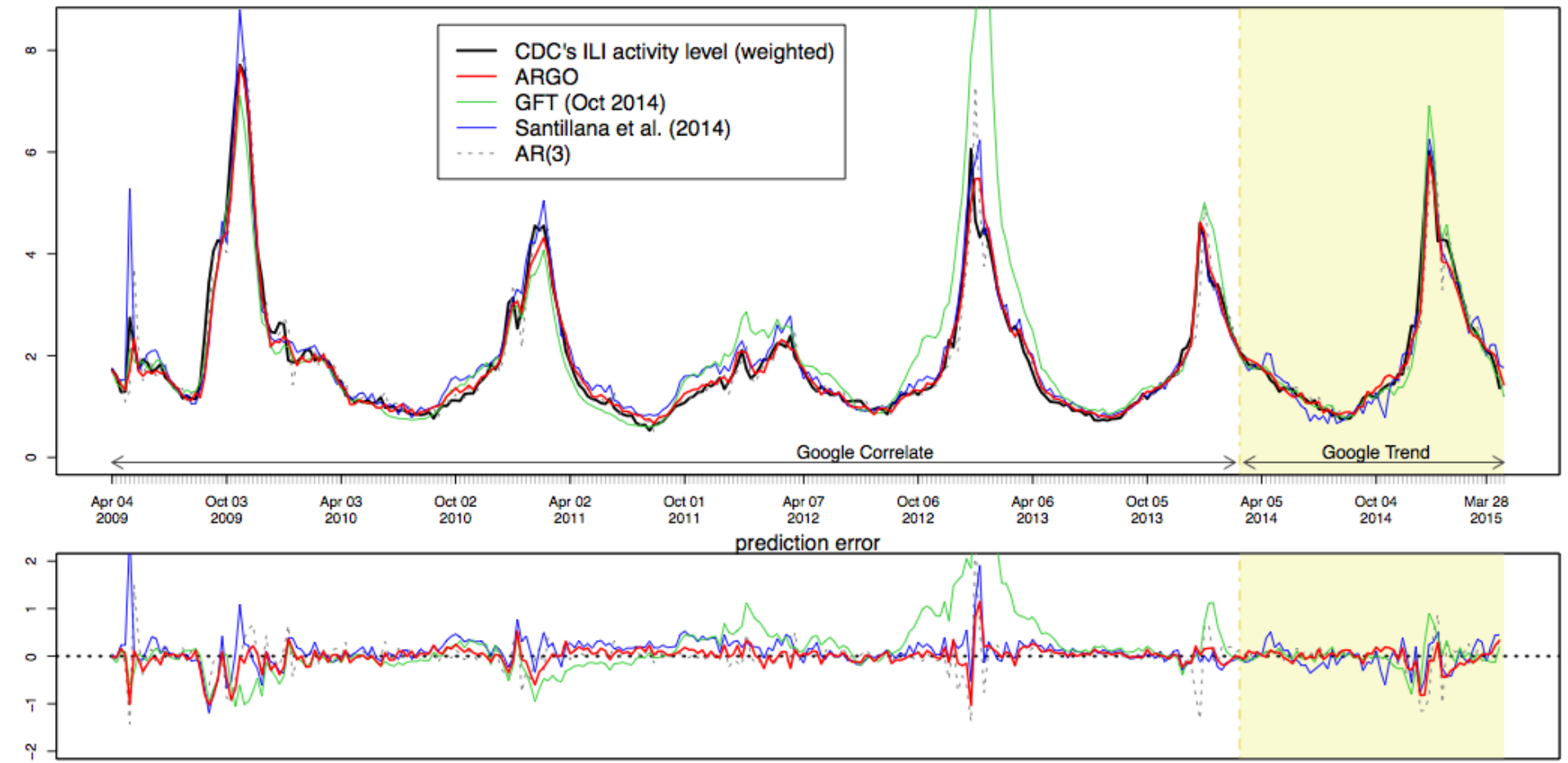
[a]Department of Statistics, Harvard University, Cambridge, MA 02138; [b]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; and [c]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA 02115

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved September 30, 2015 (received for review August 6, 2015)

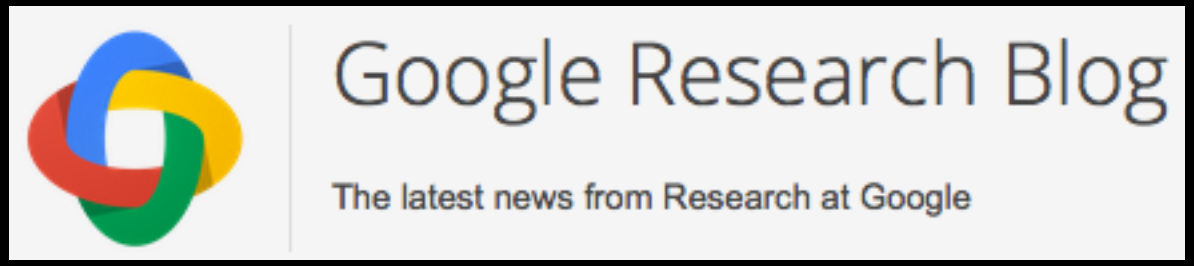
Accurate real-time tracking of influenza outbreaks helps public health officials make timely and meaningful decisions that could save lives. We propose an influenza tracking model, ARGO (AutoRegression with GOogle search data), that uses publicly available online search data. In addition to having a rigorous statistical foundation, ARGO outperforms all previously available Google-search–based tracking models, including the latest version of Google Flu Trends, even though it uses only low-quality search data as input from publicly available Google Trends and Google Correlate websites. ARGO not only incorporates the seasonality in influenza epidemics but also captures changes in people's online search behavior over time. ARGO is also flexible, self-correcting, robust, and scalable, making it a potentially powerful tool that can be used for real-time tracking of other social events at multiple temporal and spatial resolutions.

CDC's ILI reports have a delay of 1–3wk due to the time for processing and aggregating clinical information. This time lag is far from optimal for decision-making purposes. To alleviate this information gap, multiple methods combining climate, demographic, and epidemiological data with mathematical models have been proposed for real-time estimation of flu activity (18, 21–25). In recent years, methods that harness Internet-based information have also been proposed, such as Google (1), Yahoo (2), and Baidu (3) Internet searches, Twitter posts (4), Wikipedia article views (5), clinicians' queries (6), and crowdsourced self-reporting mobile apps such as Influenzanet (Europe) (26), Flutracking (Australia) (27), and Flu Near You (United States) (28). Among them, GFT has received the most attention and has inspired subsequent digital disease detection systems (3, 8,

PPLIED HEMATICS


CDC's ILI activity level (weighted)
ARGO
GFT (Oct 2014)
Santillana et al. (2014)
AR(3)
Google Correlate
Google Trend
Apr 04 2009
Oct 03 2009
Apr 03 2010
Oct 02 2010
Apr 02 2011
Oct 01 2011
Apr 07 2012
Oct 06 2012
Apr 06 2013
Oct 05 2013
Apr 05 2014
Oct 04 2014
Mar 28 2015
prediction error

Google discontinues Flu Trends indefinitely!



The Next Chapter for Flu Trends

Posted: Thursday, August 20, 2015

g+1 7

Instead of maintaining our own website going forward, we're now going to empower institutions who specialize in infectious disease research to use the data to build their own models. Starting this season, we'll provide Flu and Dengue signal data directly to partners including Columbia University's Mailman School of Public Health (to update their dashboard), Boston Children's Hospital/Harvard, and Centers for Disease Control and Prevention (CDC) Influenza Division. We will also continue to make historical Flu and Dengue estimate data available for anyone to see and analyze.

CIO

Most read:

NEWS

Google Flu Trends calls out sick, indefinitely

Google will pass along search queries related to the flu to health organizations so they can develop their own prediction models

By Fred O'Connor | Follow
IDG News Service | Aug 20, 2015 2:07 PM PT

MORE LIKE THIS ::

Google Begins Tracking Swine Flu in Mexico

Google's Panicky Flu Estimates Were Dead Wrong

VB

NEWS EVENTS RESEARCH SUBSCRIBE Sign up | Login

BIG DATA

Google discontinues Flu Trends, starts offering data to researchers

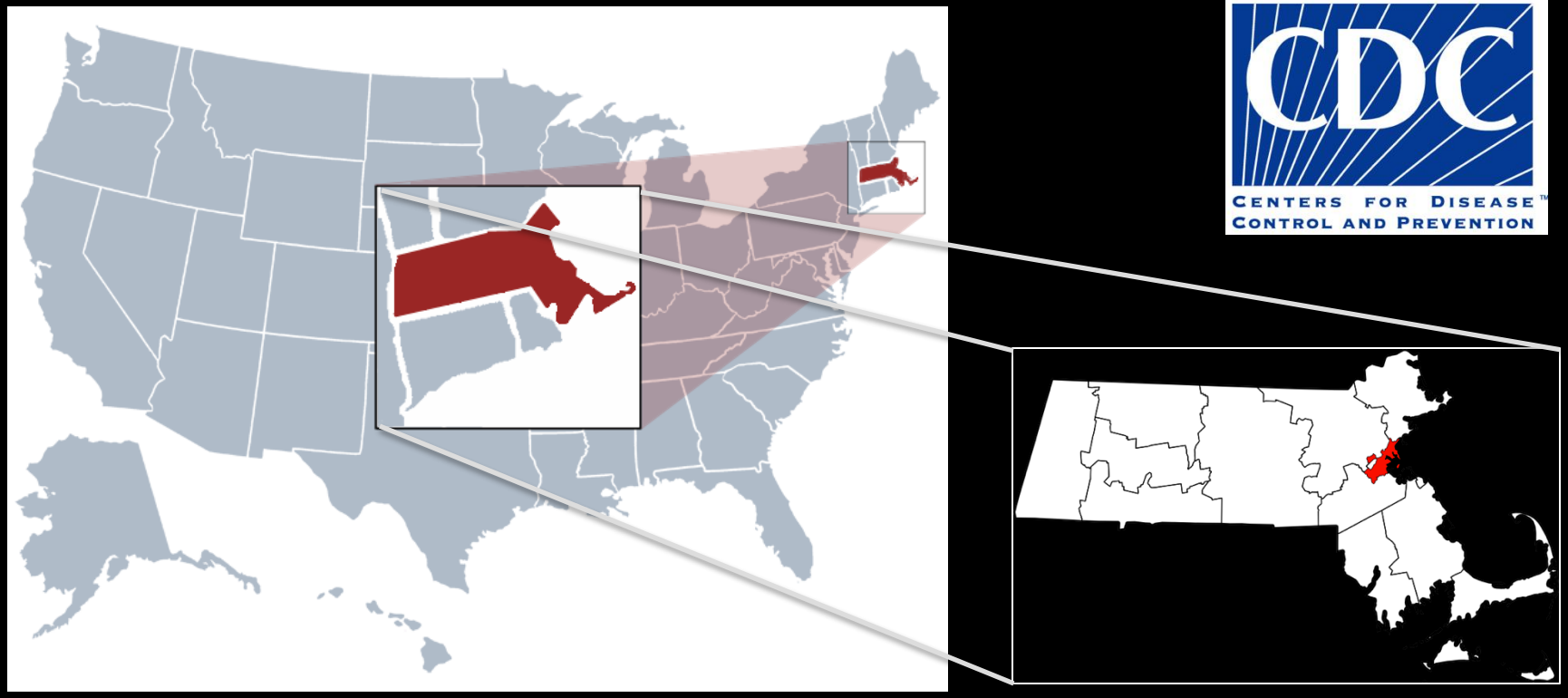
JORDAN NOVET AUGUST 20, 2015 12:17 PM

TAGS: GOOGLE, GOOGLE FLU TRENDS

In collaboration with the CDC Influenza division, we are extending our work from National and Regional predictions, to state-level and city level (Boston as a pilot)

Grant: *Centers for Disease Control and Prevention's Cooperative Agreement PPHF 11797-998G-15*

Team members: *Fred Lu, Leonardo C. Clemente*
CDC liaison and collaborator: *Matt Biggerstaff*



Prepared by Mauricio Santillana

Session 3 started here

PLOS ONE

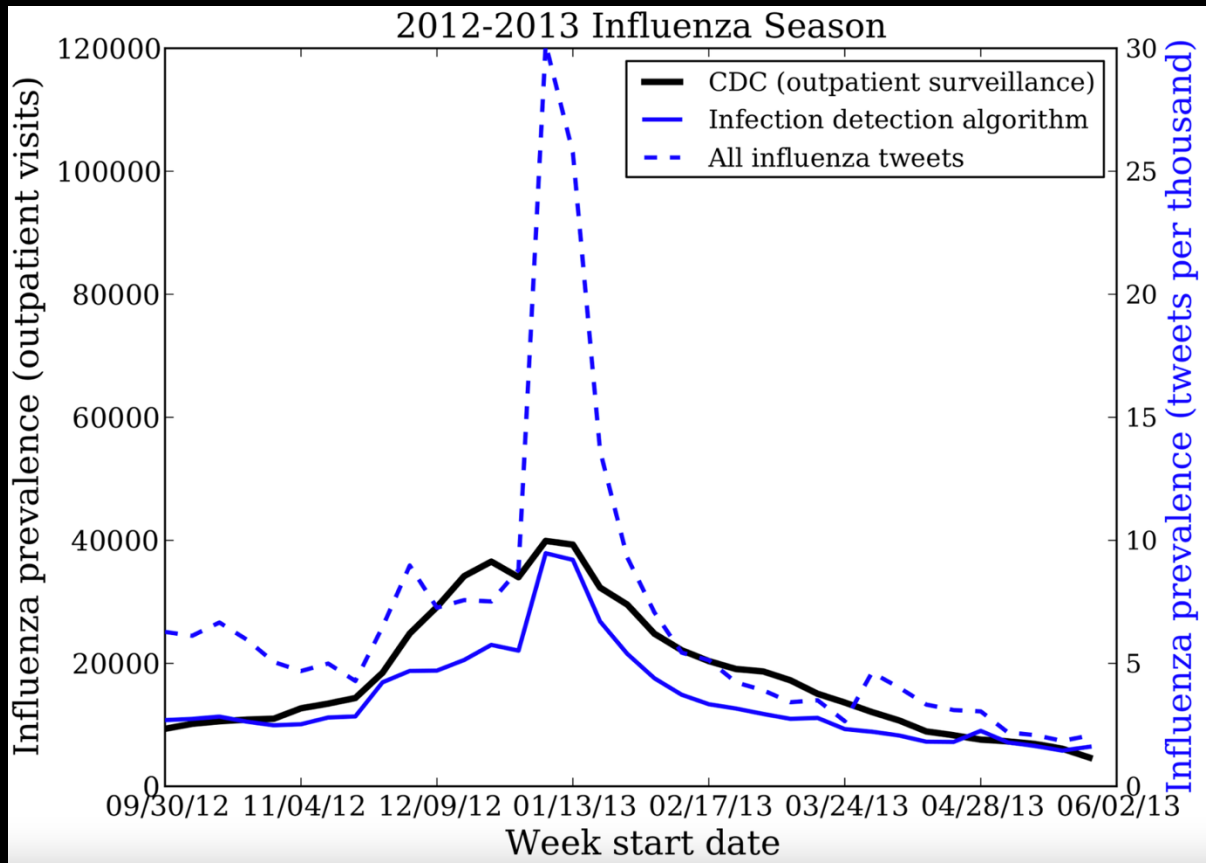
OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic

David A. Broniatowski, Michael J. Paul, Mark Dredze

Published: December 9, 2013 • https://doi.org/10.1371/journal.pone.0083672

Beyond Google searches…



PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

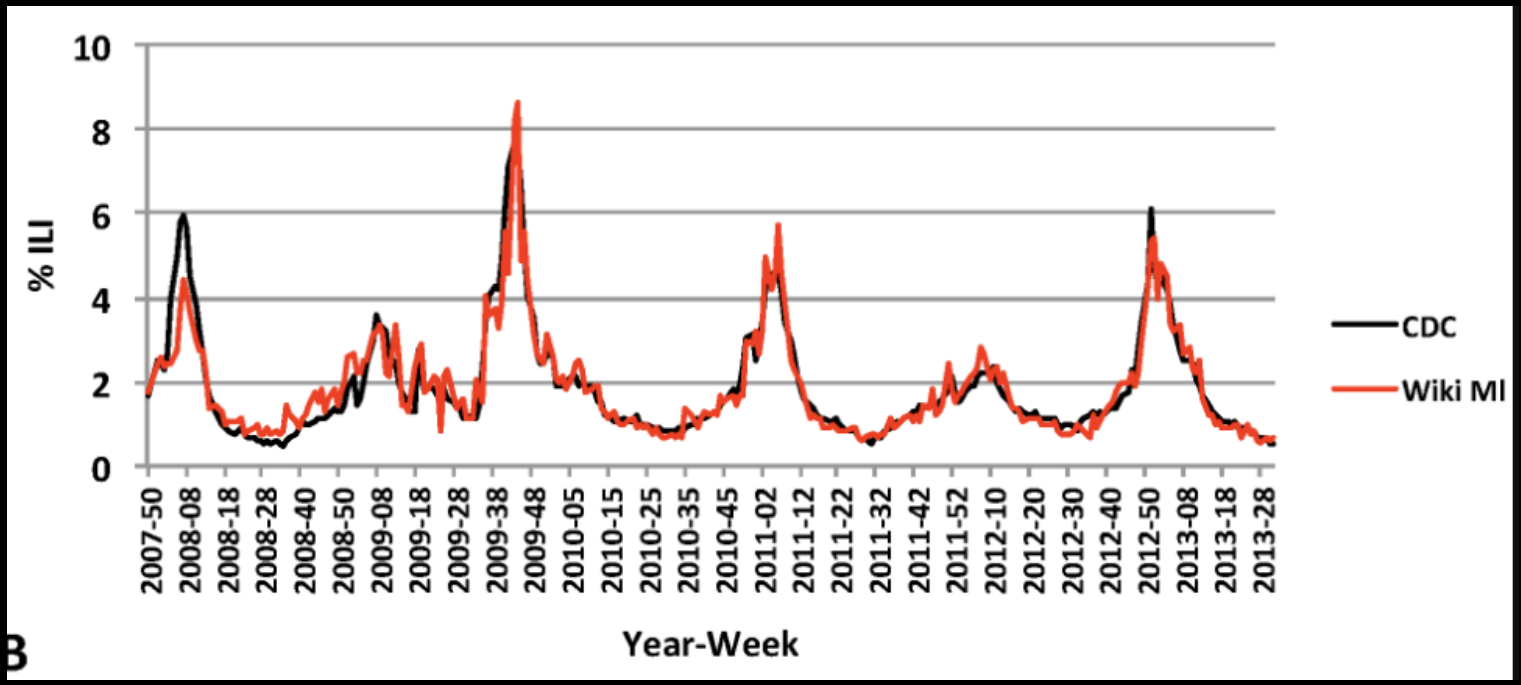
Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time

David J. McIver, John S. Brownstein

Published: April 17, 2014 • https://doi.org/10.1371/journal.pcbi.1003581

WIKIPEDIA
The Free Encyclopedia

Beyond Google searches…



PLOS COMPUTATIONAL BIOLOGY

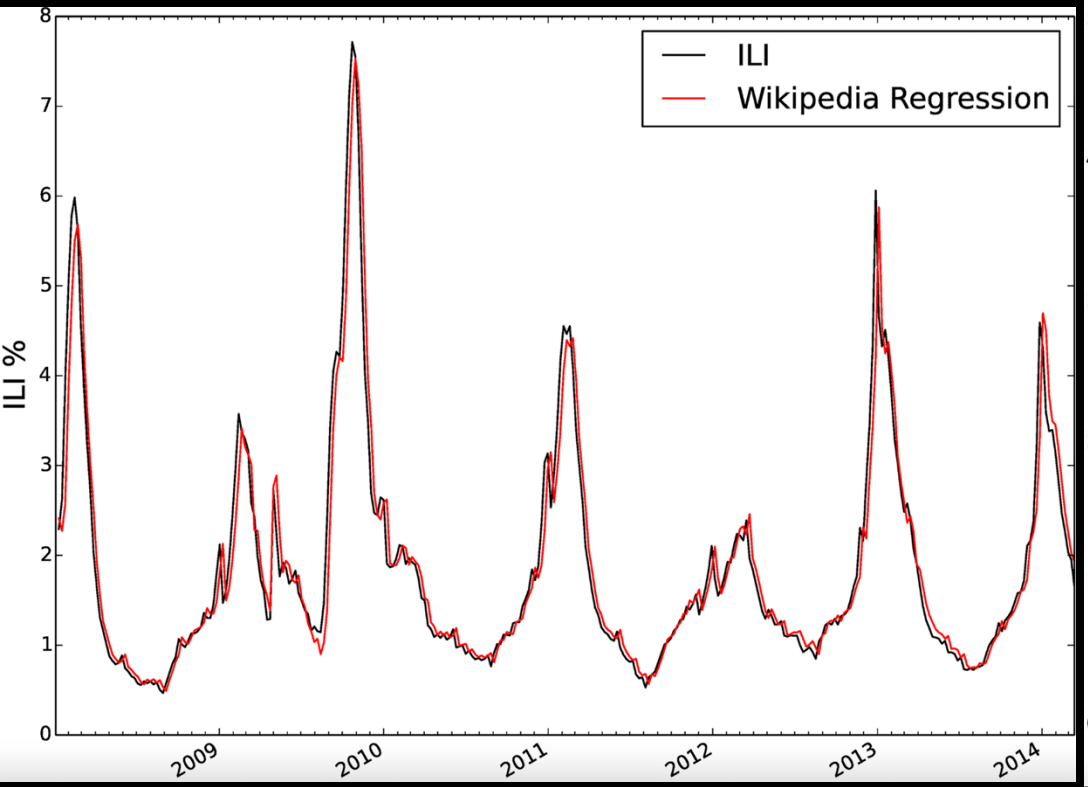
OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

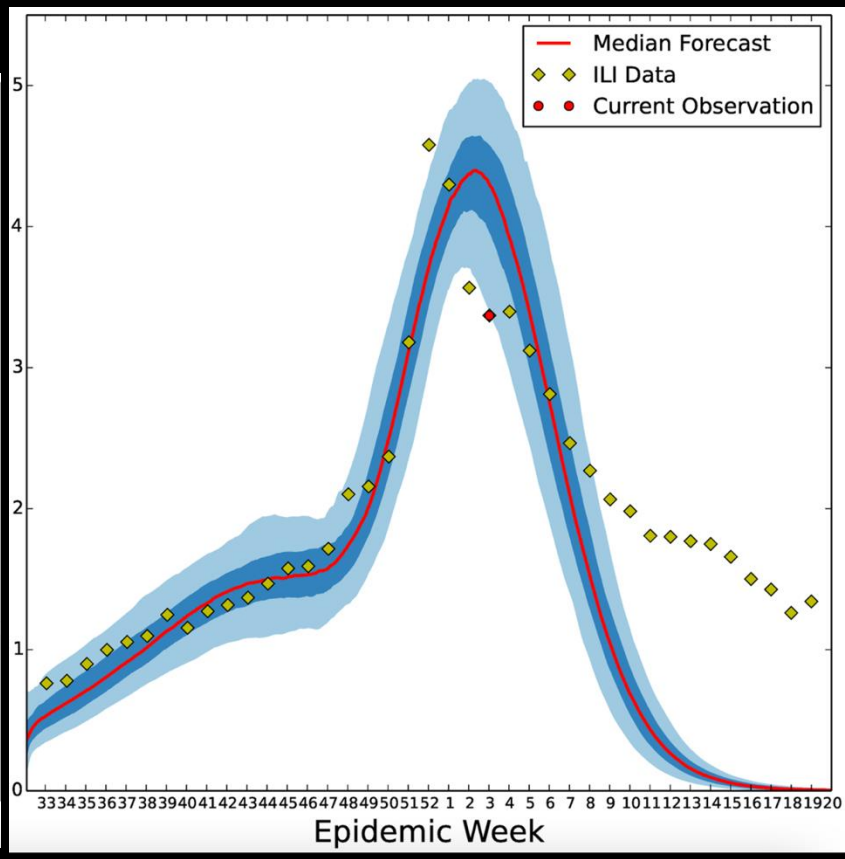
Forecasting the 2013–2014 Influenza Season Using Wikipedia

Kyle S. Hickmann, Geoffrey Fairchild, Reid Priedhorsky, Nicholas Generous, James M. Hyman, Alina Deshpande, Sara Y. Del Valle

Published: May 14, 2015 • https://doi.org/10.1371/journal.pcbi.1004239

WIKIPEDIA
The Free Encyclopedia



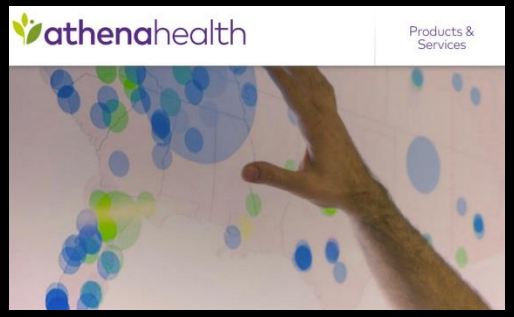


Beyond Google searches…

What are doctors searching for?

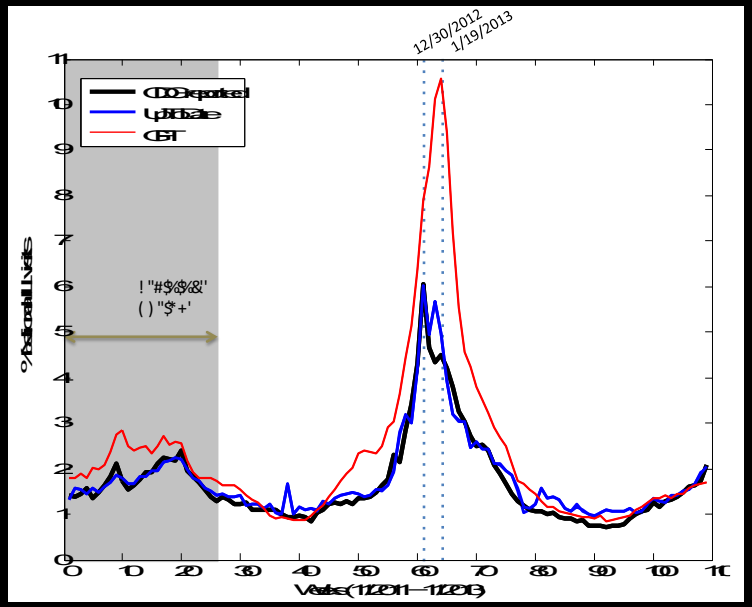
flu near you
do you have it in you?

What are people tweeting? What are they reporting on crowd-sourced disease surveillance apps?

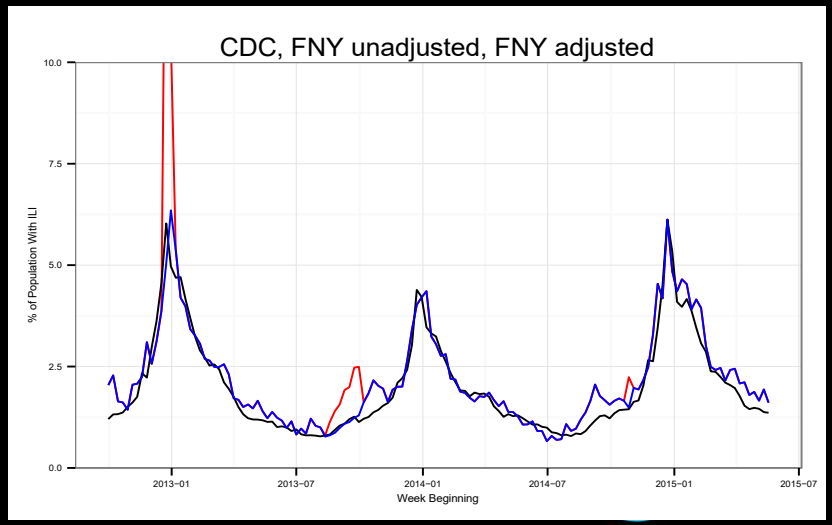


Can we use Electronic Health Records (EHR) to track disease incidence? What lab tests or medications are doctors prescribing?

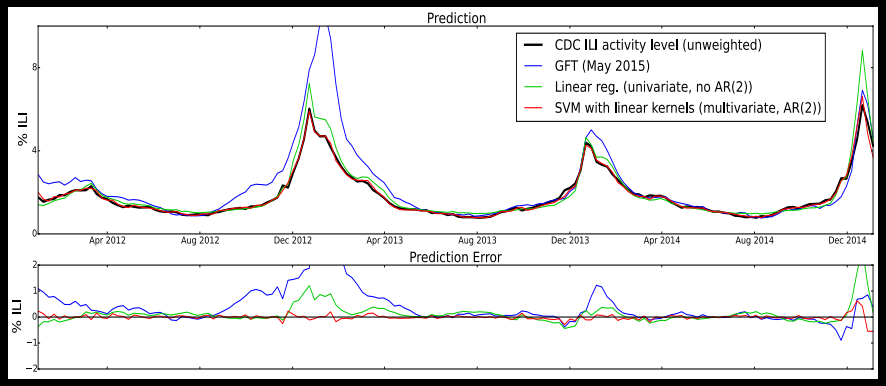
Beyond Google searches…



What are doctors searching for?

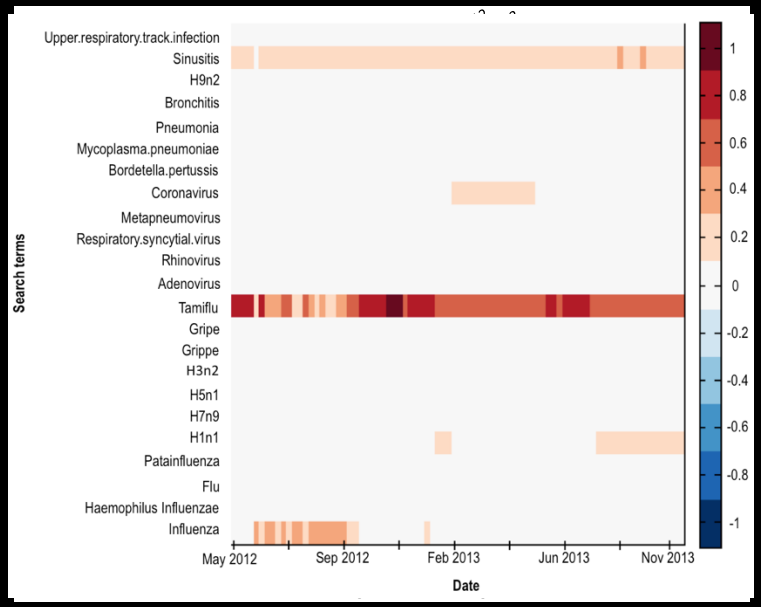


What are people tweeting? What are they reporting on crowd-sourced disease surveillance apps?



Can we use Electronic Health Records (EHR) to track disease incidence? What lab tests or medications are doctors prescribing?

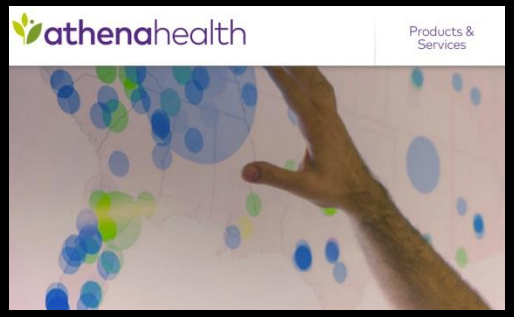
Beyond Google searches…



What are doctors searching for?

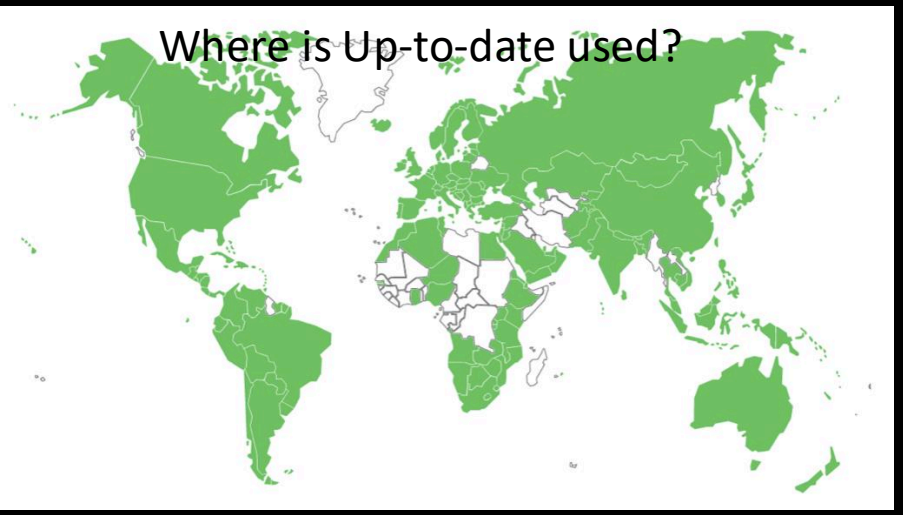
flu near you
do you have it in you?

What are people tweeting? What are they reporting on crowd-sourced disease surveillance apps?



Can we use Electronic Health Records (EHR) to track disease incidence? What lab tests or medications are doctors prescribing?

Beyond Google searches…



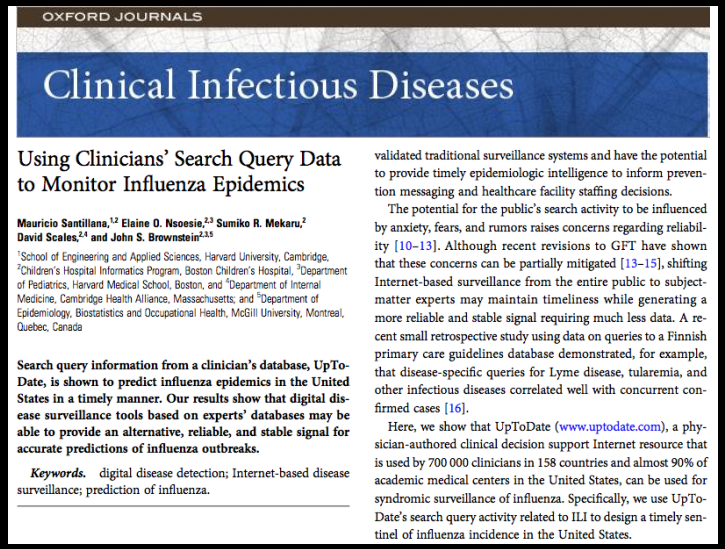
What are doctors searching for?

flu near you
do you have it in you?

What are people tweeting? What are they reporting on crowd-sourced disease surveillance apps?

Can we use Electronic Health Records (EHR) to track disease incidence? What lab tests or medications are doctors prescribing?

Beyond Google searches…



OXFORD JOURNALS

Clinical Infectious Diseases

Using Clinicians' Search Query Data to Monitor Influenza Epidemics

Mauricio Santillana,[1,2] Elaine O. Nsoesie,[2,3] Sumiko R. Mekaru,[2] David Scales,[2,4] and John S. Brownstein[2,5]

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, [2]Children's Hospital Informatics Program, Boston Children's Hospital, [3]Department of Pediatrics, Harvard Medical School, Boston, and [4]Department of Internal Medicine, Cambridge Health Alliance, Massachusetts; and [5]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

Search query information from a clinician's database, UpToDate, is shown to predict influenza epidemics in the United States in a timely manner. Our results show that digital disease surveillance tools based on experts' databases may be able to provide an alternative, reliable, and stable signal for accurate predictions of influenza outbreaks.

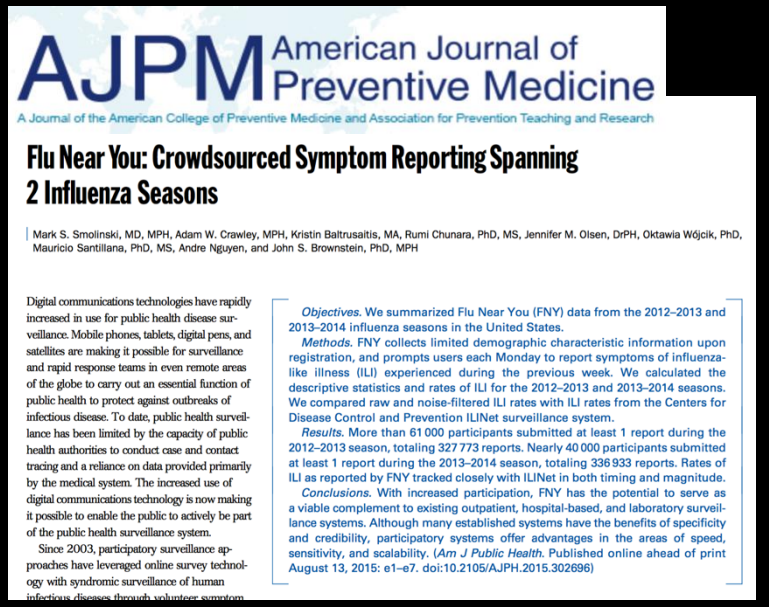
Keywords. digital disease detection; Internet-based disease surveillance; prediction of influenza.

validated traditional surveillance systems and have the potential to provide timely epidemiologic intelligence to inform prevention messaging and healthcare facility staffing decisions.

The potential for the public's search activity to be influenced by anxiety, fears, and rumors raises concerns regarding reliability [10–13]. Although recent revisions to GFT have shown that these concerns can be partially mitigated [13–15], shifting Internet-based surveillance from the entire public to subject-matter experts may maintain timeliness while generating a more reliable and stable signal requiring much less data. A recent small retrospective study using data on queries to a Finnish primary care guidelines database demonstrated, for example, that disease-specific queries for Lyme disease, tularemia, and other infectious diseases correlated well with concurrent confirmed cases [16].

Here, we show that UpToDate (www.uptodate.com), a physician-authored clinical decision support Internet resource that is used by 700 000 clinicians in 158 countries and almost 90% of academic medical centers in the United States, can be used for syndromic surveillance of influenza. Specifically, we use UpToDate's search query activity related to ILI to design a timely sentinel of influenza incidence in the United States.

What are doctors searching for?



AJPM American Journal of Preventive Medicine

A Journal of the American College of Preventive Medicine and Association for Prevention Teaching and Research

Flu Near You: Crowdsourced Symptom Reporting Spanning 2 Influenza Seasons

Mark S. Smolinski, MD, MPH, Adam W. Crawley, MPH, Kristin Baltrusaitis, MA, Rumi Chunara, PhD, MS, Jennifer M. Olsen, DrPH, Oktawia Wójcik, PhD, Mauricio Santillana, PhD, MS, Andre Nguyen, and John S. Brownstein, PhD, MPH

Digital communications technologies have rapidly increased in use for public health disease surveillance. Mobile phones, tablets, digital pens, and satellites are making it possible for surveillance and rapid response teams in even remote areas of the globe to carry out an essential function of public health to protect against outbreaks of infectious disease. To date, public health surveillance has been limited by the capacity of public health authorities to conduct case and contact tracing and a reliance on data provided primarily by the medical system. The increased use of digital communications technology is now making it possible to enable the public to actively be part of the public health surveillance system.

Since 2003, participatory surveillance approaches have leveraged online survey technology with syndromic surveillance of human infectious diseases through volunteer symptom

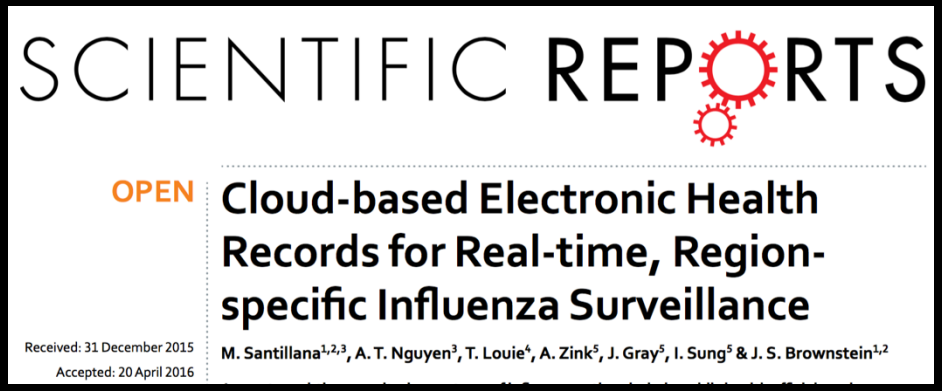
Objectives. We summarized Flu Near You (FNY) data from the 2012–2013 and 2013–2014 influenza seasons in the United States.

Methods. FNY collects limited demographic characteristic information upon registration, and prompts users each Monday to report symptoms of influenza-like illness (ILI) experienced during the previous week. We calculated the descriptive statistics and rates of ILI for the 2012–2013 and 2013–2014 seasons. We compared raw and noise-filtered ILI rates with ILI rates from the Centers for Disease Control and Prevention ILINet surveillance system.

Results. More than 61 000 participants submitted at least 1 report during the 2012–2013 season, totaling 327 773 reports. Nearly 40 000 participants submitted at least 1 report during the 2013–2014 season, totaling 336 933 reports. Rates of ILI as reported by FNY tracked closely with ILINet in both timing and magnitude.

Conclusions. With increased participation, FNY has the potential to serve as a viable complement to existing outpatient, hospital-based, and laboratory surveillance systems. Although many established systems have the benefits of specificity and credibility, participatory systems offer advantages in the areas of speed, sensitivity, and scalability. (*Am J Public Health.* Published online ahead of print August 13, 2015: e1–e7. doi:10.2105/AJPH.2015.302696)

What are people tweeting? What are they reporting on crowd-sourced disease surveillance apps?



SCIENTIFIC REPORTS

OPEN

Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance

Received: 31 December 2015
Accepted: 20 April 2016

M. Santillana[1,2,3], A. T. Nguyen[3], T. Louie[4], A. Zink[5], J. Gray[5], I. Sung[5] & J. S. Brownstein[1,2]

Can we use Electronic Health Records (EHR) to track disease incidence? What lab tests or medications are doctors prescribing?

Disease: **Influenza**

Goal: **short-term forecasting**

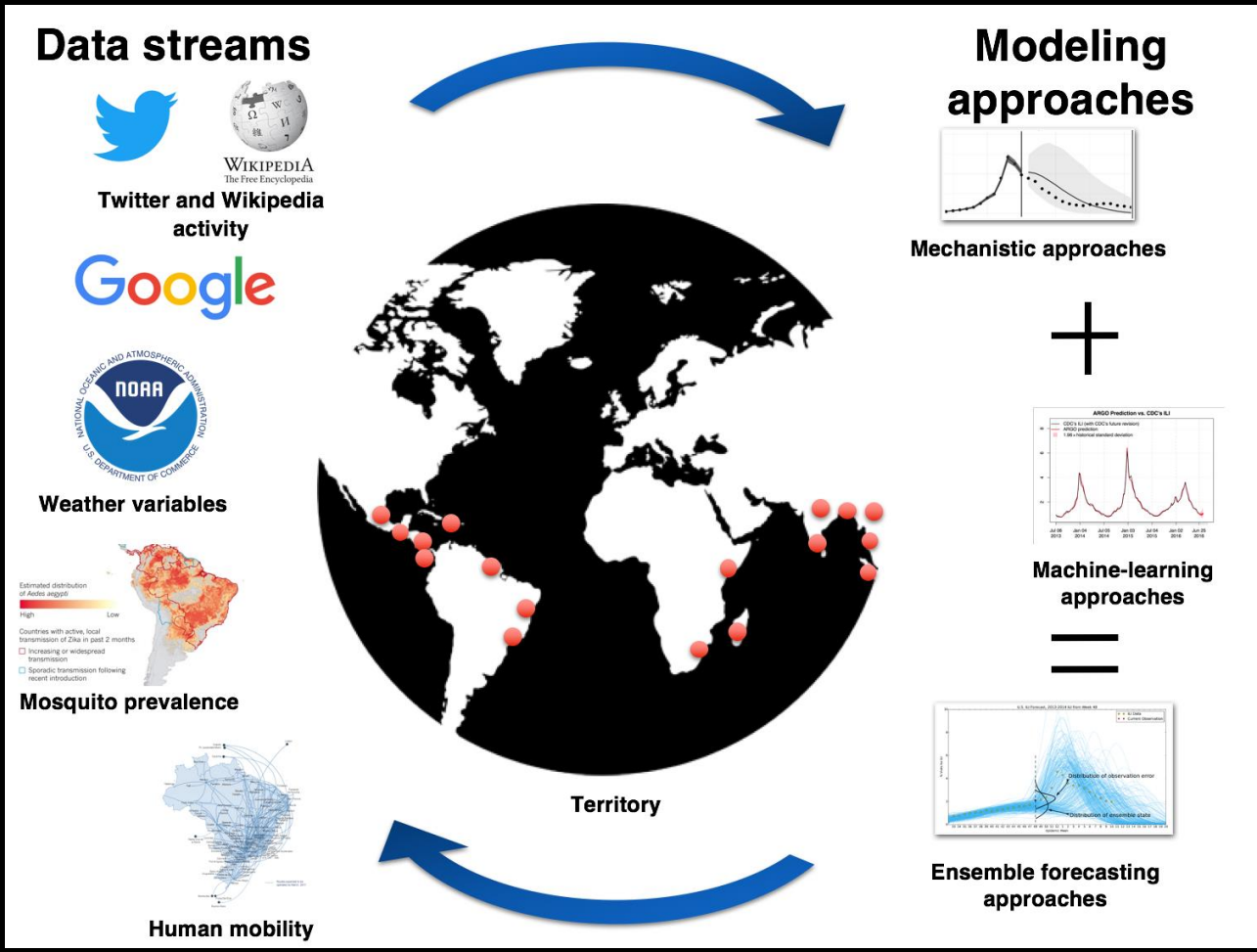
Location: **United States** (Data rich, wealthy country)

Spatial resolution: **Country**

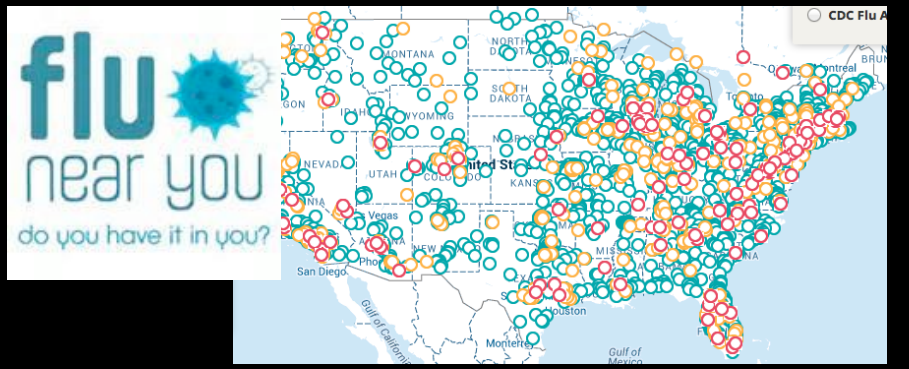
Method: **Machine learning**

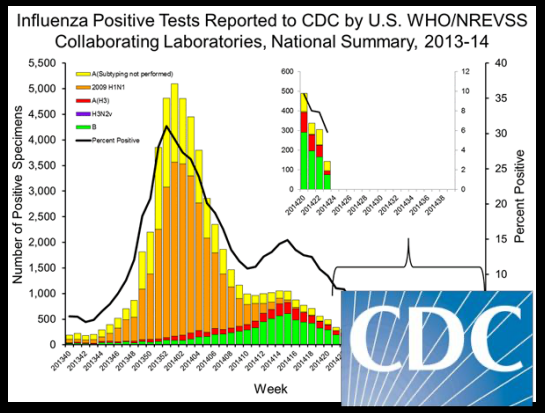
Input data sources:

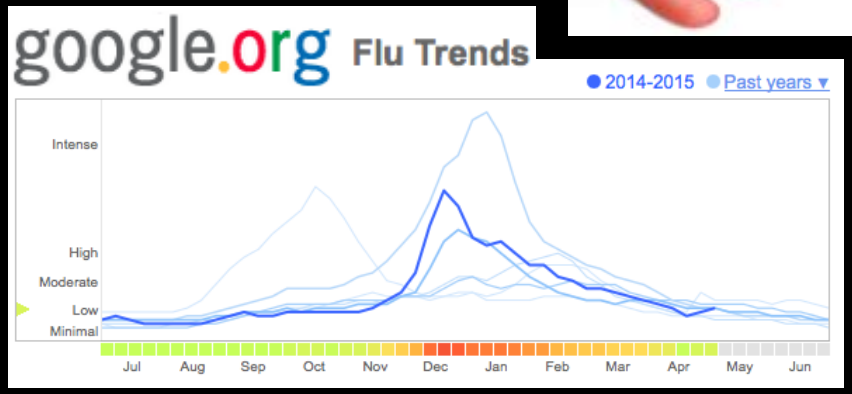
- **Historical flu activity**
- **Google search activity**
- **Electronic Health records**
- **Crowd sourced information**

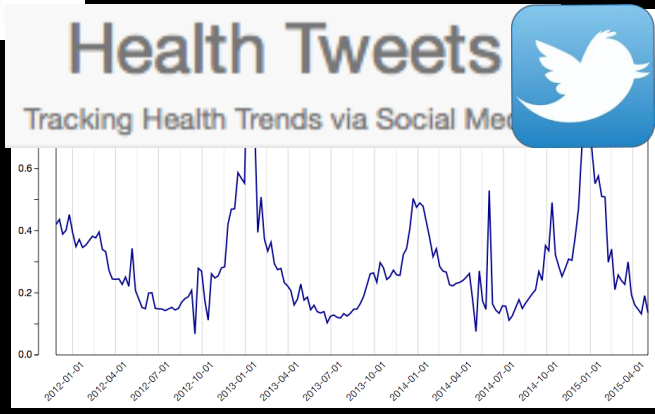


Ensemble approaches yield more accurate and more robust real-time and forecast flu estimates









Performance of individual data sources

	CORR	RMSE (%ILI)	Rel RMSE (%)	RMAE (%)	Hit Rate
FNY	0.948	0.385	15.9	39.3	65.9
ATH	0.977	0.351	14.1	36.7	**77.7**
GT	0.978	0.245	13.3	42.9	65.9
GFT	0.980	0.333	12.3	35.3	75.3
TWT	0.937	0.414	15.1	50.1	62.4
CDC Baseline	0.930	0.501	18.2	46.7	68.2
CDC Virology	0.923	-	-	-	69.4

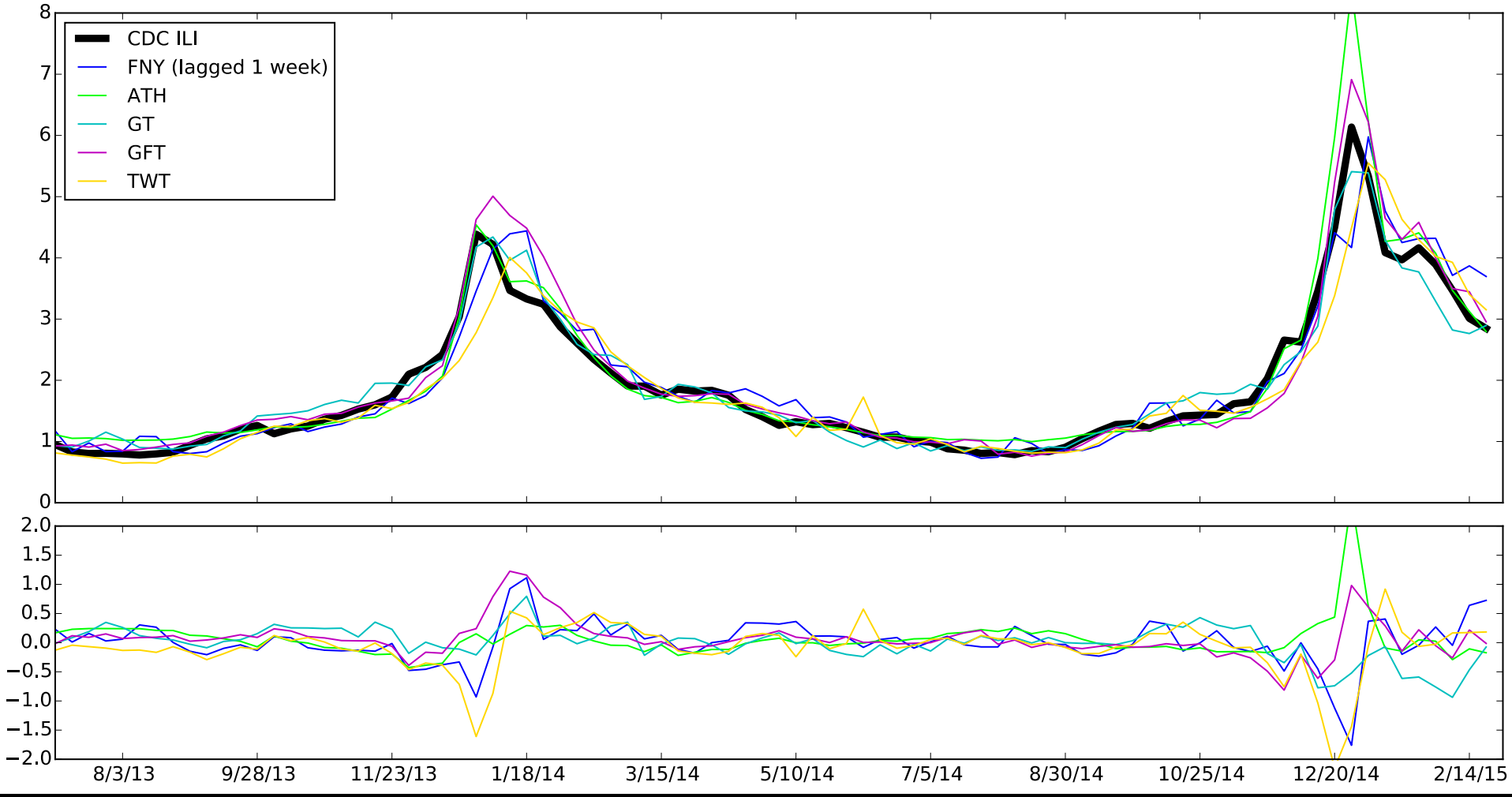
Santillana et al. PLoS Computational Biology, 2015

Performance ensemble

	CORR	RMSE (%ILI)	Rel RMSE (%)	RMAE (%)	Hit Rate
FNY	0.948	0.385	15.9	39.3	65.9
ATH	0.977	0.351	14.1	36.7	**77.7**
GT	0.978	0.245	13.3	42.9	65.9
GFT	0.980	0.333	12.3	35.3	75.3
TWT	0.937	0.414	15.1	50.1	62.4
CDC Baseline	0.930	0.501	18.2	46.7	68.2
CDC Virology	0.923	-	-	-	69.4
SVM (RBF)	**0.989**	**0.176**	**8.27**	**23.6**	69.4

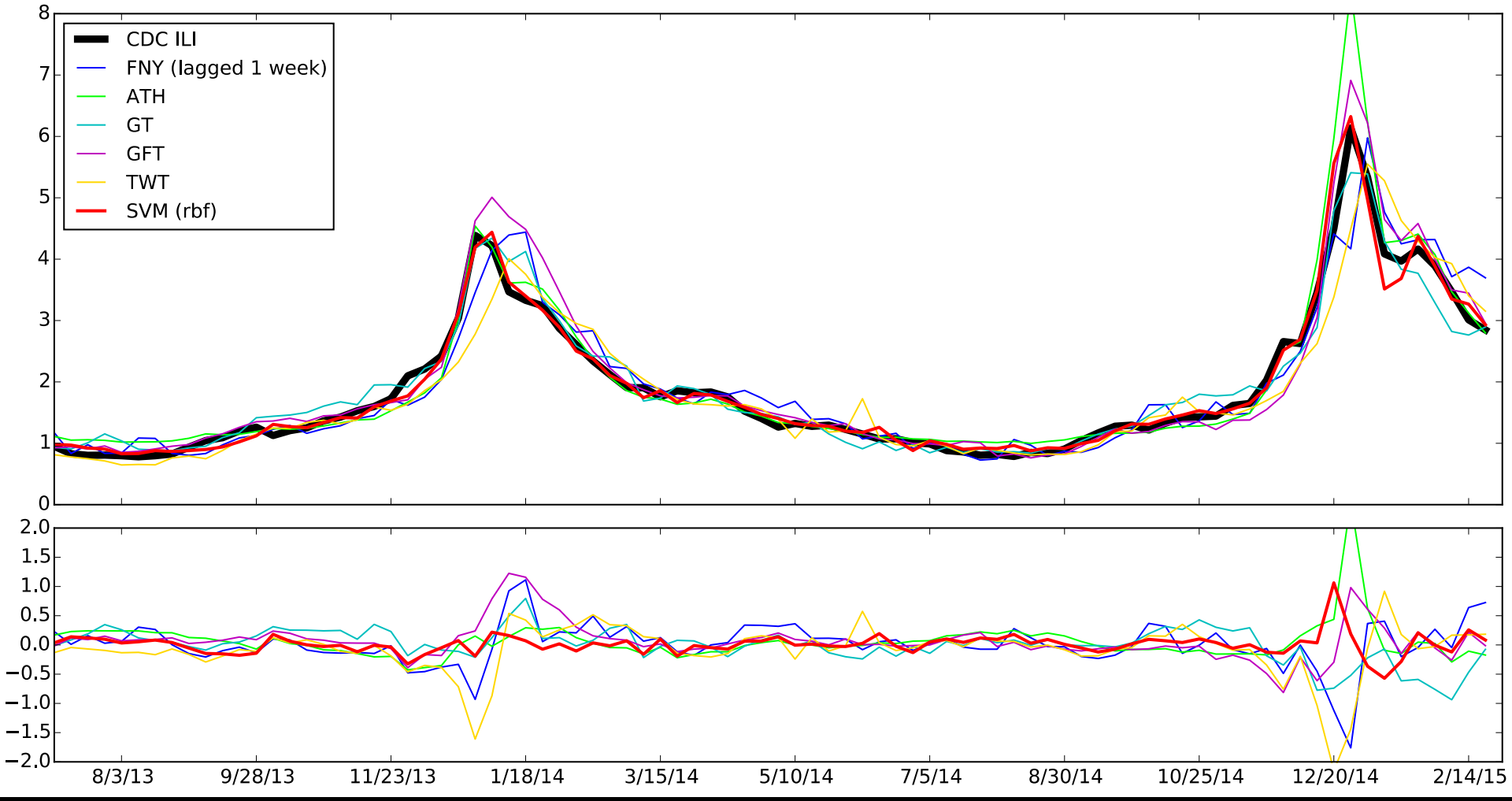
Santillana et al. PLoS Computational Biology, 2015

Performance of individual data sources



Santillana et al. PLoS Computational Biology, 2015

Performance ensemble



Santillana et al. PLoS Computational Biology, 2015

Ensemble approaches yield more accurate and more robust real-time and forecast flu estimates

Yang et al. BMC Infectious Diseases (2017) 17:332
DOI 10.1186/s12879-017-2424-7

BMC Infectious Diseases

RESEARCH ARTICLE Open Access

Using electronic health records and Internet search information for accurate influenza forecasting

Shihao Yang[1], Mauricio Santillana[2,3*], John S. Brownstein[2,3], Josh Gray[4], Stewart Richardson[4] and S. C. Kou[1*]

Abstract

Background: Accurate influenza activity forecasting helps public health officials prepare and allocate resources for unusual influenza activity. Traditional flu surveillance systems, such as the Centers for Disease Control and Prevention's (CDC) influenza-like illnesses reports, lag behind real-time by one to 2 weeks, whereas information contained in cloud-based electronic health records (EHR) and in Internet users' search activity is typically available in near real-time. We present a method that combines the information from these two data sources with historical flu activity to produce national flu forecasts for the United States up to 4 weeks ahead of the publication of CDC's flu reports.

Methods: We extend a method originally designed to track flu using Google searches, named ARGO, to combine information from EHR and Internet searches with historical flu activities. Our regularized multivariate regression model dynamically selects the most appropriate variables for flu prediction every week. The model is assessed for the flu seasons within the time period 2013–2016 using multiple metrics including root mean squared error (RMSE).

Results: Our method reduces the RMSE of the publicly available alternative (Healthmap flutrends) method by 33, 20, 17 and 21%, for the four time horizons: real-time, one, two, and 3 weeks ahead, respectively. Such accuracy improvements are statistically significant at the 5% level. Our real-time estimates correctly identified the peak timing and magnitude of the studied flu seasons.

Conclusions: Our method significantly reduces the prediction error when compared to historical publicly available Internet-based prediction systems, demonstrating that: (1) the method to combine data sources is as important as data quality; (2) effectively extracting information from a cloud-based EHR and Internet search activity leads to accurate forecast of flu.

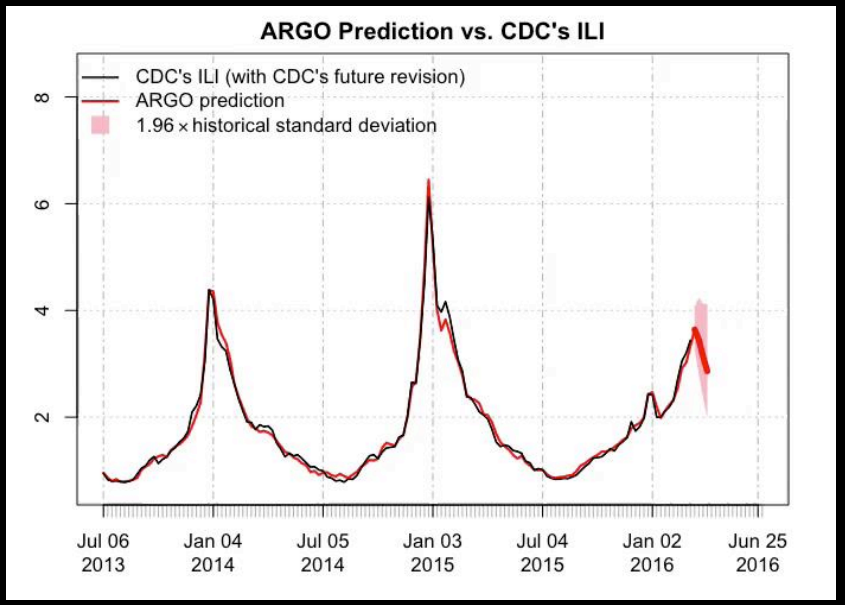
Keywords: Influenza-like illnesses reports, Digital disease detection, Dynamic error reduction, Validation test, Autoregression

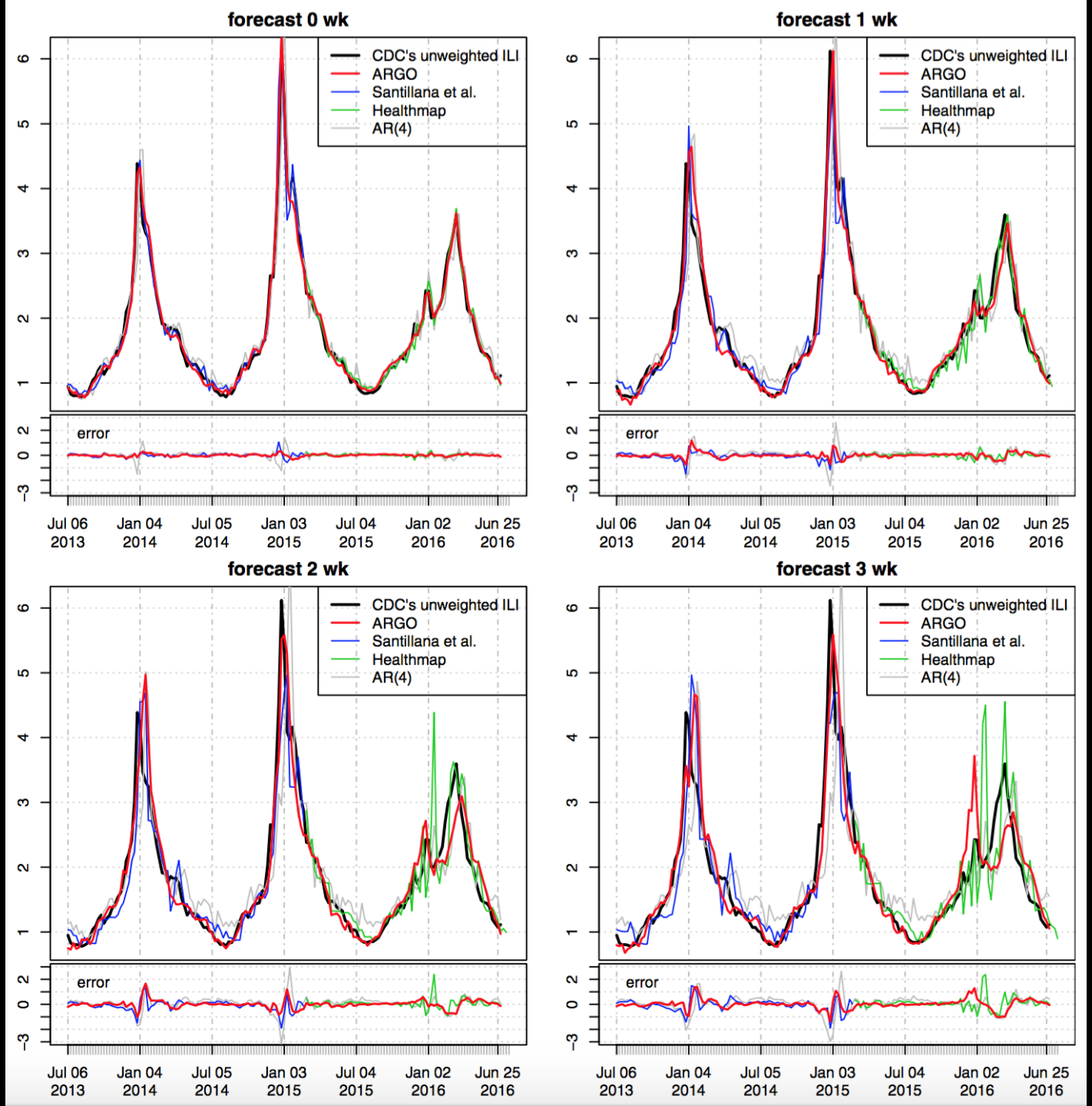
PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance

Mauricio Santillana[1,2,3]*, André T. Nguyen[1], Mark Dredze[4], Michael J. Paul[5], Elaine O. Nsoesie[6,7], John S. Brownstein[2,3]




forecast 0 wk
forecast 1 wk
forecast 2 wk
forecast 3 wk
CDC's unweighted ILI
ARGO
Santillana et al.
Healthmap
AR(4)
error
1
2
3
4
5
6
−3
0
2
Jul 06 2013
Jan 04 2014
Jul 05 2014
Jan 03 2015
Jul 04 2015
Jan 02 2016
Jun 25 2016

nature COMMUNICATIONS

Article | OPEN | Published: 11 January 2019

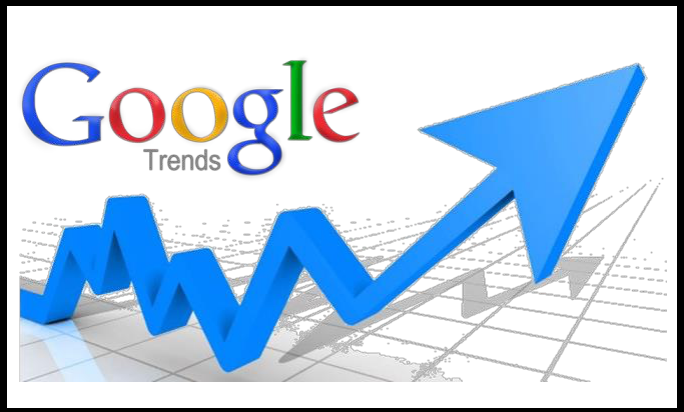
Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches

Fred S. Lu ✉, Mohammad W. Hattab, Cesar Leonardo Clemente, Matthew Biggerstaff & Mauricio Santillana ✉

Nature Communications **10**, Article number: 147 (2019) | Download Citation ↓

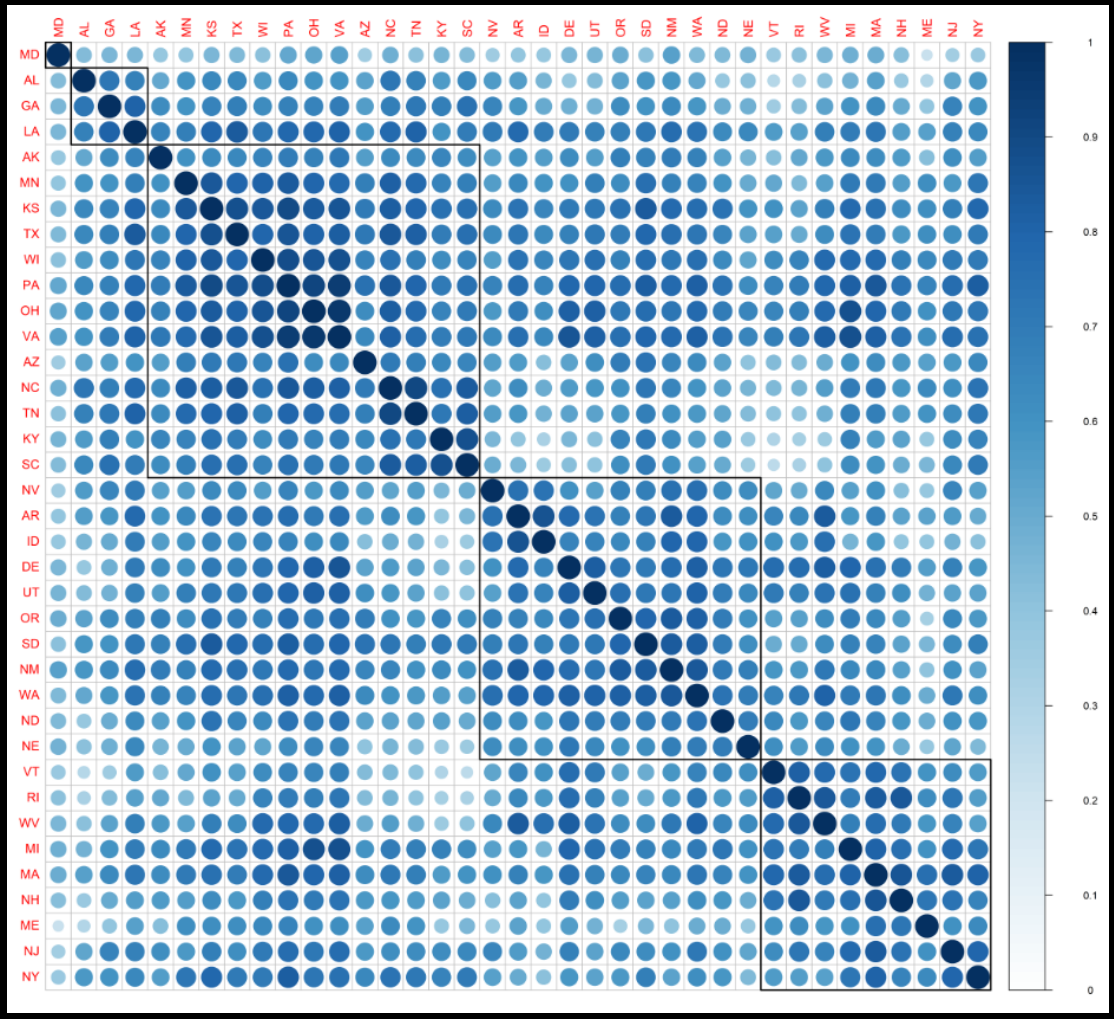
Spatial-temporal synchronicities

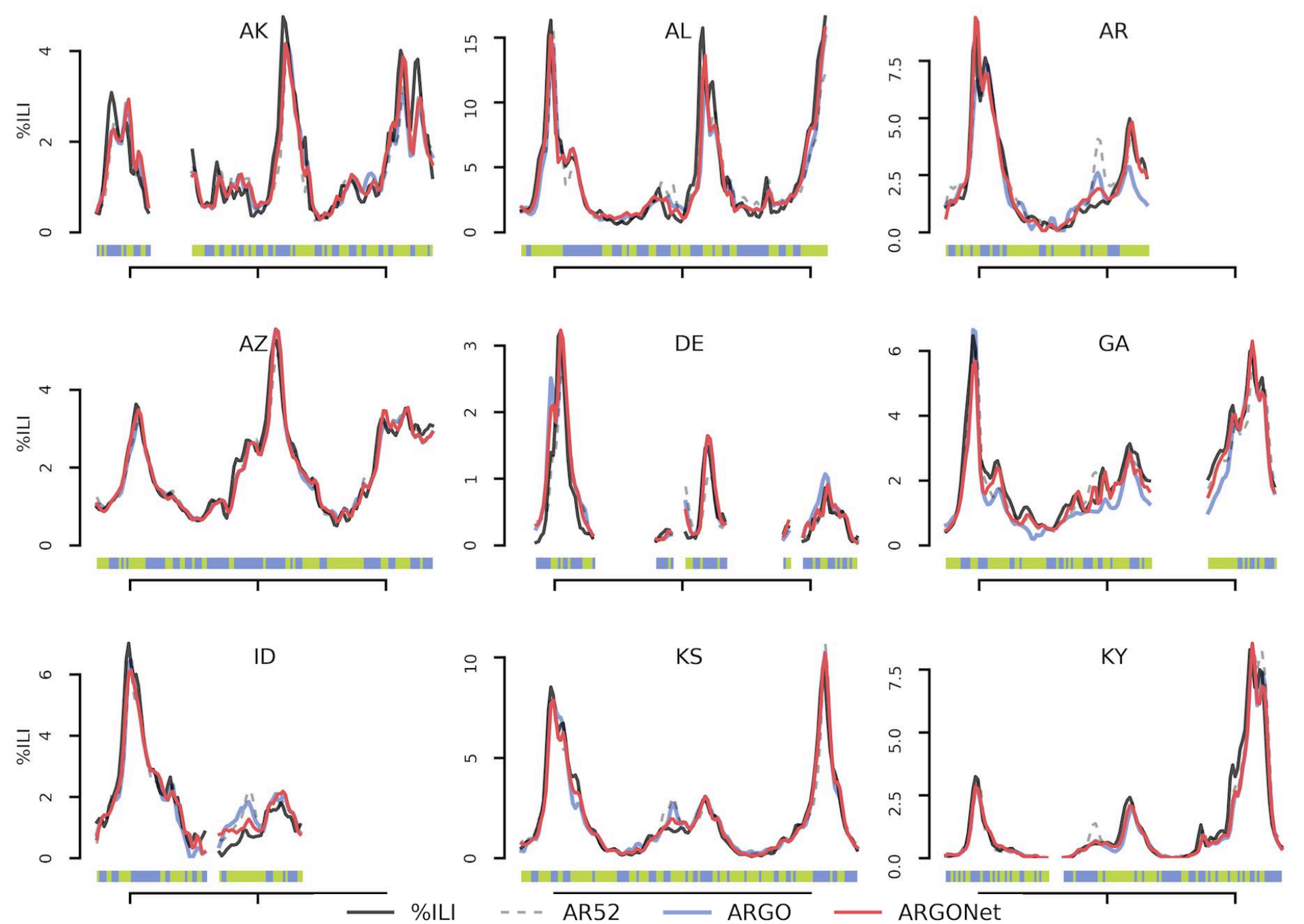
Flu-related Google search information

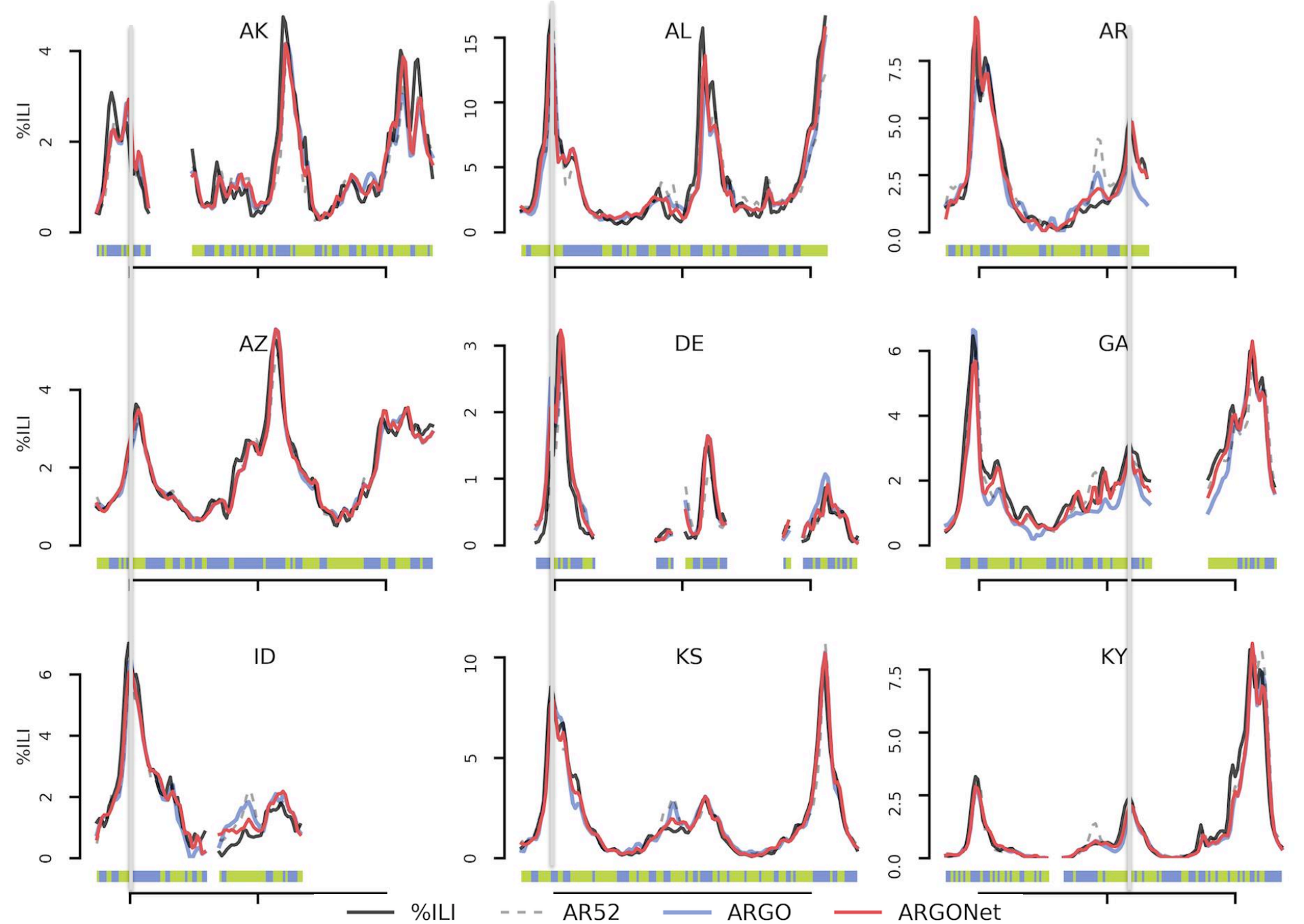


Lu F, Hattab M, Clemente L, **Santillana M.** *Improved state-level influenza activity nowcasting in the United States leveraging Internet-based data sources and network approaches via ARGONet*. Nature Communications. 2019; 10 (147)

Heat map of pairwise %ILI **correlations** between all states.
Boxes denote clusters of highly correlated states.




AK
AL
AR
AZ
DE
GA
ID
KS
KY
%ILI
%ILI
AR52
ARGO
ARGONet
ARGO selected
Net selected


AK
AL
AR
AZ
DE
GA
ID
KS
KY
%ILI
%ILI
%ILI
%ILI
AR52
ARGO
ARGONet
ARGO selected
Net selected

Refining the spatial resolution…

Tracking Flu using twitter
(Daily analysis in NYC)

Work with R. Nagar, Q. Yuan, C. Freifeld, A. Nojima, R. Chunara, and J. S. Brownstein

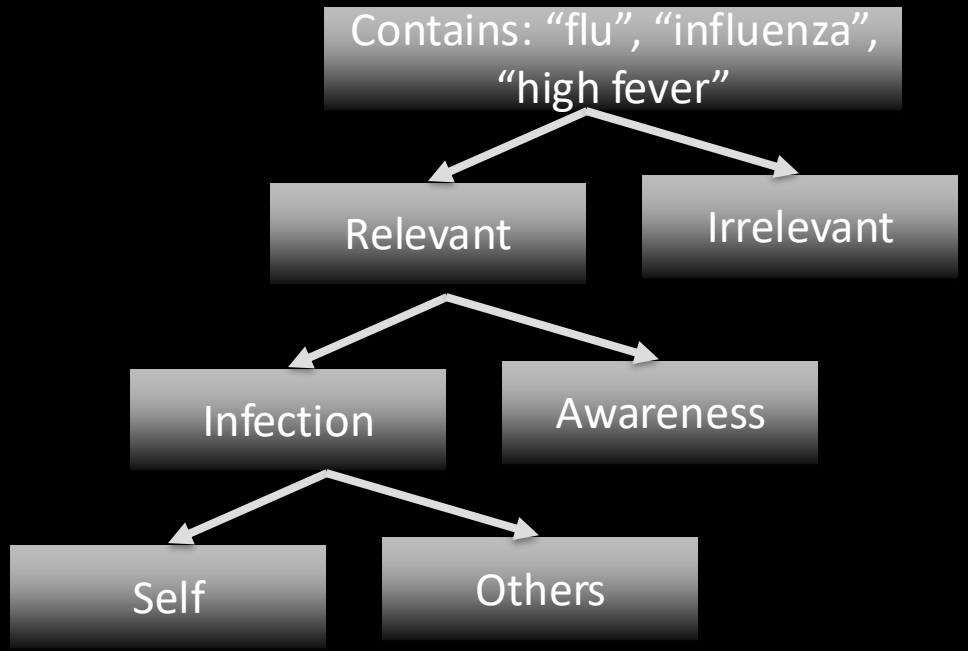
Natural Language Processing
(Using geo-located tweets)

1. Identified tweets containing “flu”, “influenza”, “gripe”, “high fever”
2. Classified tweets in categories

Table 1. Examples of Classified Tweets

Label	Example Tweets
Irrelevant	The flu shot prevents hangovers, so going all out w/5 fingers of...wait, it's "no drinking" that prevents hangovers? #toolate
	Romney: I would not put no flu zones over Syria. Military is not necessary in the conflict #debate2012
RISH	This flu is kicking my butt,2nd day off work. Hopefully I can win the battle because I'm losing sick days & that might hurt my pockets later
	Maldito flu sueltame!!! ðŸ˜«ðŸ˜¡
RISM	Uhhh I think I might be getting the flu /:
	Creo que me va a dar Gripe :'(
RISL	Finally getting over a miserae flu.
	@boyXsupreme ik ik it's awful. the past two weeks darci and I have had the flu but thank god we're done with it. get lots of rest + tlc ðŸ’•
RIOH	When @CiaraAnnex3 has the flu... Love you but stay the hell away <3
	Running on no sleep my poor daughter has the flu
RIOM	@giannarussso YOU PROB GOTS THE FLU!!, ariannas has it
	@YanieseRivera damn girl, do you have the flu?
RIOL	On day 6, son's #flu is gone. He threw open side door and screamed to the outside, "FREEDOM"! Then shoveled snow. I am miserable on day 3.
	@_AlexAlford she's good too, fortunately she never actually got the flu. just a fever for a day or two
RASH	I'd rather get the flu than get the flu shot, #JustSayin. No needles for me.
	The flu is an epidemic here and I volunteer at a preschool twice a week. If I don't get he flu it will be a miracle.
RASM	I survived more than a week in NYC without contracting the flu! Let's hope the plane ride home won't break me. Trying to stay healthy here.
	So glad to be back in NYC, but stay away from me you Flu filled city.
RASL	Ah yea... flu shot acquired! (@ Duane Reade) http://t.co/RSR2BI11
	Just got a flu shot and Tdap booster at @onemedical â€" if youâ€™ll be in close contact with an infant, consider taking these vaccines.
RAOH	Flu infections sweep America hospitalizing thousands and leaving 18 children dead of complications, ... http://t.co/Yf9eQikm
	19,000 flu cases across NY this week. I'm not leaving my house.
RAOM	Wash your hand. America sick girl. #influenza #SICKENING #cleanup http://t.co/MKdASaV9
	The latest figures from the CDC show that flu cases are still rising in the west. Listen to our newscast: http://t.co/0kCiyoP8
RAOL	#patient advice. #flu vaccine not only protects you but your community Too. Less outbreaks. U may survive the flu, but sicker people may not
	â€œ@grubstreetny: Hereâ€™s How New York Chefs Beat the Flu http://t.co/bV3pEZ2Câ€ uuuuu for real?

Categories



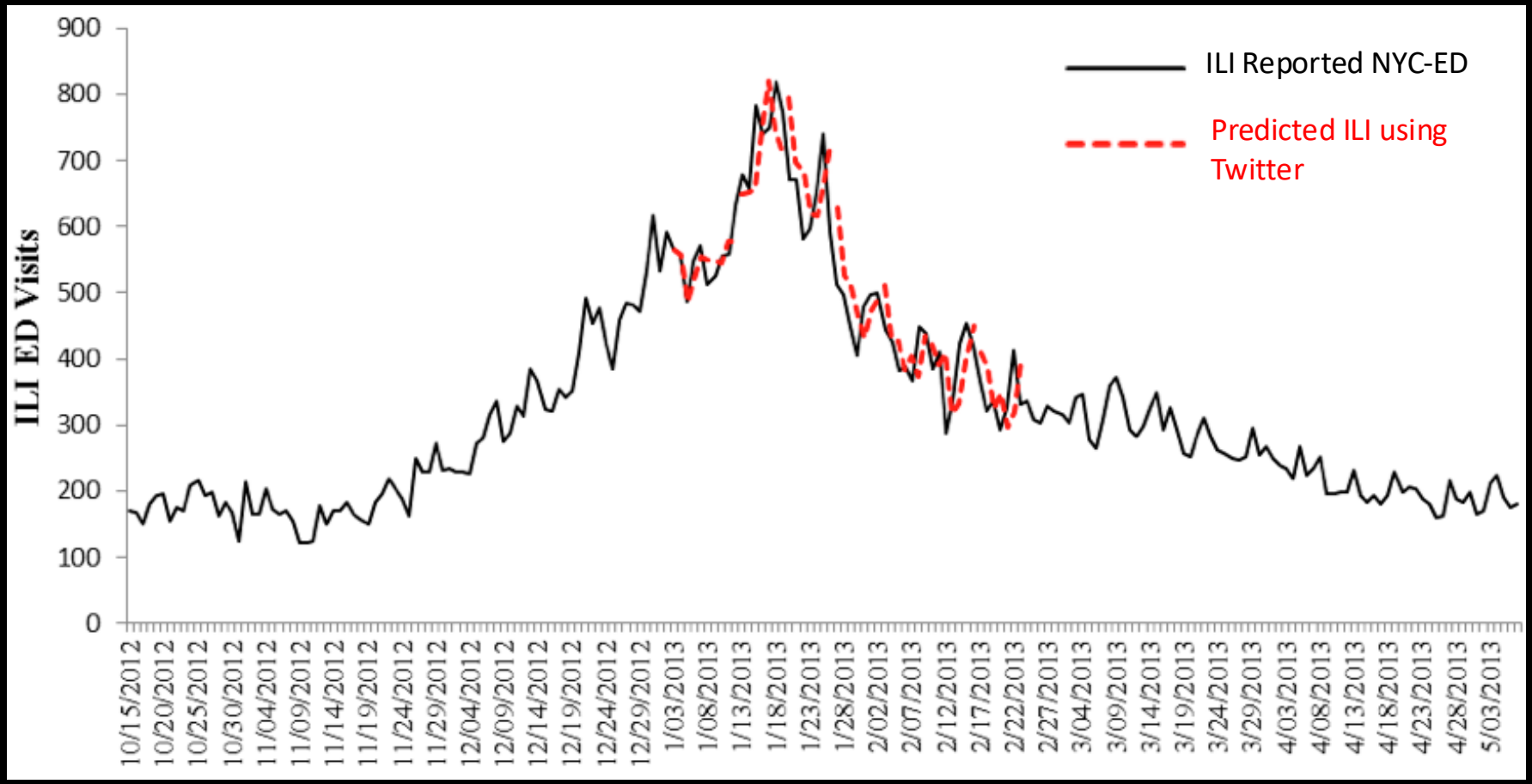
First experiment: was done by hand…

Nagar et al. (2014) Journal of Medical Internet Research. In press

Table 1. Examples of Classified Tweets

Label	Example Tweets
Irrelevant	The flu shot prevents hangovers, so going all out w/5 fingers of...wait, it's "no drinking" that prevents hangovers? #toolate
	Romney: I would not put no flu zones over Syria. Military is not necessary in the conflict #debate2012
RISH	This flu is kicking my butt,2nd day off work. Hopefully I can win the battle because I'm losing sick days & that might hurt my pockets later
	Maldito flu sueltame!!! ðŸ˜«ðŸ˜¡
RISM	Uhhh I think I might be getting the flu /:
	Creo que me va a dar Gripe :'(
RISL	Finally getting over a miserae flu.
	@boyXsupreme ik ik it's awful. the past two weeks darci and I have had the flu but thank god we're done with it. get lots of rest + tlc ðŸ’•
RIOH	When @CiaraAnnex3 has the flu... Love you but stay the hell away <3
	Running on no sleep my poor daughter has the flu
RIOM	@giannarussso YOU PROB GOTS THE FLU!!, ariannas has it
	@YanieseRivera damn girl, do you have the flu?
RIOL	On day 6, son's #flu is gone. He threw open side door and screamed to the outside, "FREEDOM"! Then shoveled snow. I am miserable on day 3.
	@_AlexAlford she's good too, fortunately she never actually got the flu. just a fever for a day or two
RASH	I'd rather get the flu than get the flu shot, #JustSayin. No needles for me.
	The flu is an epidemic here and I volunteer at a preschool twice a week. If I don't get he flu it will be a miracle.
RASM	I survived more than a week in NYC without contracting the flu! Let's hope the plane ride home won't break me. Trying to stay healthy here.
	So glad to be back in NYC, but stay away from me you Flu filled city.
RASL	Ah yea... flu shot acquired! (@ Duane Reade) http://t.co/RSR2Bl11
	Just got a flu shot and Tdap booster at @onemedical â€" if youâ€™ll be in close contact with an infant, consider taking these vaccines.
RAOH	Flu infections sweep America hospitalizing thousands and leaving 18 children dead of complications, ... http://t.co/Yf9eQikm
	19,000 flu cases across NY this week. I'm not leaving my house.
RAOM	Wash your hand. America sick girl. #influenza #SICKENING #cleanup http://t.co/MKdASaV9
	The latest figures from the CDC show that flu cases are still rising in the west. Listen to our newscast: http://t.co/0kCiyoP8
RAOL	#patient advice. #flu vaccine not only protects you but your community Too. Less outbreaks. U may survive the flu, but sicker people may not

Daily ILI visits (as reported by the NYC emergency department) compared to predicted ILI using twitter data

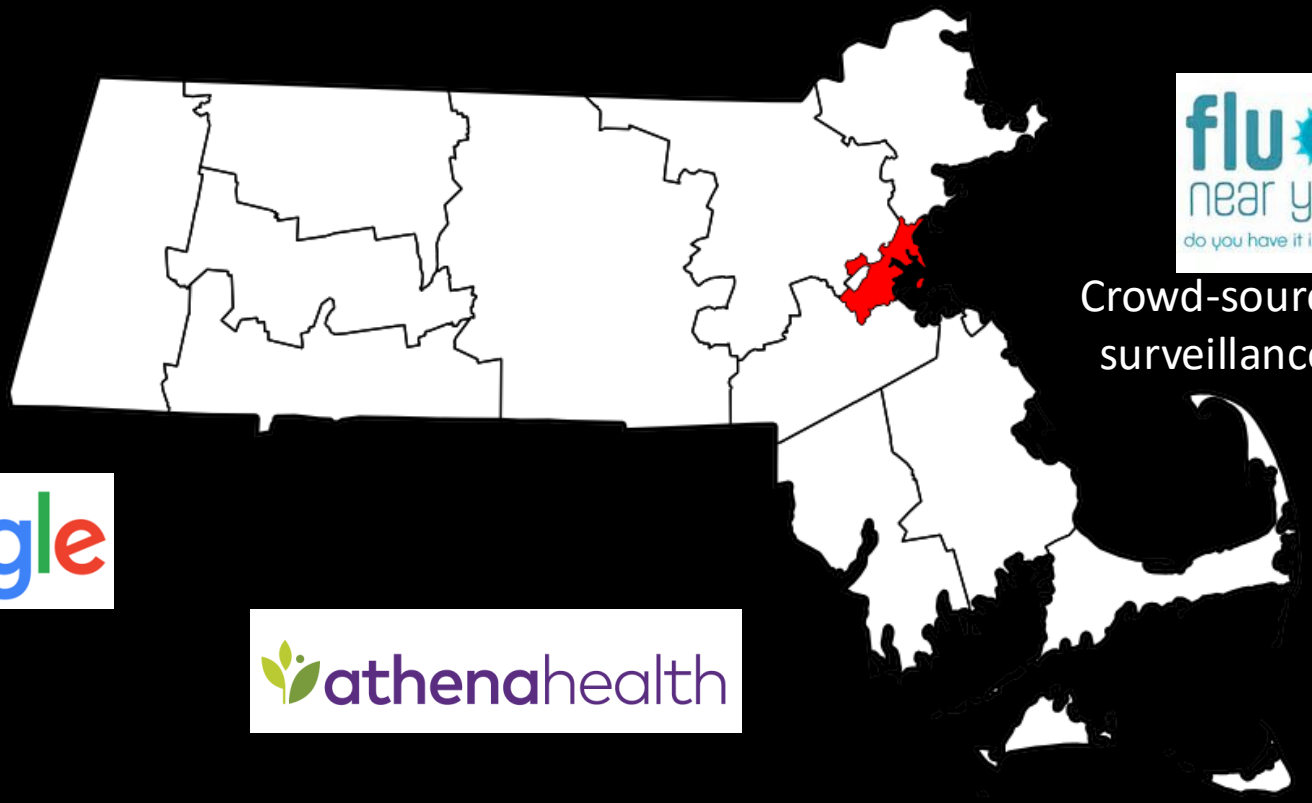


Nagar et al. (2014) Journal of Medical Internet Research. In press

We will extend out methodology to finer spatial resolutions. (Massachusetts and Boston)

Highlights: (a) dynamic-moving training window, (b) automatic feature selection, (c) ensemble approach

Twitter



Crowd-sourced disease surveillance platform

Lu F, Hou S, Baltrusaitis K, Shah M, Leskovec J, Sosic R, Hawkins J, Brownstein JS, Conidi G, Gunn J, ..., Santillana M. Accurate influenza monitoring and forecasting in the Boston metropolis using novel Internet data streams. Journal of Medical Internet Research. 2018;4 (1) :e4.7

Published on 09.01.18 in Vol 4, No 1 (2018): Jan-Mar

This paper is in the following e-collection/theme issue:

Infoveillance, Infodemiology and Digital Disease Surveillance Infodemiology and Infoveillance

Original Paper

Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis

Fred Sun Lu[1], AB ; Suqin Hou[2], MS ; Kristin Baltrusaitis[3], MS ; Manan Shah[4] ; Jure Leskovec[4,5], PhD ; Rok Sosic[4], PhD ; Jared Hawkins[1,6], MMSc, PhD ; John Brownstein[1,6], PhD ; Giuseppe Conidi[7], MPH ; Julia Gunn[7], RN, MPH ; Josh Gray[8], MBA ; Anna Zink[8], BA ; Mauricio Santillana[1,6], MS, PhD

[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, United States

[2]Harvard Chan School of Public Health, Harvard University, Boston, MA, United States

[3]Department of Biostatistics, Boston University School of Public Health, Boston, MA, United States

[4]Computer Science Department, Stanford University, Stanford, CA, United States

[5]Chan Zuckerberg Biohub, San Francisco, CA, United States

[6]Department of Pediatrics, Harvard Medical School, Boston, MA, United States

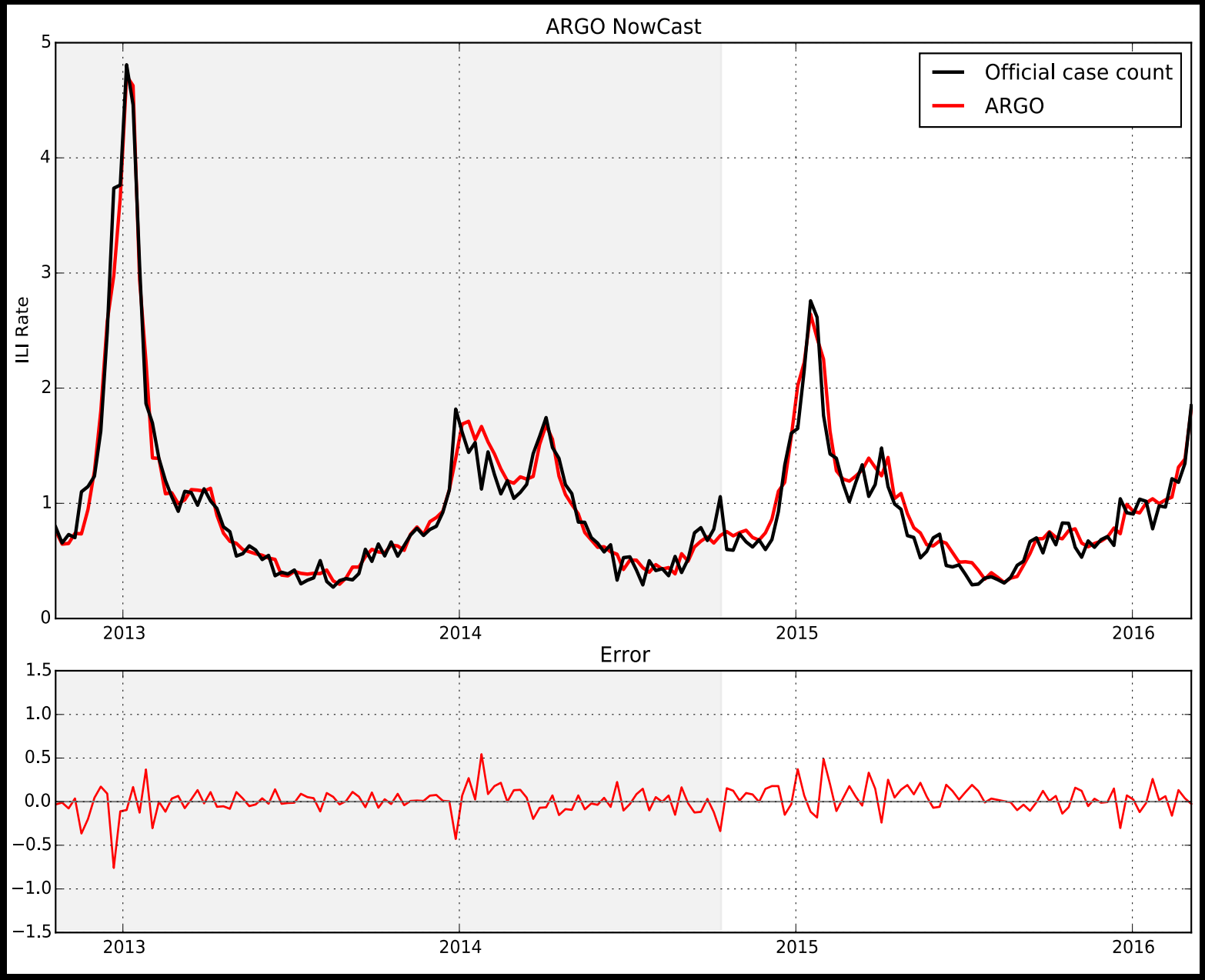
[7]Boston Public Health Commission, Boston, MA, United States

[8]athenaResearch, athenahealth, Watertown, MA, United States

Corresponding Author:

Mauricio Santillana, MS, PhD

Using multiple data sources to track flu in **Boston**

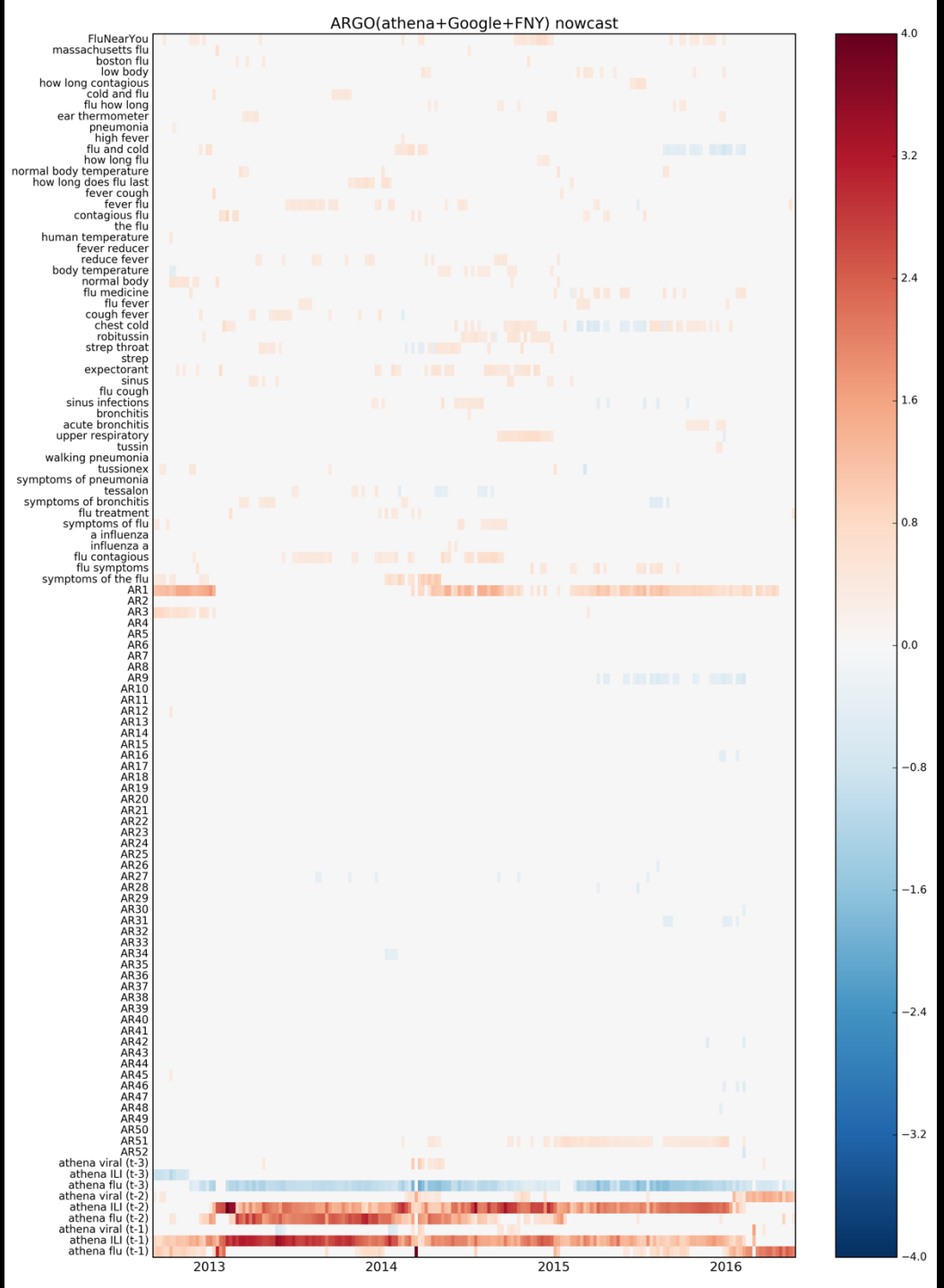


Plots produced by Fred Lu and Suqin Hou

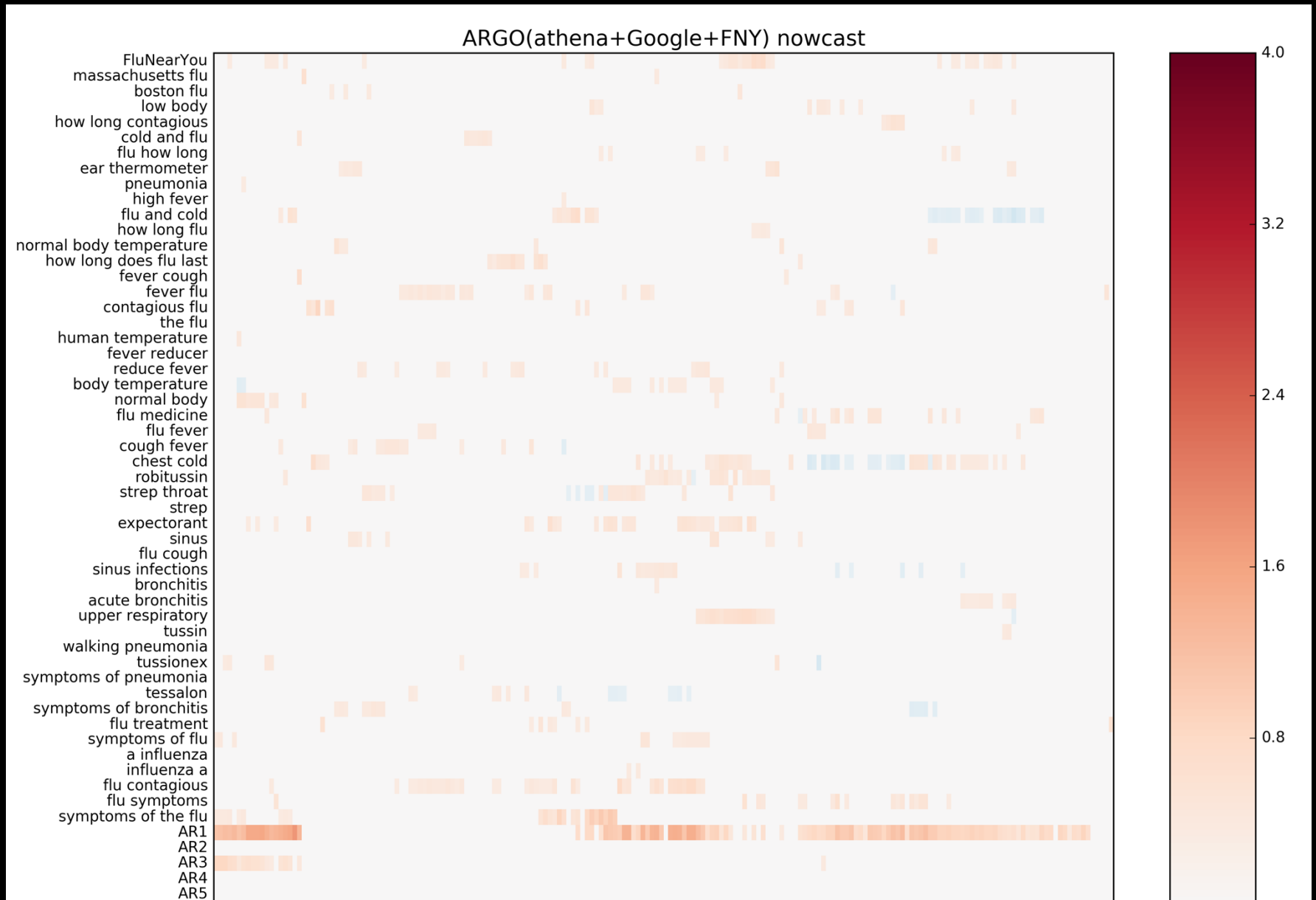
Using multiple data sources to **forecast** flu in **Boston**

Plots produced by Fred Lu and Suqin Hou

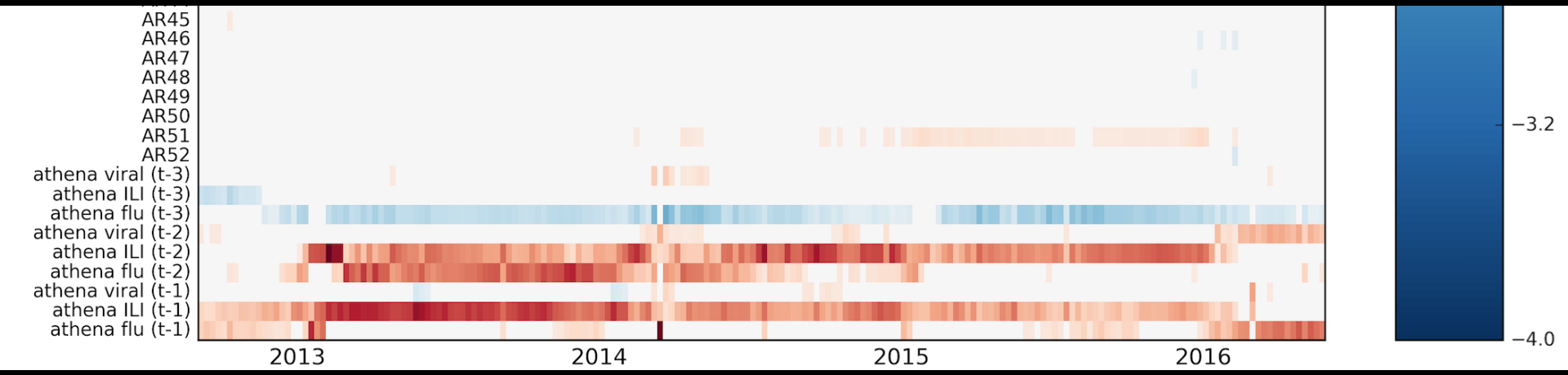
When combined, what are the strongest predictors?



When combined, what are the strongest predictors?



When combined, what are the strongest predictors?



Hyper-local predictions

Can we predict daily emergency department visits in a hospital?

OXFORD ACADEMIC

JAMIA
A SCHOLARLY JOURNAL OF INFORMATICS IN HEALTH AND BIOMEDICINE

Volume 26, Issue 12
December 2019

< Previous Next >

Internet search query data improve forecasts of daily emergency department volume

Sam Tideman ✉, Mauricio Santillana, Jonathan Bickel, Ben Reis

Journal of the American Medical Informatics Association, Volume 26, Issue 12, December 2019, Pages 1574–1583, https://doi.org/10.1093/jamia/ocz154
Published: 17 September 2019 **Article history** ▾

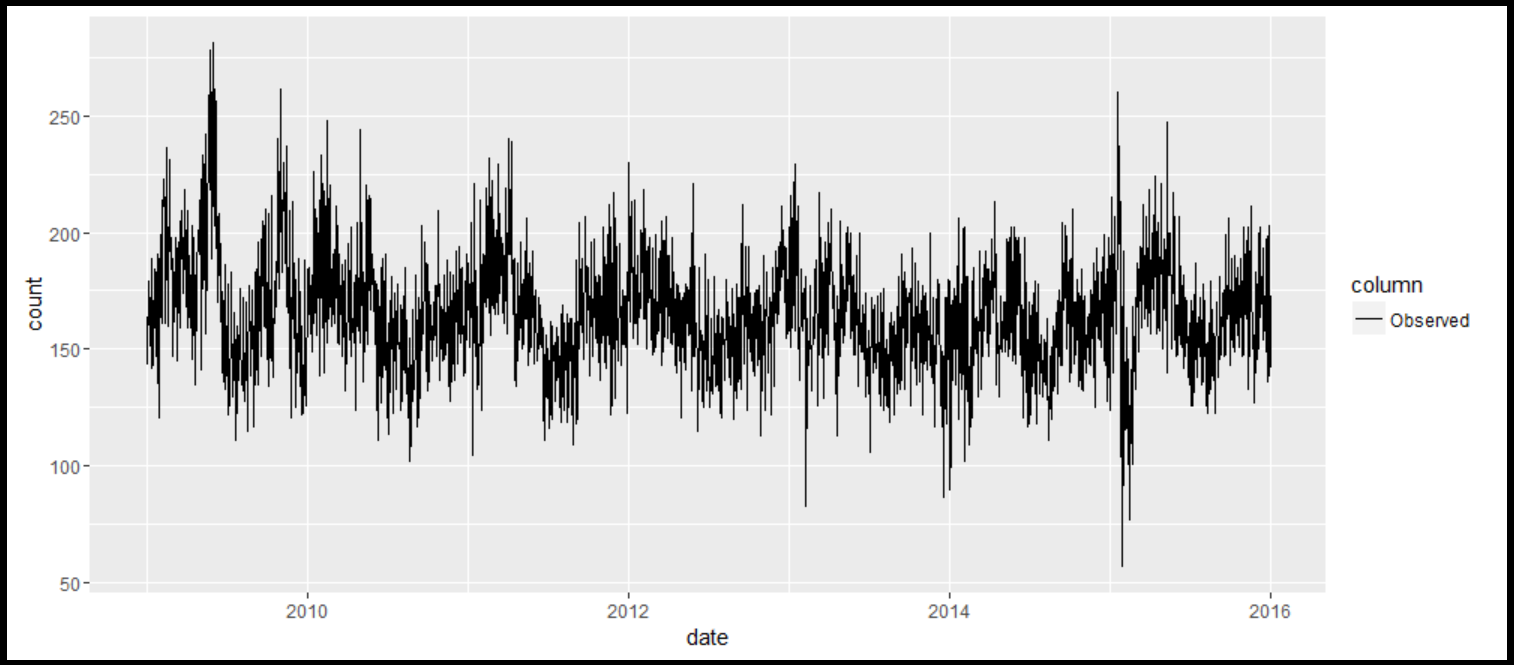
❝ Cite 🔑 Permissions Share ▾

Abstract

Objective

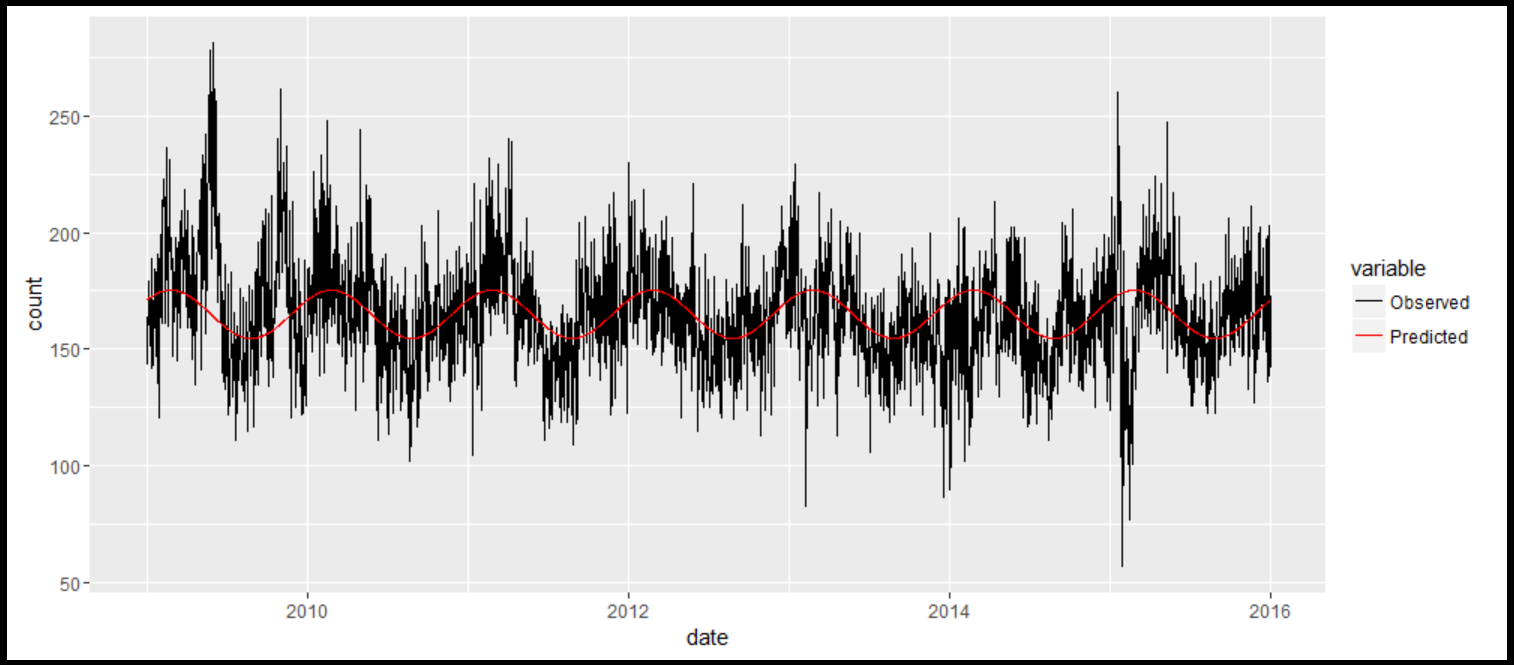
Emergency departments (EDs) are increasingly overcrowded. Forecasting patient visit volume is challenging. Reliable and accurate forecasting strategies may help improve resource allocation and mitigate the effects of overcrowding. Patterns related to weather, day of the week, season, and holidays have been previously used to forecast ED visits. Internet search activity has proven useful for predicting disease trends and offers a new opportunity to improve ED visit forecasting. This study tests whether Google search data and relevant statistical methods can improve the accuracy of ED volume forecasting compared with traditional data sources.

Daily Visits 2009-2015



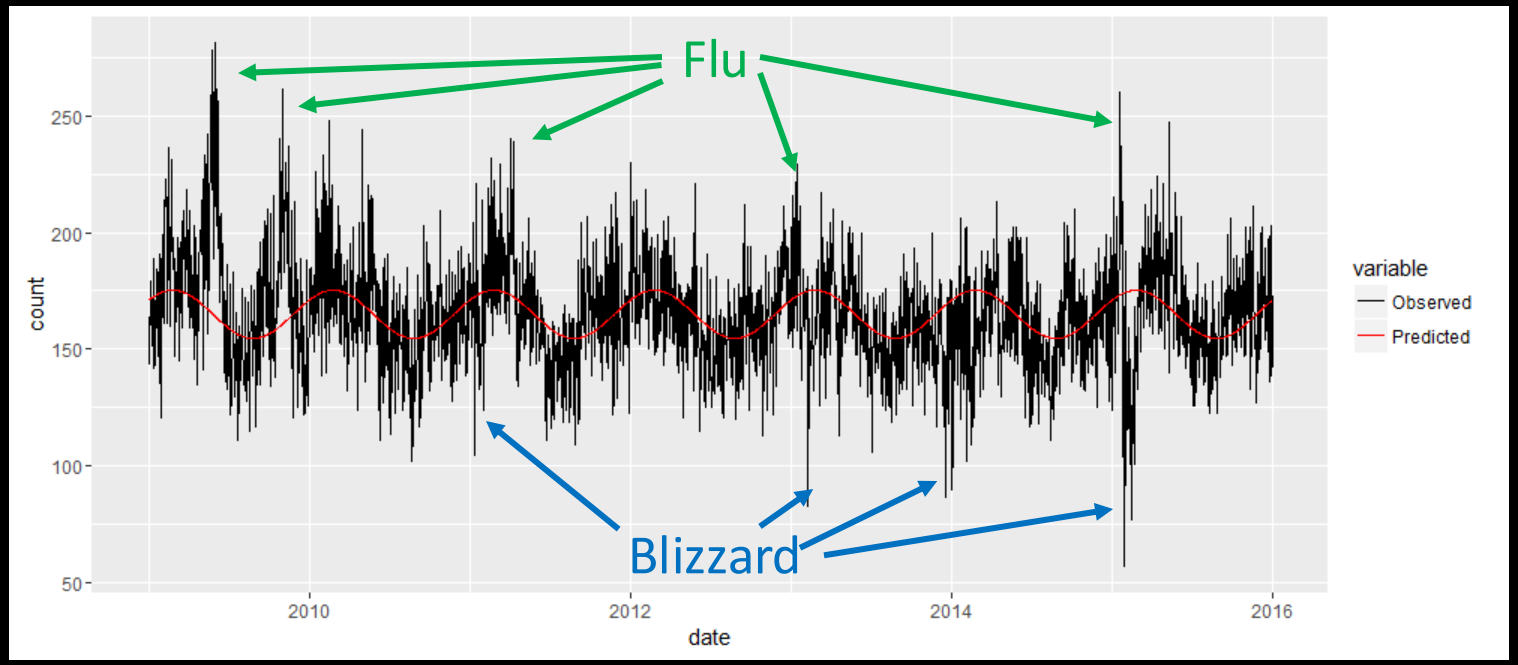
In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Seasonal Trend



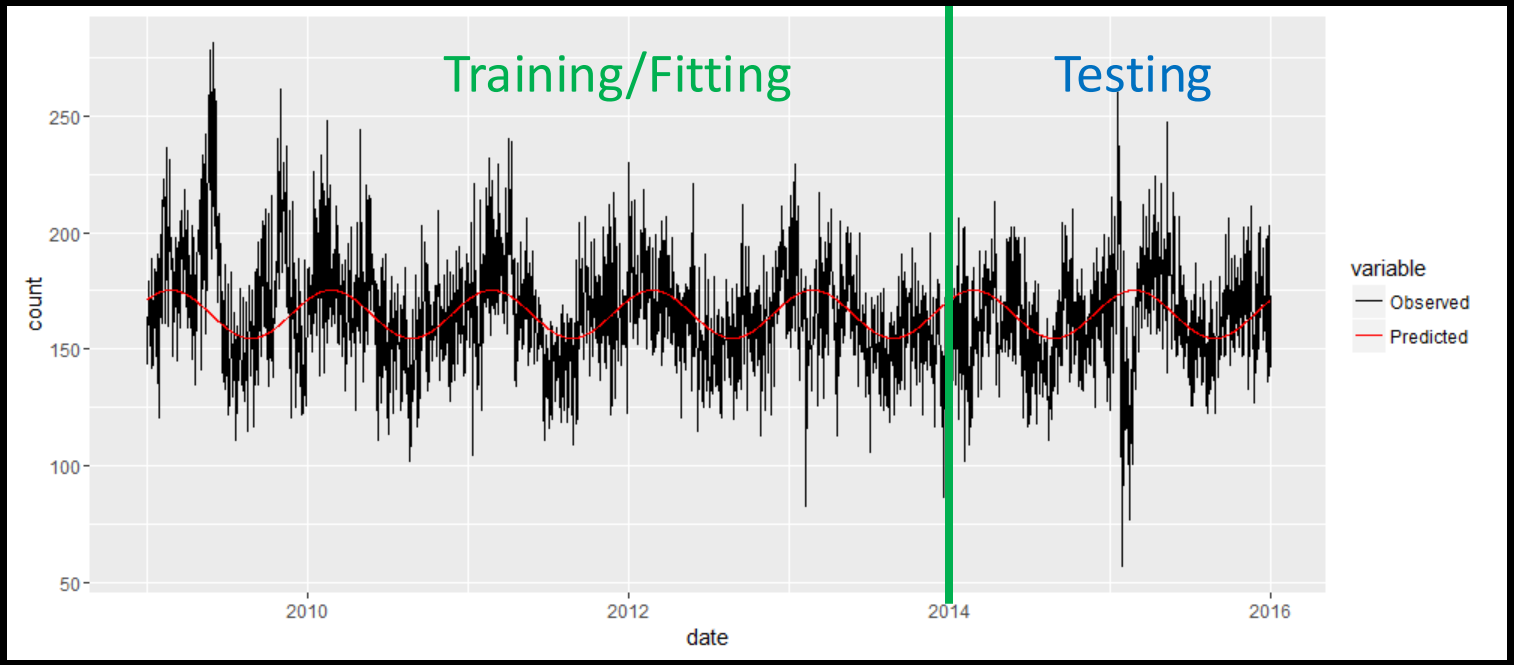
In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Noticeable Events



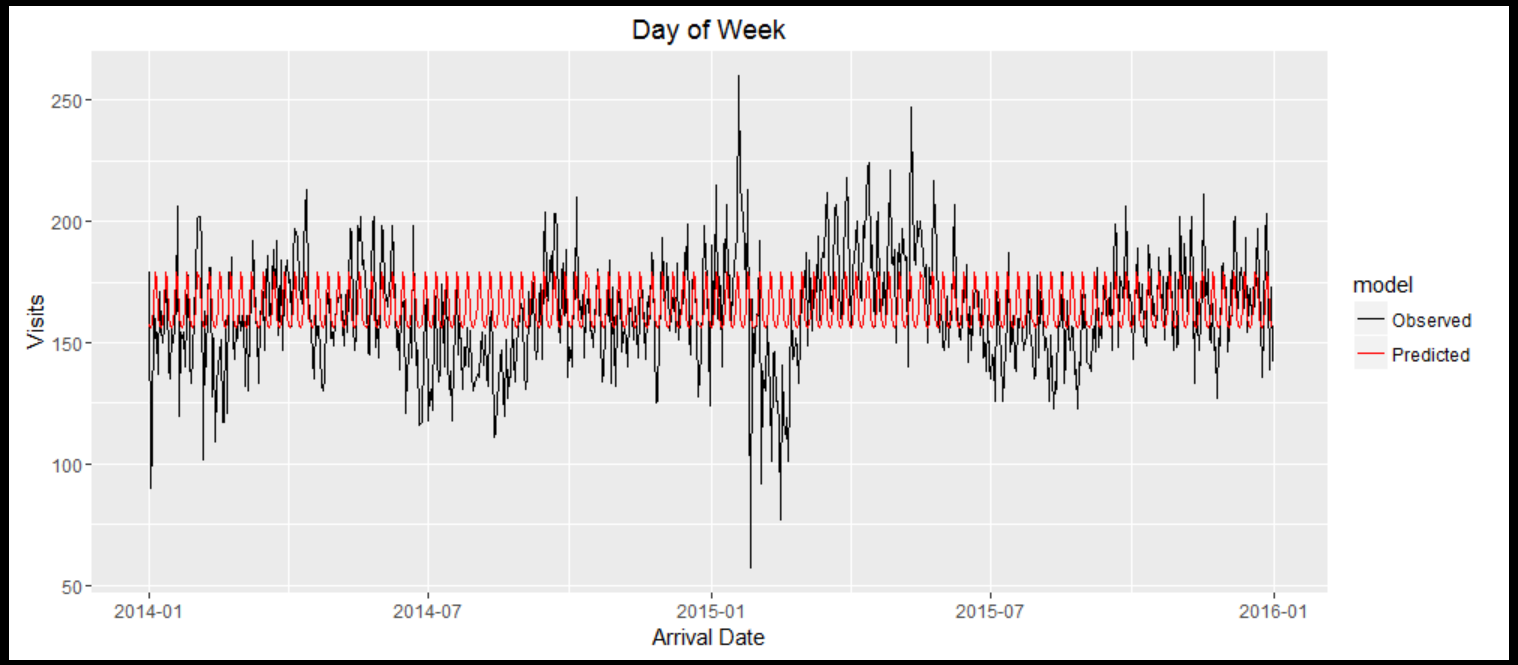
In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Split data for modeling



In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Current Staffing model = Day of Week

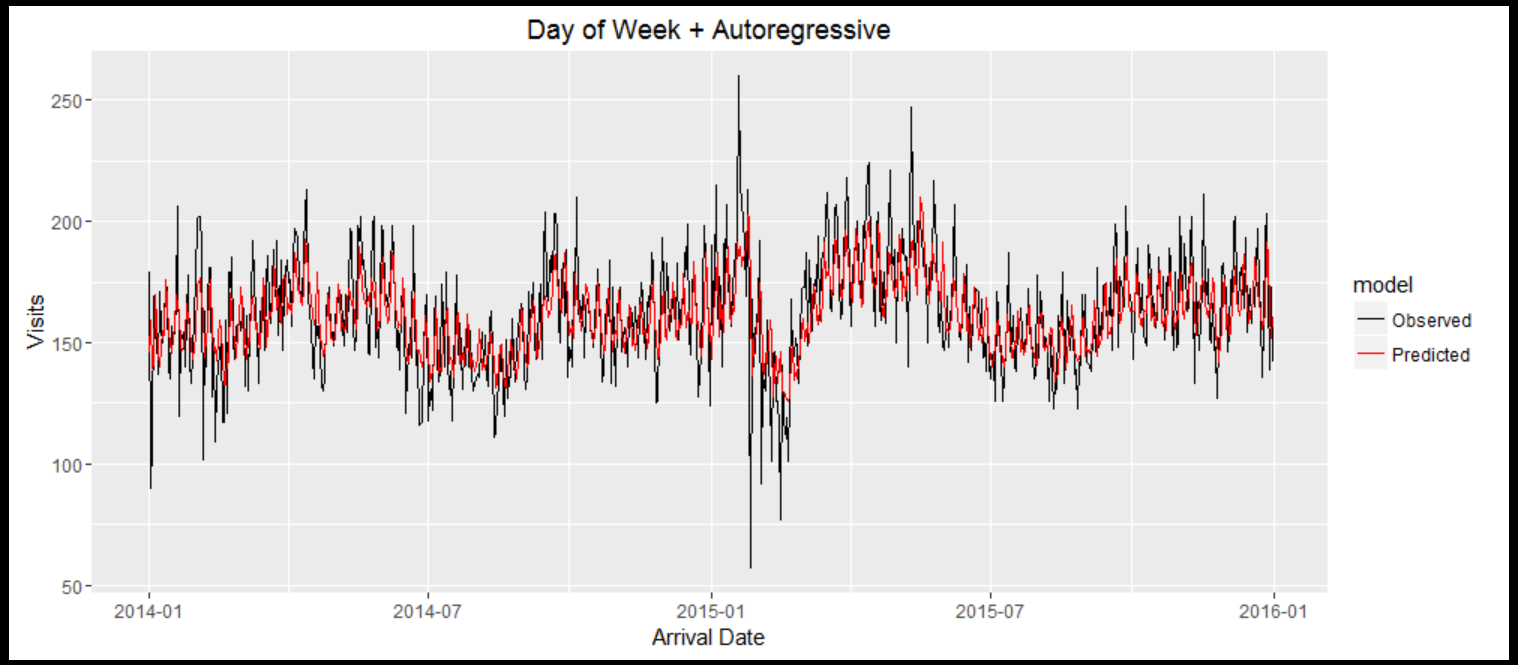


MAPE = 11.0%

Percent of days with bad staffing= 11.2%

In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Add in Auto regression

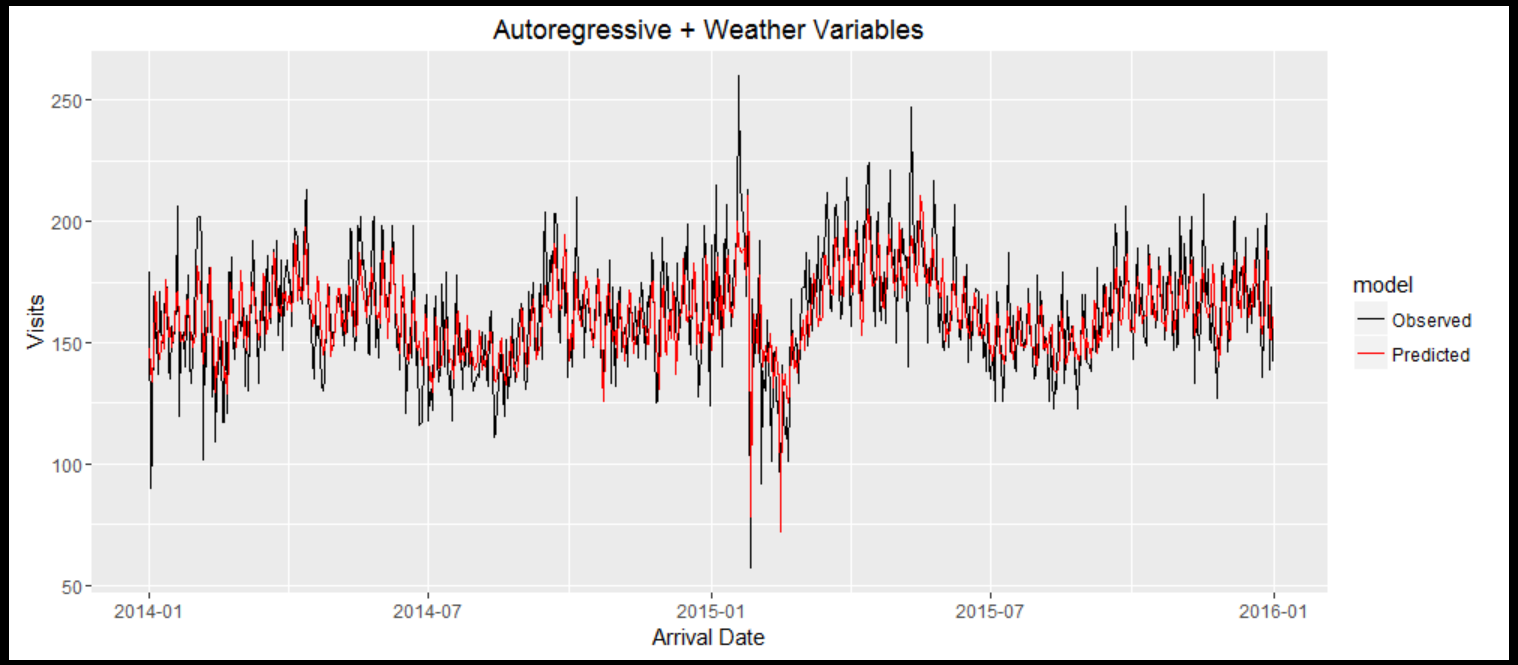


MAPE = 8.4%

Percent of days with bad staffing= 4.9%

In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Add in Weather Data

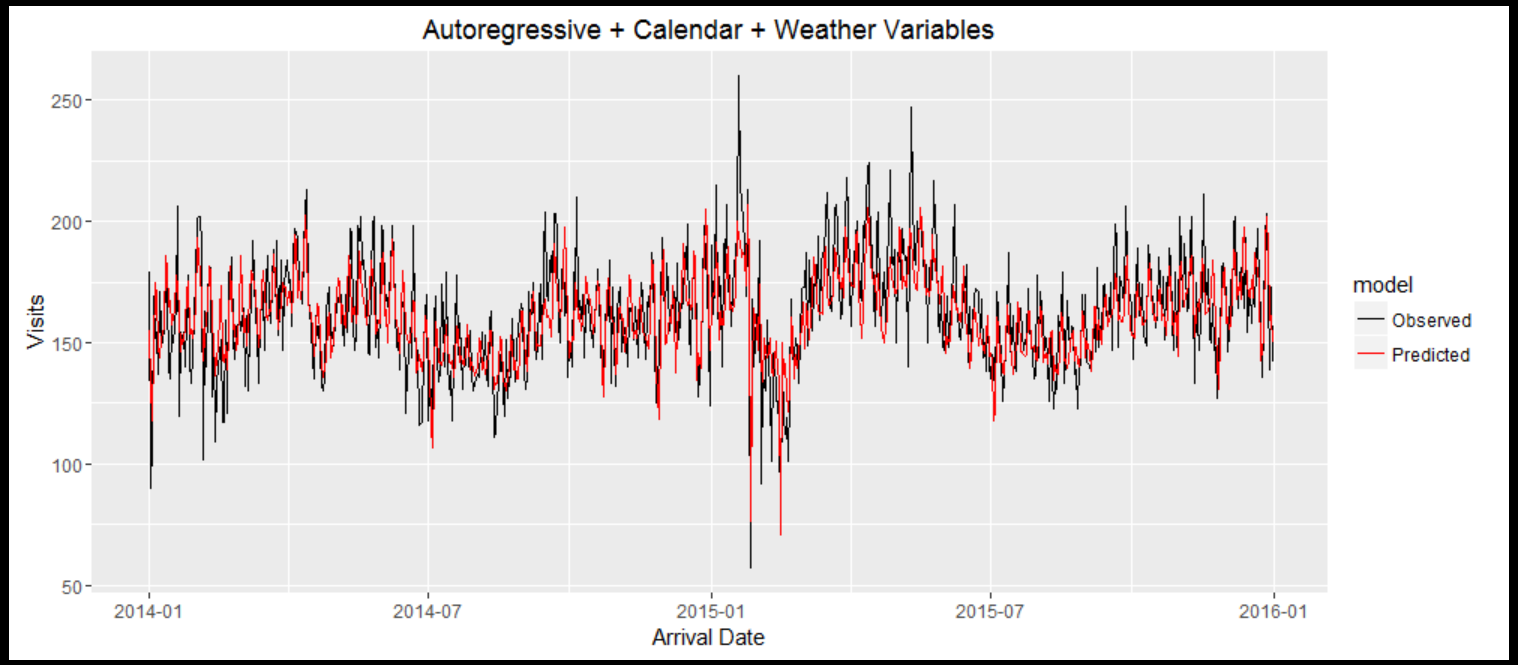


MAPE = 7.9%

Percent of days with bad staffing= 4.8%

In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Add in Calendar Data

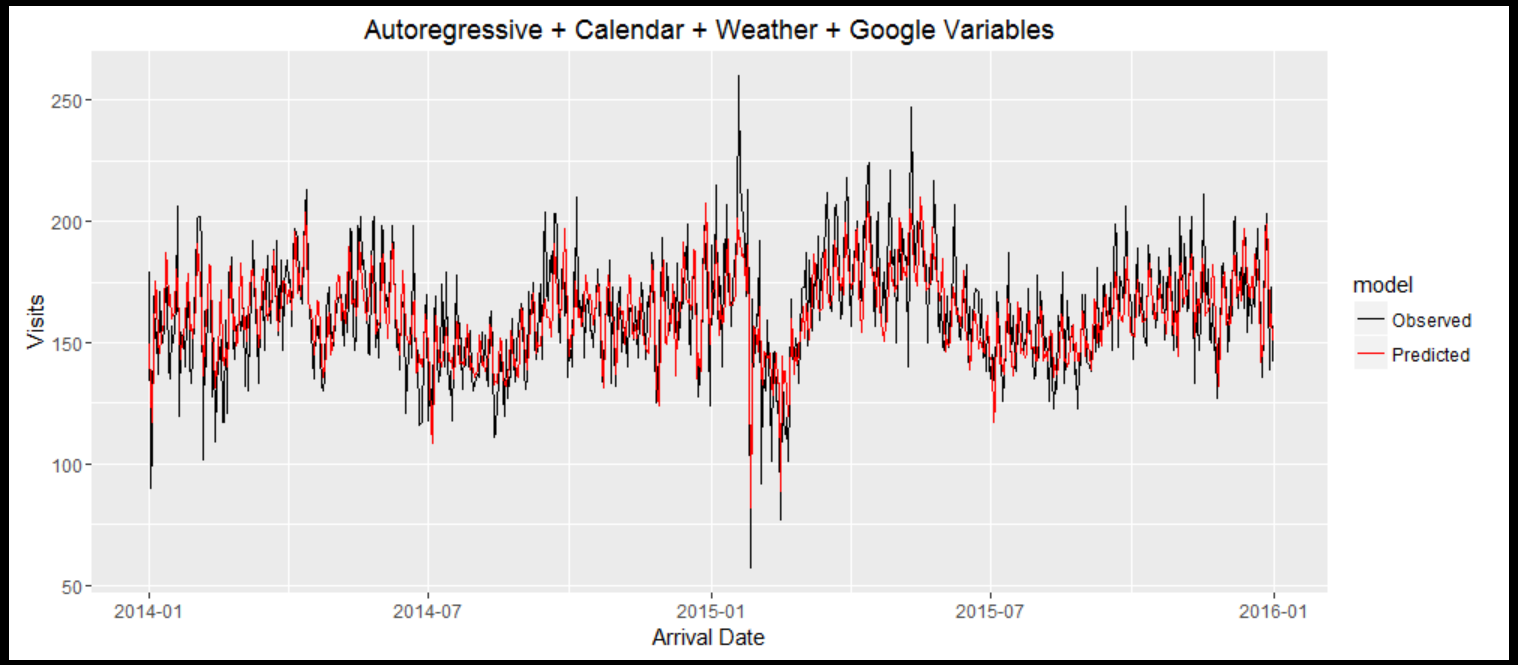


MAPE = 7.7%

Percent of days with bad staffing= 3.8%

In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Add in Google Data



MAPE = 7.6%

Percent of days with bad staffing= 3.3%

In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Toward the use of neural networks for influenza prediction at multiple spatial resolutions

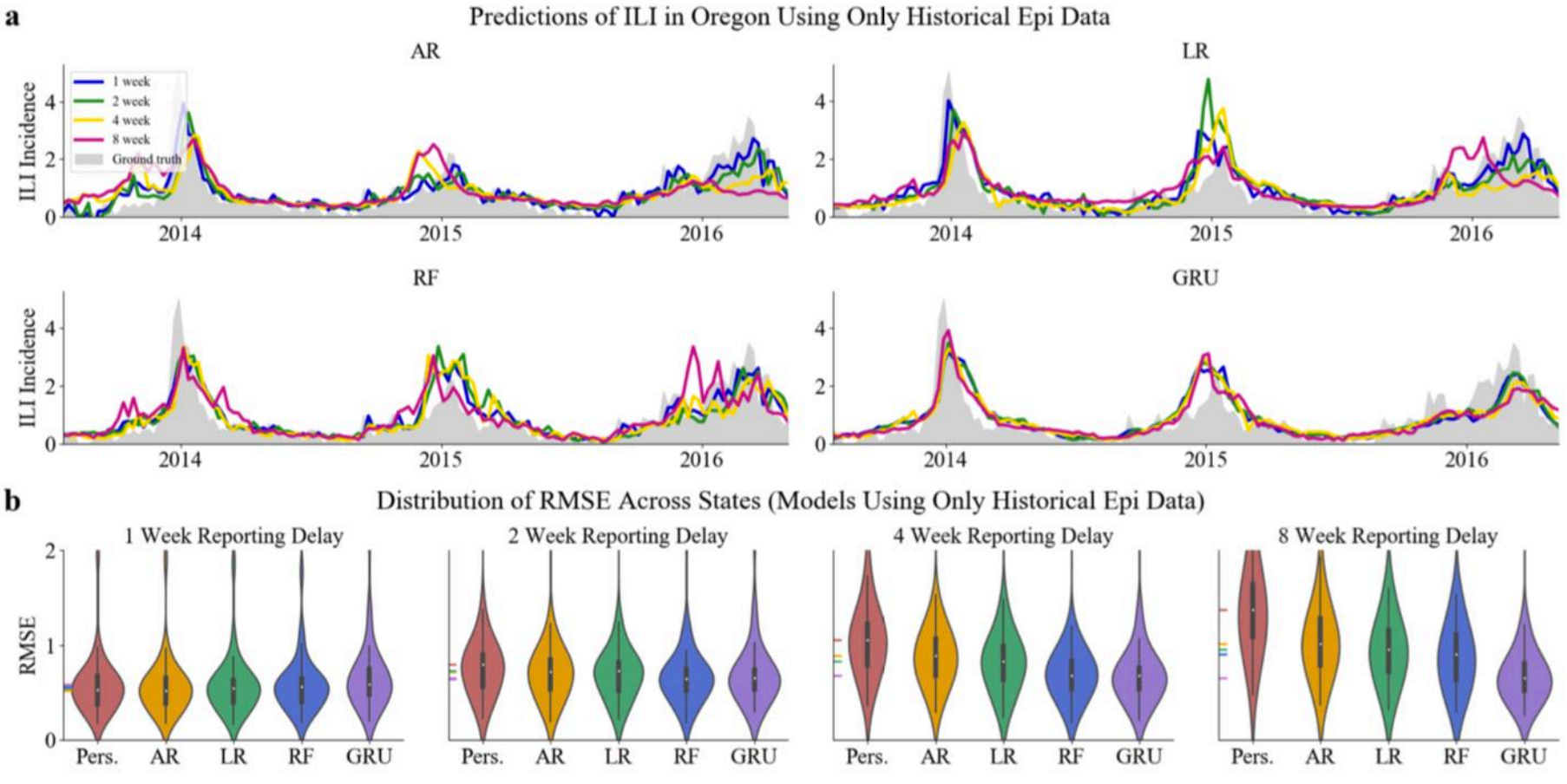
Emily L. Aiken[1,*], Andre T. Nguyen[2,3], Cecile Viboud[4] and Mauricio Santillana[1,5,6,*]
+ See all authors and affiliations

Science Advances 16 Jun 2021:
Vol. 7, no. 25, eabb1237
DOI: 10.1126/sciadv.abb1237

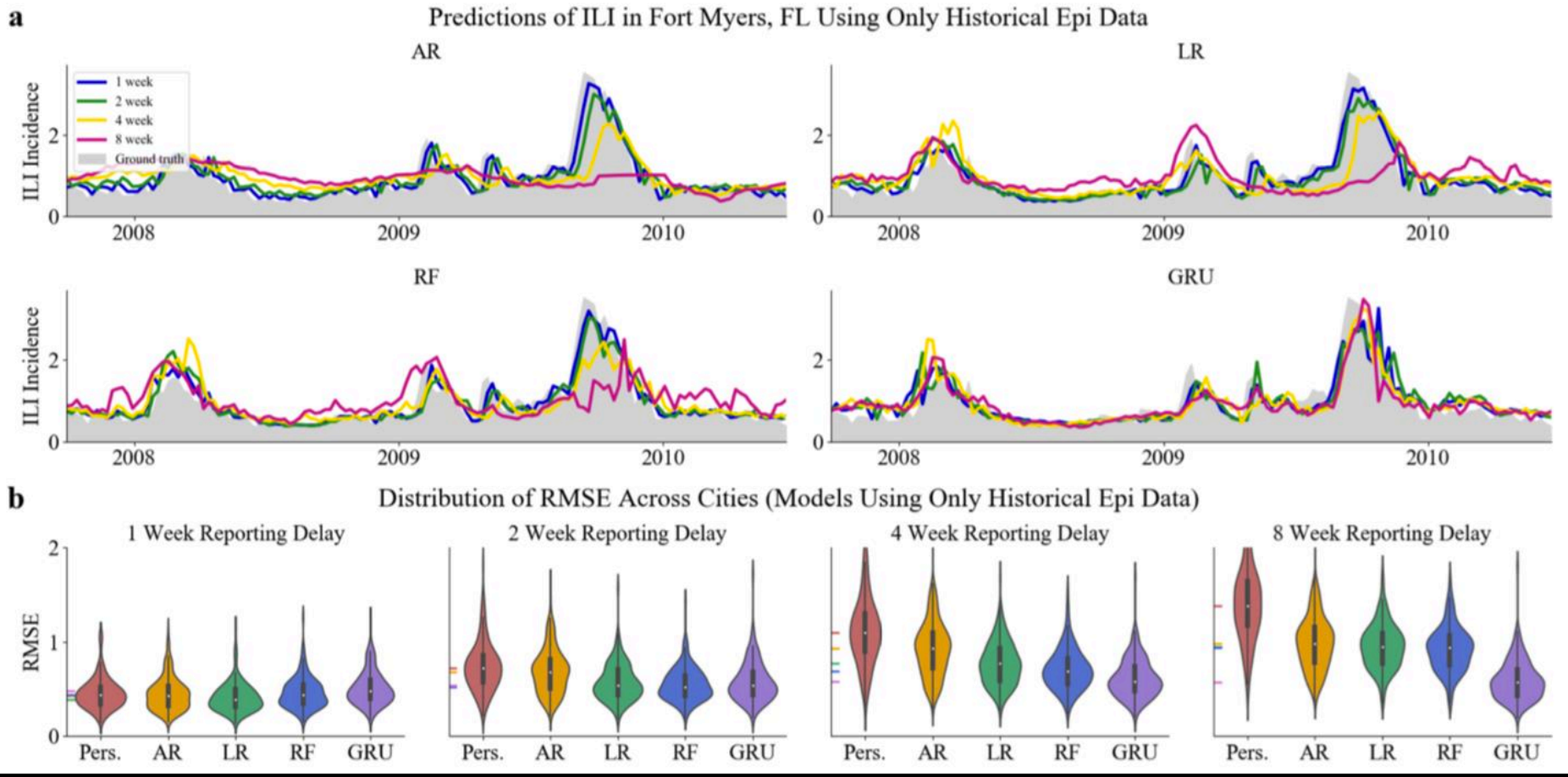
Abstract

Mitigating the effects of disease outbreaks with timely and effective interventions requires accurate real-time surveillance and forecasting of disease activity, but traditional health care–based surveillance systems are limited by inherent reporting delays. Machine learning methods have the potential to fill this temporal "data gap," but work to date in this area has focused on relatively simple methods and coarse geographic resolutions (state level and above). We evaluate the predictive performance of a gated recurrent unit neural network approach in comparison with baseline machine learning methods for estimating influenza activity in the United States at the state and city levels and experiment with the inclusion of real-time Internet search data. We find that the neural network approach improves upon baseline models for long time horizons of prediction but is not improved by real-time internet search data. We conduct a thorough analysis of feature importances in all considered models for interpretability purposes.

State-level predictions

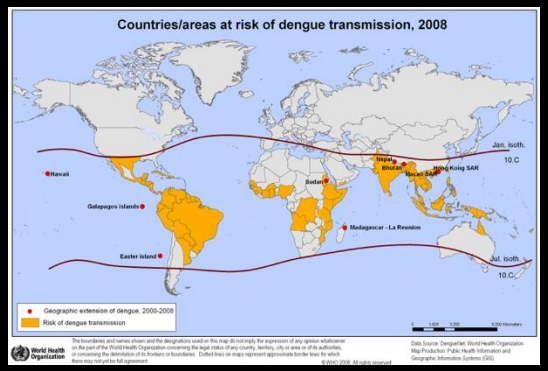


City-level predictions



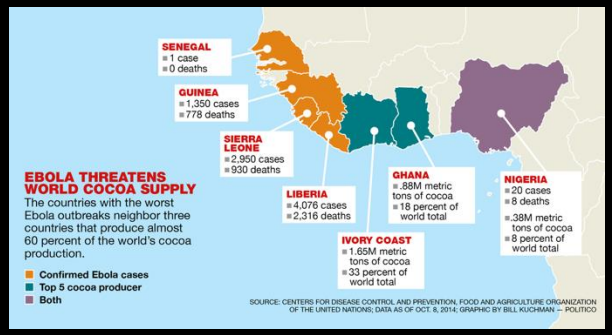
Part 2. Success stories in tracking and forecasting Flu, Zika, Dengue, Ebola in data-poor medium- to low-income countries.

Dengue, Zika, and Flu, AMR



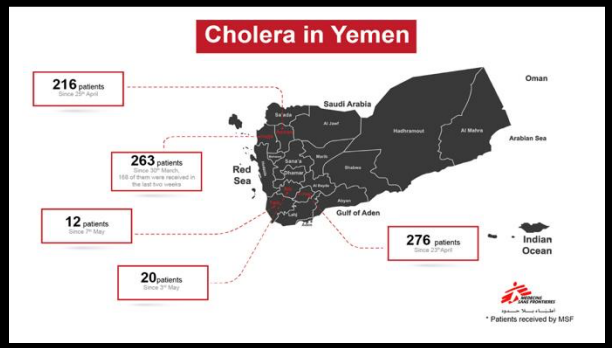
- Latin America (Flu, Zika, Dengue)
- South-east Asia (Dengue)

Ebola



- West Africa

Cholera



- Middle East

Can these methodologies yield accurate estimates of flu in Low to middle income countries?
Yes, in selected countries where enough historical flu activity has been recorded over time

Latin America

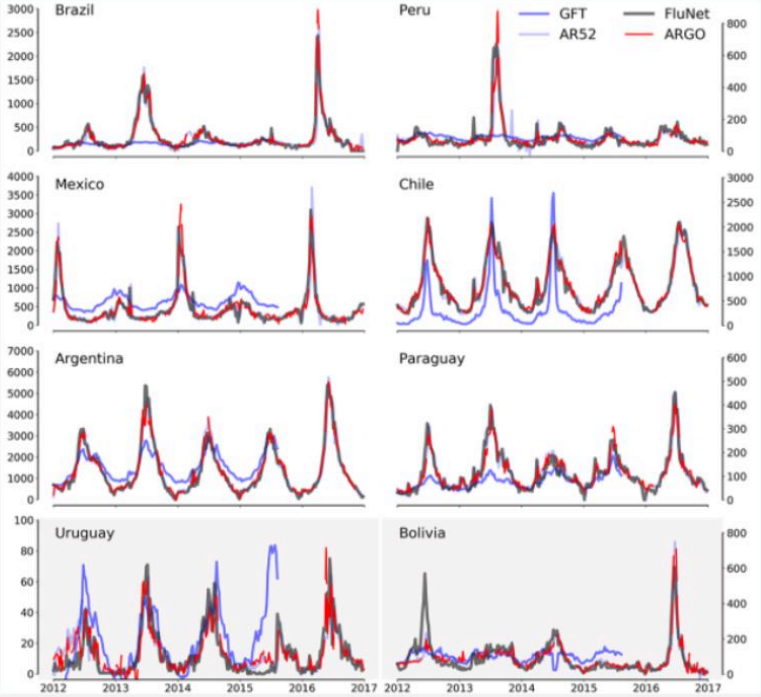


Figure 1. Graphical representation of the number of processed specimens (NPSs) as reported by WHO's FluNet (black), along with the NPS estimates generated by ARGO (red), AR (light blue), and Google Flu Trends (GFT; blue), over the whole study period of January 1, 2012 to December 25, 2016.
View this figure

JMIR Publications
Advancing Digital Health & Open Science

Articles | Search articles

JMIR Public Health and Surveillance | Journal Information | Browse Journal | Subm

Published on 4.4.2019 in Vol 5, No 2 (2019): Apr-Jun

Preprints (earlier versions) of this paper are available at https://preprints.jmir.org/preprint/12214, first published September 14, 2018.

Improved Real-Time Influenza Surveillance: Using Internet Search Data in Eight Latin American Countries

Leonardo Clemente [1,2]; Fred Lu [2]; Mauricio Santillana [2,3]

Article | Authors | Cited by | Tweetations | Metrics

- Abstract
- Introduction
- Methods
- Results
- Discussion
- Abbreviations
- Copyright

Abstract

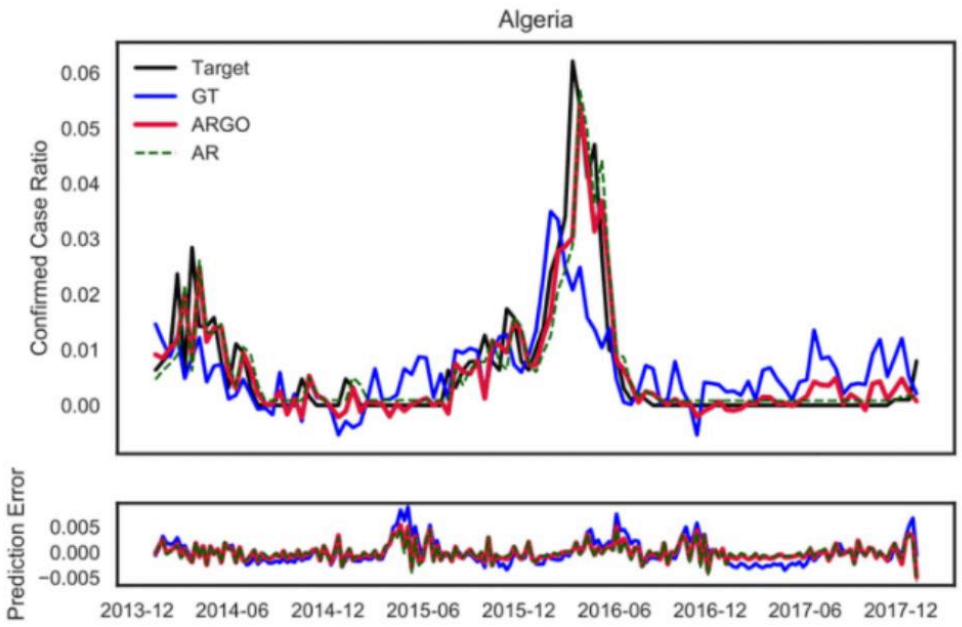
Background:
Novel influenza surveillance systems that leverage Internet-based real-time data sources including Internet search frequencies, social-network information, and crowd-sourced flu surveillance tools have shown improved accuracy over the past few years in data-rich countries like the United States. These systems not only track flu activity accurately, but they also report flu estimates a week or more ahead of the publication of reports produced by healthcare-based systems, such as those implemented and managed by the Centers for Disease Control and Prevention. Previous work has shown that the predictive capabilities of novel flu surveillance systems, like Google Flu Trends

https://publichealth.jmir.org/2019/2/e12214/

Can these methodologies yield accurate estimates of flu in Low to middle income countries?

Yes, in selected countries where enough historical flu activity has been recorded over time

Africa



https://gatesopenresearch.org/articles/3-1653

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Advances in using Internet searches to track dengue

Shihao Yang, Samuel C. Kou, Fred Lu, John S. Brownstein, Nicholas Brooke, Mauricio Santillana

Published: July 20, 2017 • https://doi.org/10.1371/journal.pcbi.1005607

Abstract

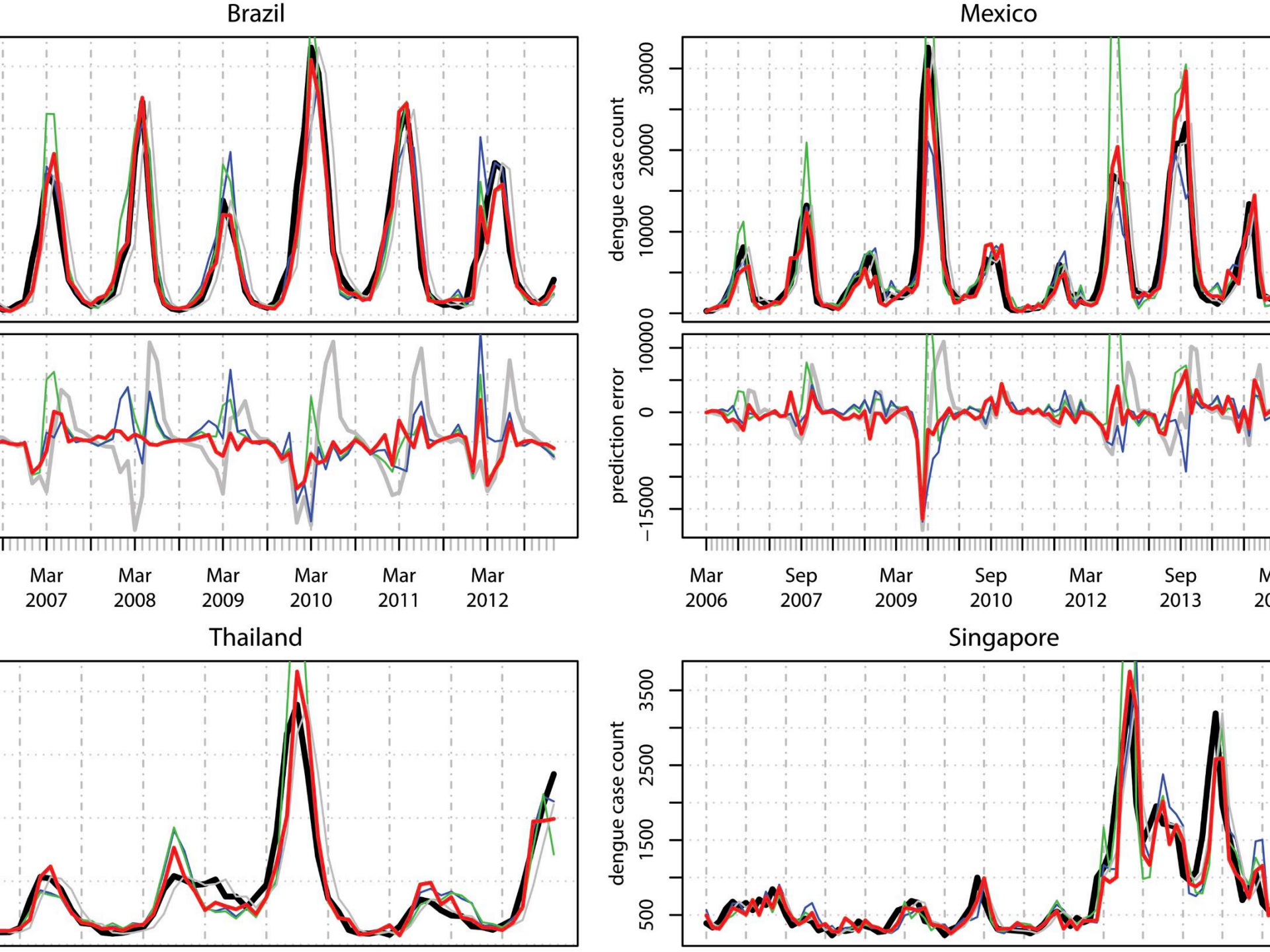
Dengue is a mosquito-borne disease that threatens over half of the world's population. Despite being endemic to more than 100 countries, government-led efforts and tools for timely identification and tracking of new infections are still lacking in many affected areas. Multiple methodologies that leverage the use of Internet-based data sources have been proposed as a way to complement dengue surveillance efforts. Among these, dengue-related Google search trends have been shown to correlate with dengue activity. We extend a methodological framework, initially proposed and validated for flu surveillance, to produce near real-time estimates of dengue cases in five countries/states: Mexico, Brazil, Thailand, Singapore and Taiwan. Our result shows that our modeling framework can be used to improve the tracking of dengue activity in multiple locations around the world.

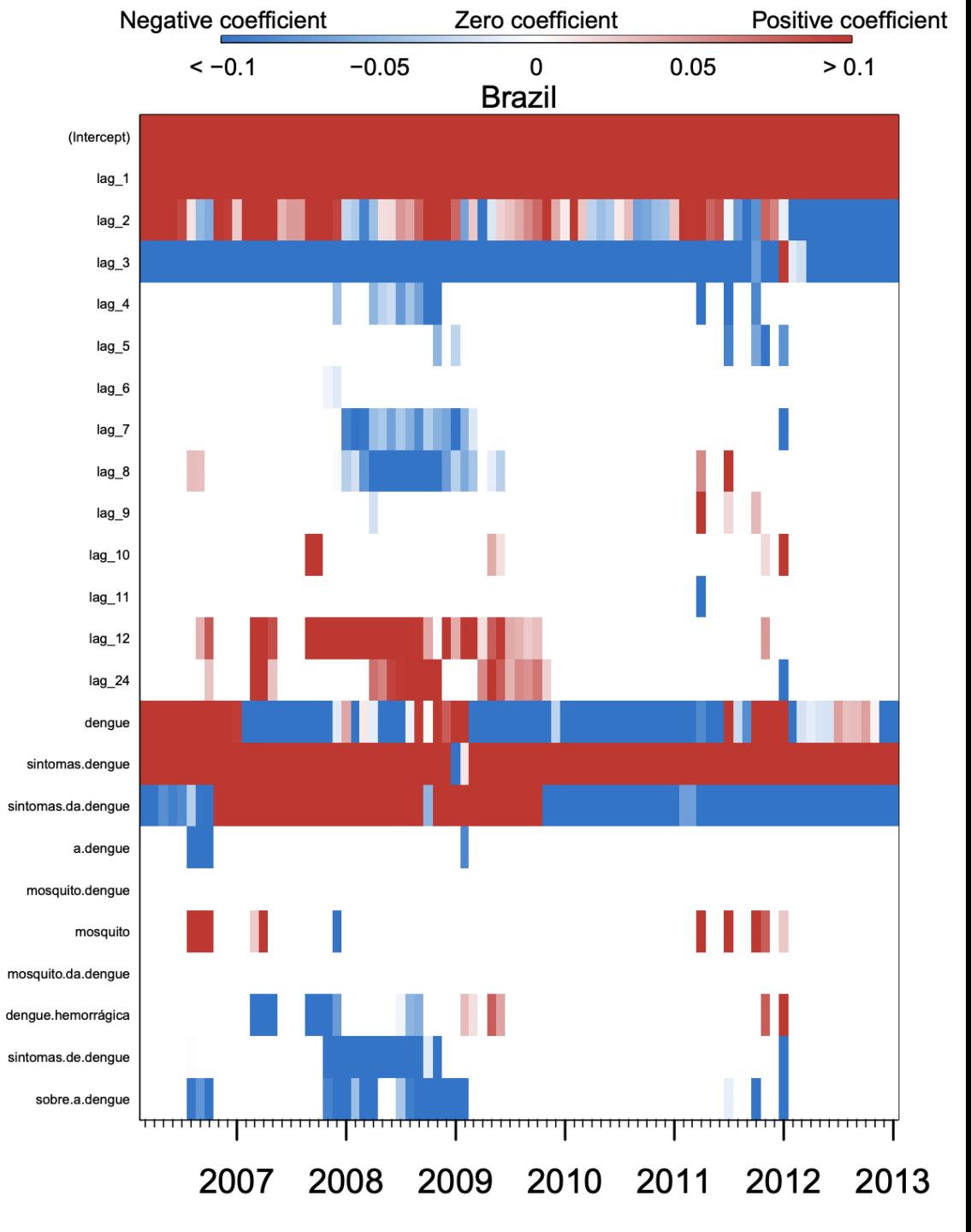
Author summary

As communicable diseases spread in our societies, people frequently turn to the Internet to search for medical information. In recent years, multiple research teams have investigated how to utilize Internet users' search activity to track infectious diseases around our planet. In this article, we show that a methodology, originally developed to track flu in the US, can be extended to improve dengue surveillance in multiple countries/states where dengue has been observed in the last several years. Our result suggests that our methodology performs best in dengue-endemic areas with high number of yearly cases and with sustained seasonal incidence.

Table A. Query terms used for each country/state

Brazil	Mexico	Thailand	Singapore	Taiwan
dengue	dengue	โรคไข้เลือดออก	dengue	登革熱
sintomas.dengue	dengue.dengue.dengue	อาการ.โรค.ไข้เลือดออก	dengue.fever	登革熱噴藥
mosquito	el.dengue	ไข้เลือดออก	dengue.symptoms	出血性登革熱
sintomas.da.dengue	dengue.sintomas	โรค.ไข้เลือดออก	dengue.singapore	埃及斑蚊
a.dengue	sintomas.del.dengue	การ.ป้องกัน.ไข้เลือดออก	symptoms.dengue.fever	登格熱
mosquito.dengue	dengue.hemorragico	อาการ.ของ.ไข้เลือดออก	symptoms.of.dengue	防蚊液
mosquito.da.dengue	sintomas.de.dengue	สาเหตุ.ไข้เลือดออก	dengue.fever.singapore	白線斑蚊
dengue.hemorrágica	que.es.dengue	โครงการ.ไข้เลือดออก	dengue.mosquito	登革樂
sintomas.de.dengue	dengue.clasico	สถานการณ์.โรค.ไข้เลือดออก	mosquito	dengue fever
sobre.a.dengue	dengue.mosquito	สถานการณ์.ไข้เลือดออก	dengue.in.singapore	蚊子叮


Brazil
Mexico
dengue case count
prediction error
30000
20000
10000
100000
0
−15000
Mar 2007
Mar 2008
Mar 2009
Mar 2010
Mar 2011
Mar 2012
Mar 2006
Sep 2007
Mar 2009
Sep 2010
Mar 2012
Sep 2013
Thailand
Singapore
3500
2500
1500
500


Negative coefficient
Zero coefficient
Positive coefficient
< −0.1
−0.05
0
0.05
> 0.1
Brazil
(Intercept)
lag_1
lag_2
lag_3
lag_4
lag_5
lag_6
lag_7
lag_8
lag_9
lag_10
lag_11
lag_12
lag_24
dengue
sintomas.dengue
sintomas.da.dengue
a.dengue
mosquito.dengue
mosquito
mosquito.da.dengue
dengue.hemorrágica
sintomas.de.dengue
sobre.a.dengue
2007
2008
2009
2010
2011
2012
2013

PLOS NEGLECTED TROPICAL DISEASES

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Predicting dengue incidence leveraging internet-based data sources. A case study in 20 cities in Brazil

Gal Koplewitz, Fred Lu, Leonardo Clemente, Caroline Buckee, Mauricio Santillana

Version 2 Published: January 24, 2022 • https://doi.org/10.1371/journal.pntd.0010071

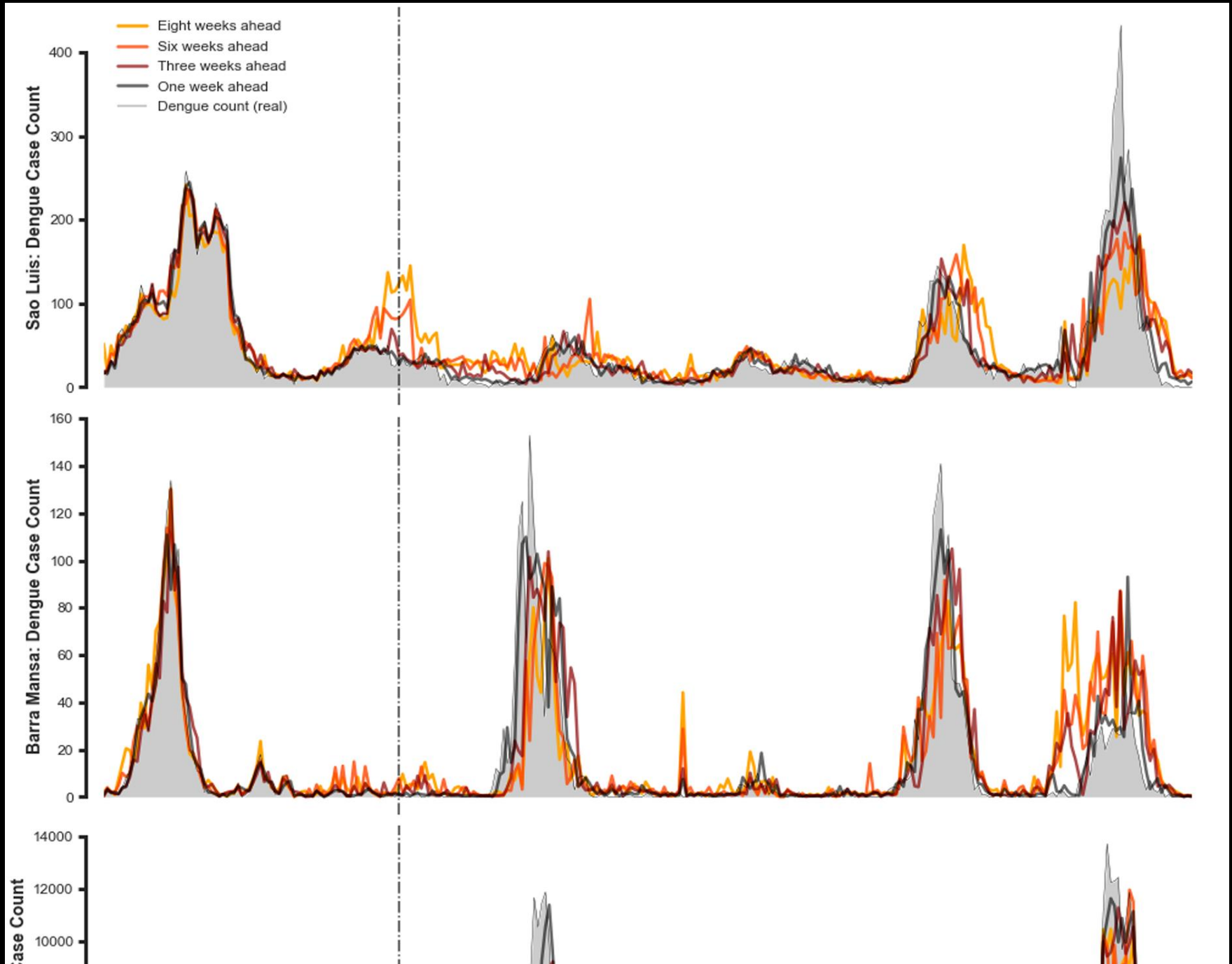
See the preprint

Article | Authors | Metrics | Comments | Media Coverage

Abstract
Author summary
Introduction
Materials and methods
Results
Discussion
Supporting information
Acknowledgments
References

Abstract

The dengue virus affects millions of people every year worldwide, causing large epidemic outbreaks that disrupt people's lives and severely strain healthcare systems. In the absence of a reliable vaccine against dengue or an effective treatment to manage the illness in humans, most efforts to combat dengue infections have focused on preventing its vectors, mainly the Aedes aegypti mosquito, from flourishing across the world. These mosquito-control strategies need reliable disease activity surveillance systems to be deployed. Despite significant efforts to estimate dengue incidence using a variety of data sources and methods, little work has been done to understand the relative contribution of the different data sources to improved prediction. Additionally, scholarship on the topic had initially focused on prediction systems at the national- and state-levels, and much remains to be done at the finer spatial resolutions at which health policy interventions often occur. We develop a methodological framework to assess and compare dengue incidence estimates at the city level, and evaluate the performance of a collection of models on 20 different cities in Brazil. The data sources we use towards this end are weekly incidence counts from prior years (seasonal autoregressive terms), weekly-aggregated weather variables, and real-time internet search data. We find that both random forest-based models and LASSO regression-based models effectively leverage these multiple data sources to produce accurate predictions, and that while the performance between them is comparable on average, the former method produces fewer extreme outliers, and can thus be considered more robust. For real-time predictions that assume long delays (6–8 weeks) in the availability of epidemiological data, we find that real-time internet search data


Eight weeks ahead
Six weeks ahead
Three weeks ahead
One week ahead
Dengue count (real)
Sao Luis: Dengue Case Count
0
100
200
300
400
Barra Mansa: Dengue Case Count
0
20
40
60
80
100
120
140
160
Case Count
10000
12000
14000

PLOS | NEGLECTED TROPICAL DISEASES

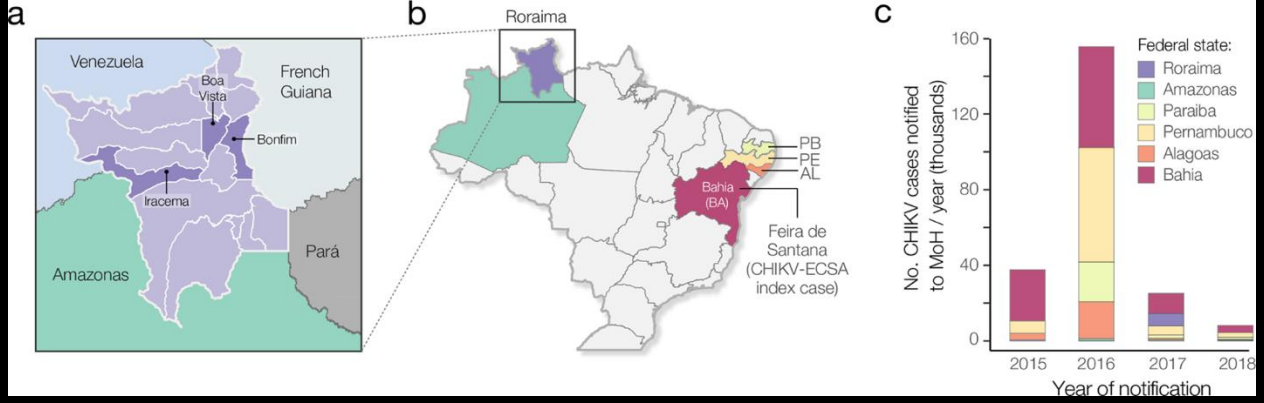
OPEN ACCESS PEER-REVIEWED

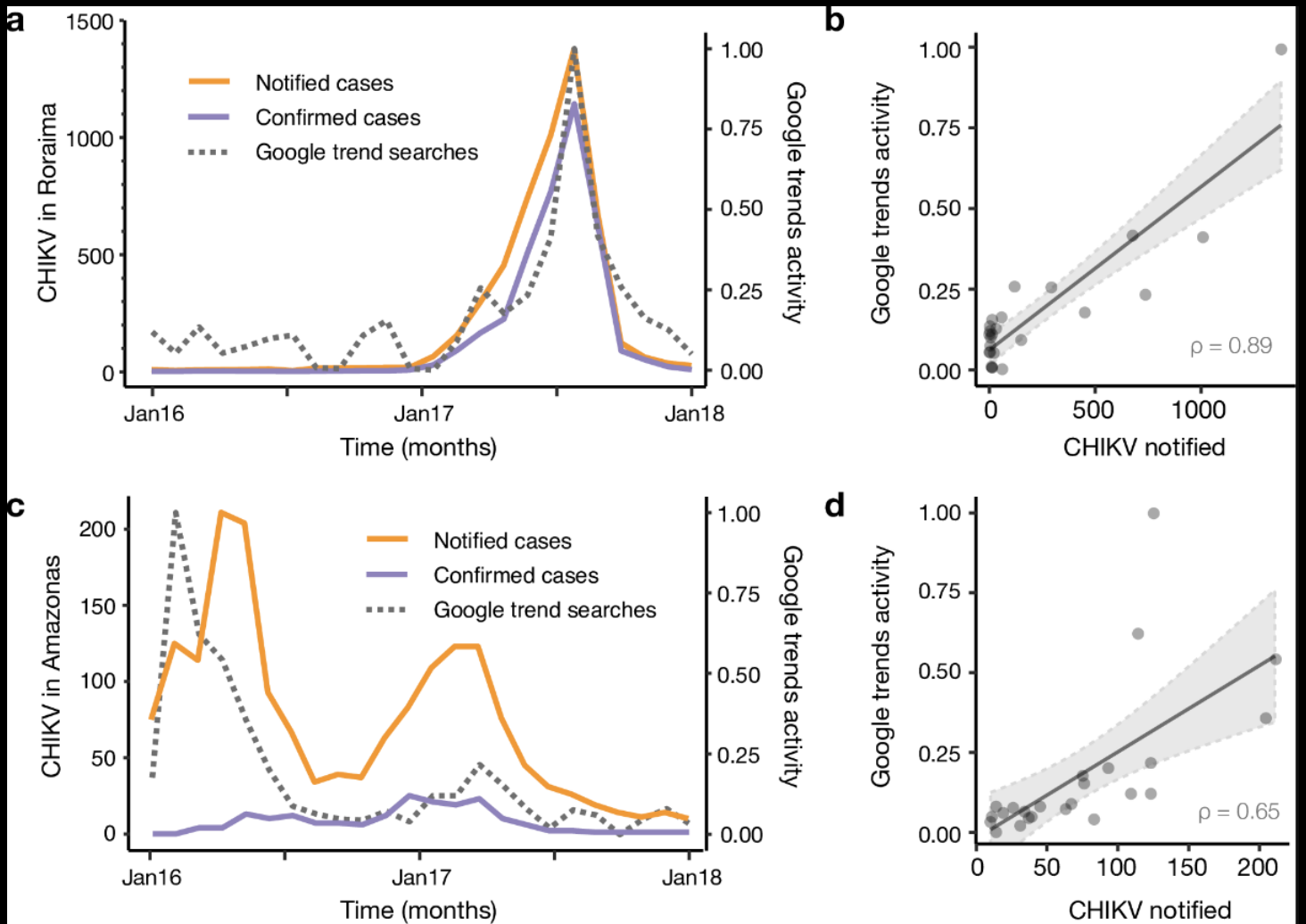
RESEARCH ARTICLE

Genomic, epidemiological and digital surveillance of Chikungunya virus in the Brazilian Amazon

Felipe Gomes Naveca, Ingra Claro, Marta Giovanetti, Jaqueline Goes de Jesus, Joilson Xavier, Felipe Campos de Melo Iani, Valdinete Alves do Nascimento, Victor Costa de Souza, Paola Paz Silveira, José Lourenço, Mauricio Santillana, Moritz U. G. Kraemer, Josh Quick, [...], Nuno Rodrigues Faria [view all]

Published: March 7, 2019 • https://doi.org/10.1371/journal.pntd.0007065


a
Venezuela
French Guiana
Boa Vista
Bonfim
Iracema
Amazonas
Pará
b
Roraima
PB
PE
AL
Bahia (BA)
Feira de Santana (CHIKV-ECSA index case)
c
No. CHIKV cases notified to MoH / year (thousands)
Federal state:
Roraima
Amazonas
Paraiba
Pernambuco
Alagoas
Bahia
2015
2016
2017
2018
Year of notification


a
CHIKV in Roraima
Notified cases
Confirmed cases
Google trend searches
Google trends activity
Jan16
Jan17
Jan18
Time (months)
b
Google trends activity
ρ = 0.89
CHIKV notified
c
CHIKV in Amazonas
Notified cases
Confirmed cases
Google trend searches
Google trends activity
Jan16
Jan17
Jan18
Time (months)
d
Google trends activity
ρ = 0.65
CHIKV notified

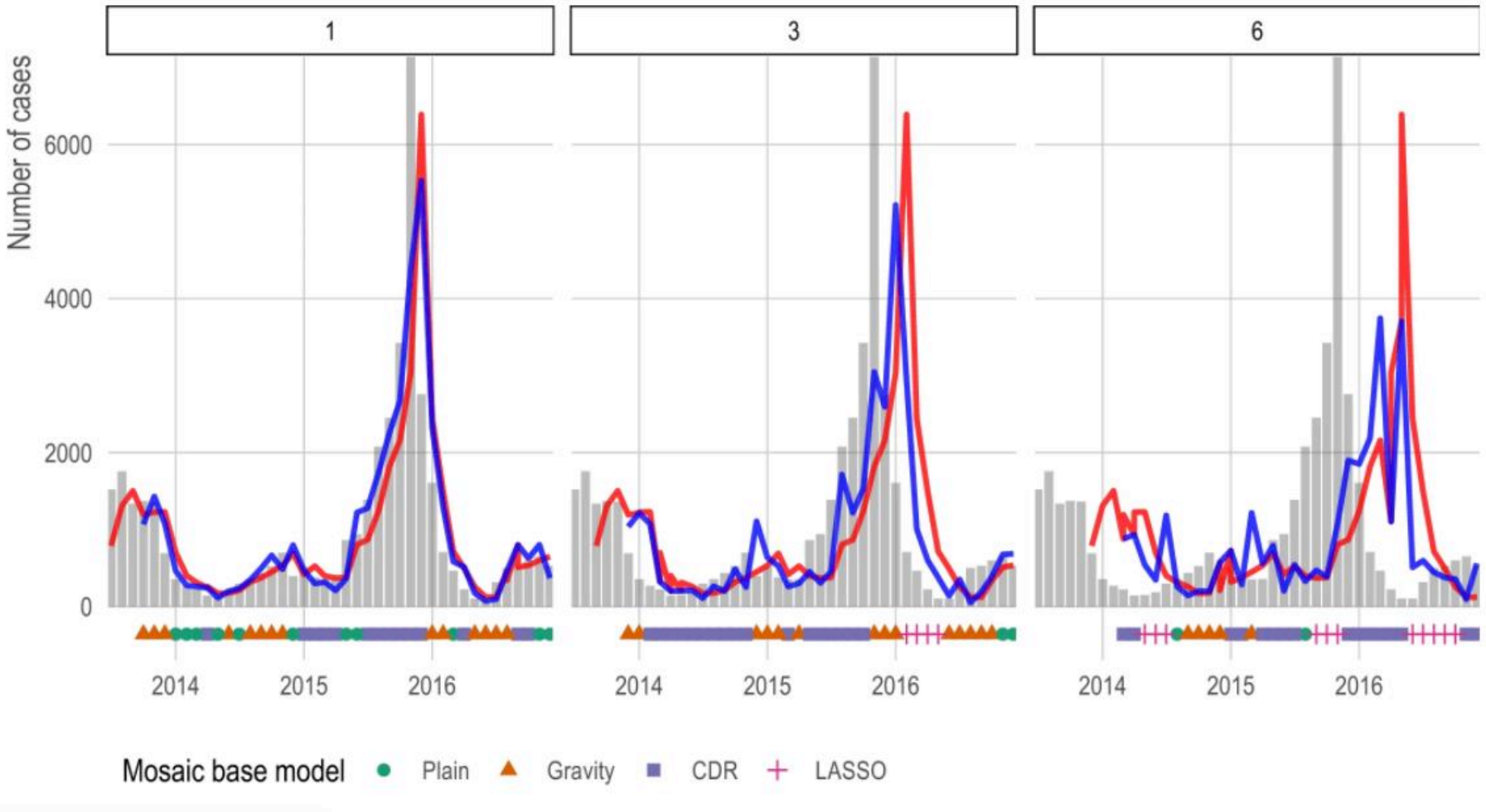
scientific reports

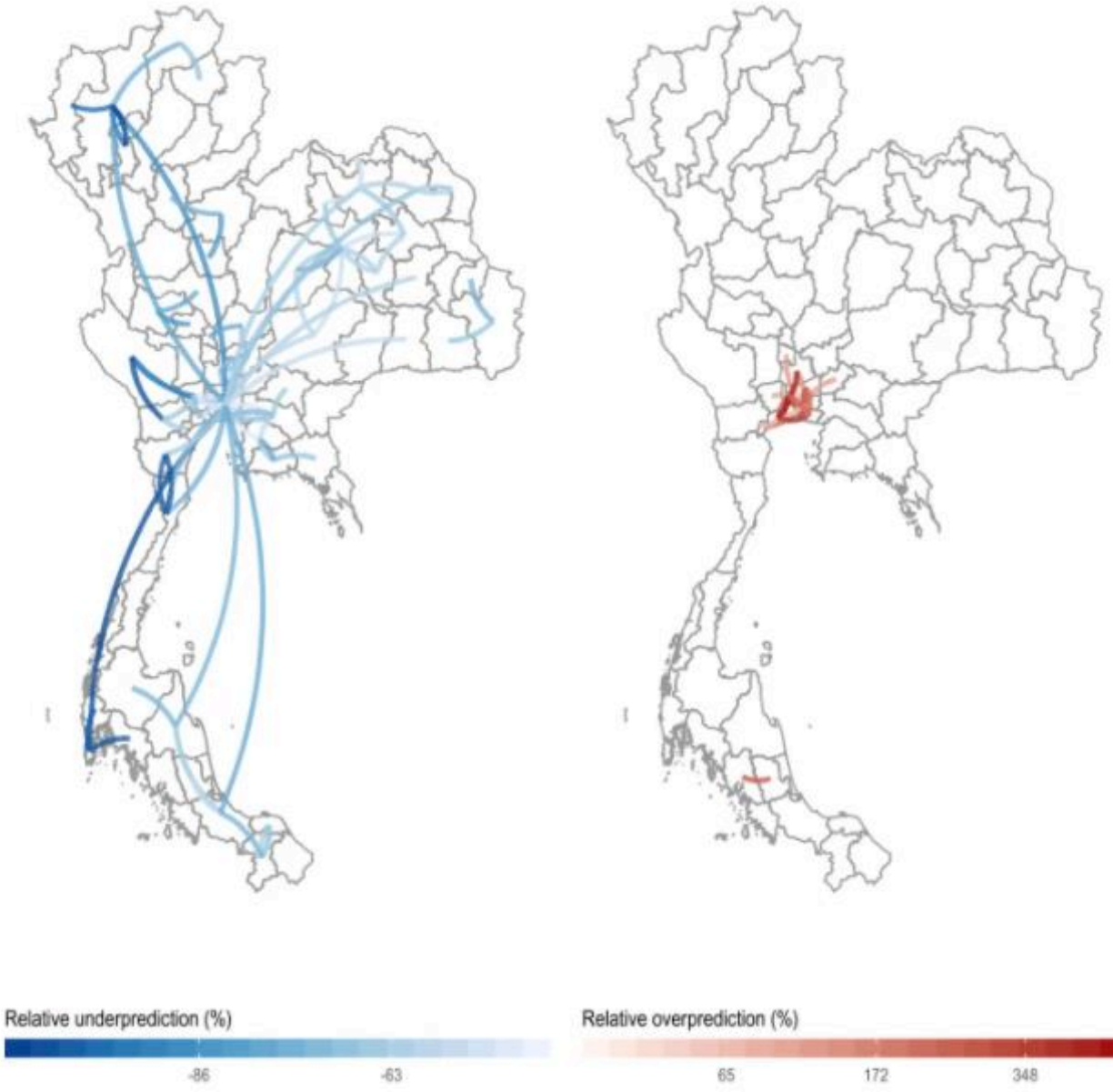
Article | Open Access | Published: 13 January 2021

Incorporating human mobility data improves forecasts of Dengue fever in Thailand

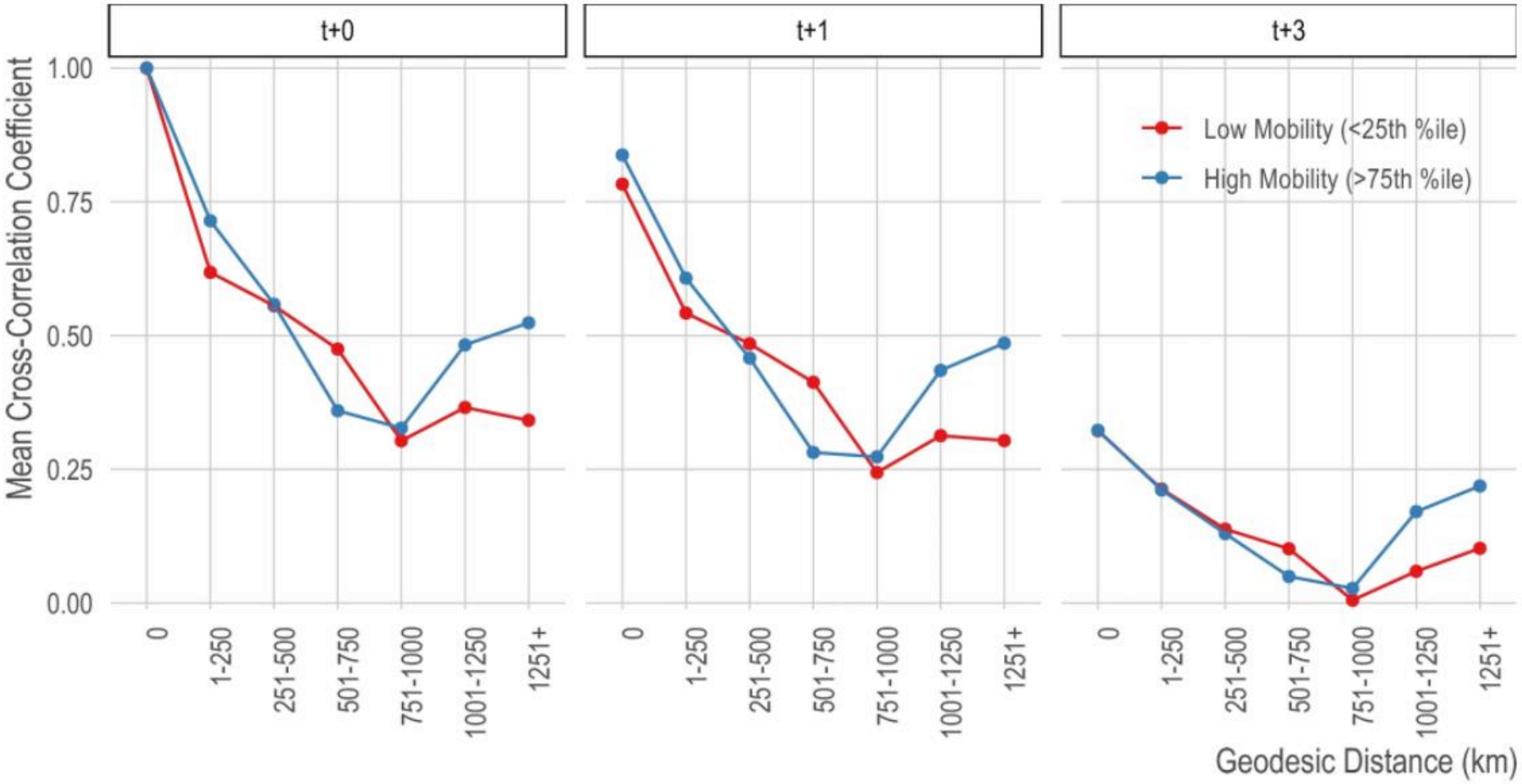
Mathew V. Kiang, Mauricio Santillana, Jarvis T. Chen, Jukka-Pekka Onnela, Nancy Krieger, Kenth Engø-Monsen, Nattwut Ekapirat, Darin Areechokchai, Preecha Prempree, Richard J. Maude & Caroline O. Buckee ✉

Scientific Reports **11**, Article number: 923 (2021) | Cite this article


1
3
6
Number of cases
6000
4000
2000
0
2014
2015
2016
Mosaic base model
Plain
Gravity
CDR
LASSO



Under- and over-prediction of outlier travel. Relative under-prediction (left) and over-prediction (right) comparing observed mobility data (from CDRs) to estimated mobility data from the best fit gravity model. We defined relative prediction error as 100%*(PredictedTrips – ObservedTrips)/ObservedTrips. We highlight only observations with Cook's distance greater than



Correlation of province-level dengue by distance, at different time lags. We show the mean cross-correlation coefficient (y-axis) for pairs of provinces at binned distances (x-axis; 0 indicates correlation of an area with itself) for synchronous dengue (left panel) and lagged by 1 month (middle panel) and 3 months (right panel). The lines are separated based on the connectivity of pairs of provinces where the red line shows the bottom quartile of provinces in terms of incoming and outgoing travel and the blue line shows the top quartile. Bangkok, an important travel hub, is in the approximate center of Thailand and between 700 and 800 km from all other provinces, therefore the last two distance categories do not include Bangkok.

Every hour

24/7

93

2,000

Public & private sources

alerts per day, precisely placed in

10,000+

locations

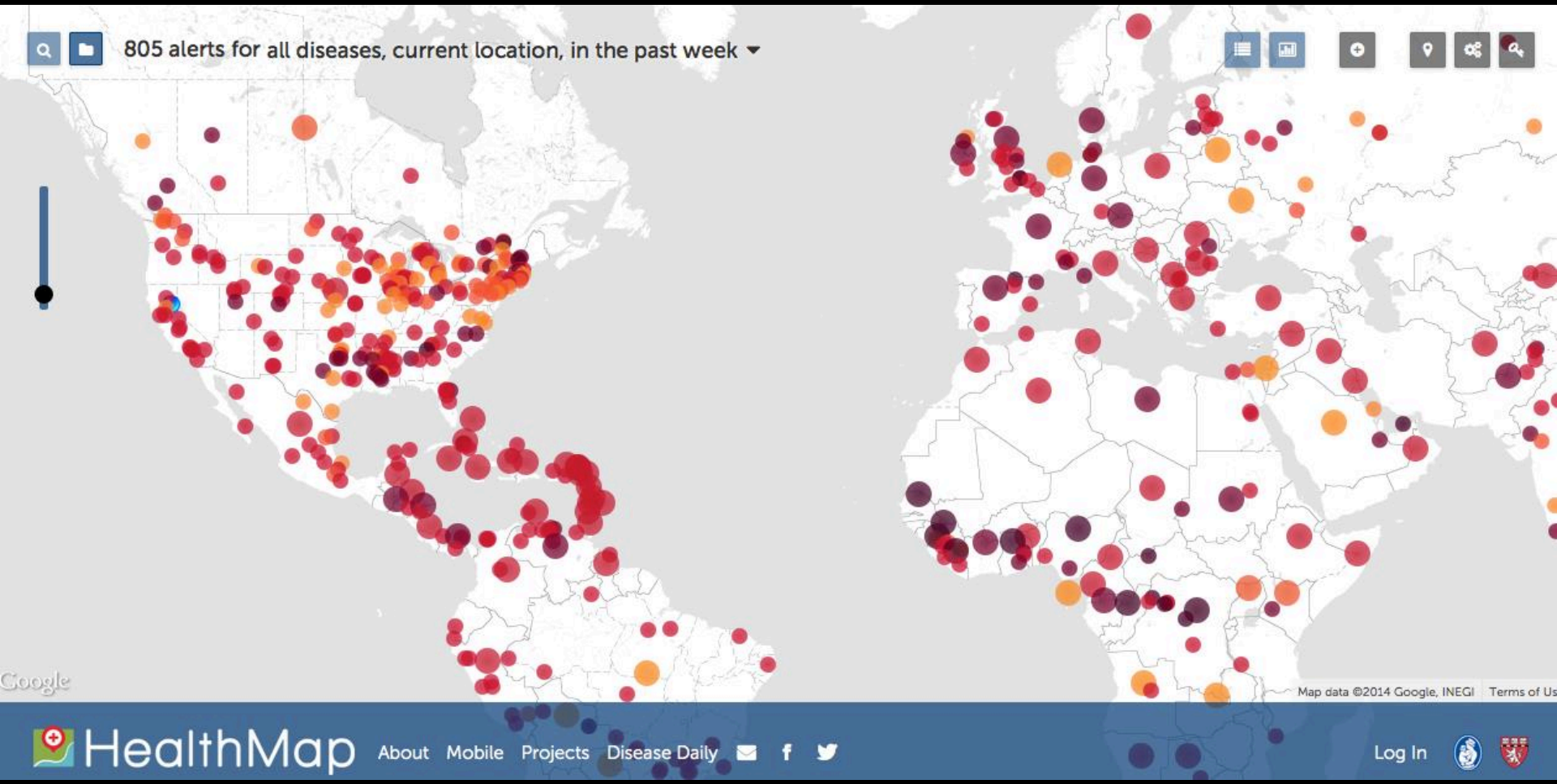
100,000

Web

15

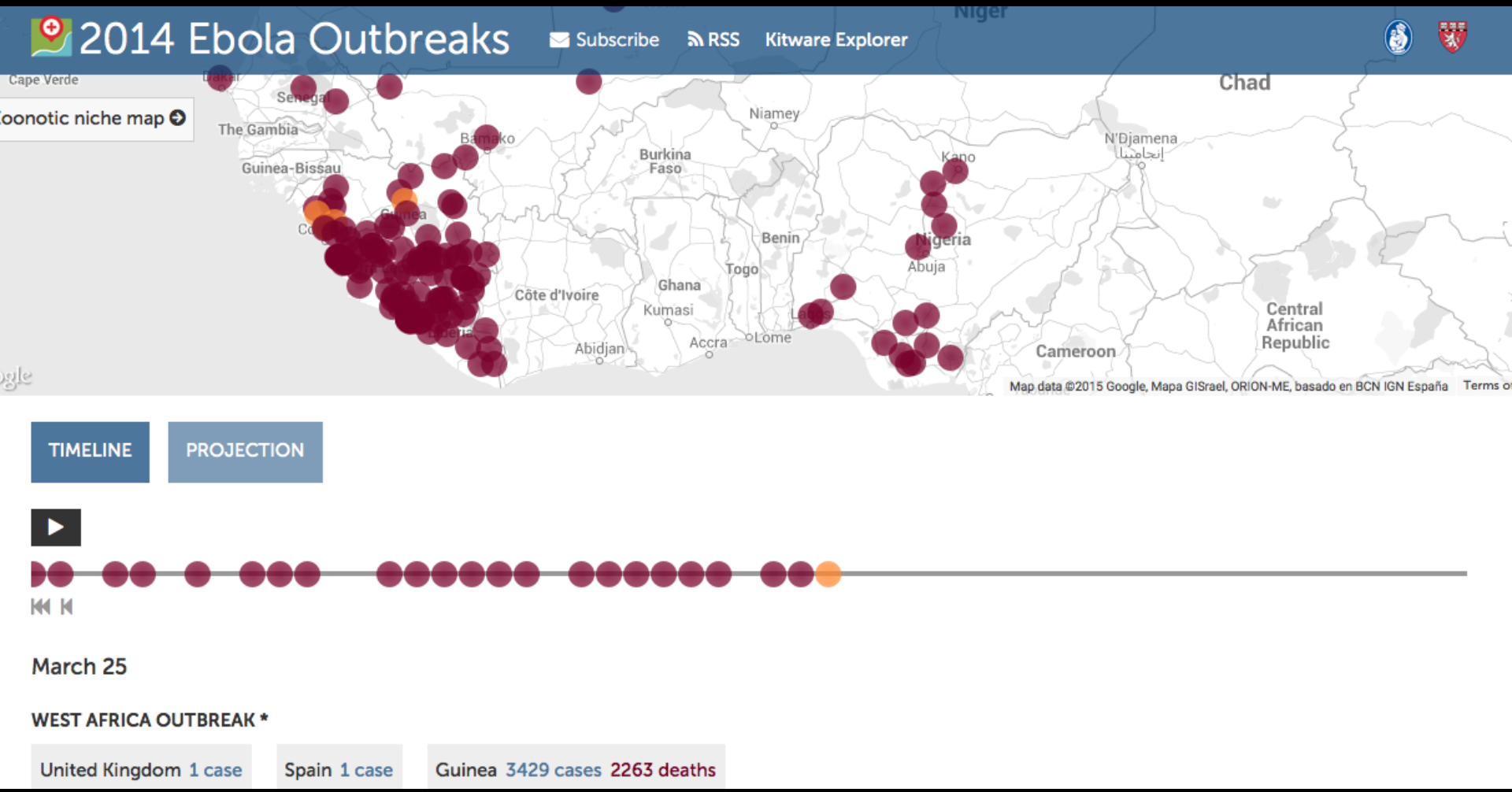
Languages

HealthMap brings together disparate data sources, including online news aggregators, eyewitness reports, expert-curated discussions and validated official reports, to achieve a unified and comprehensive view of the current global state of infectious diseases and their effect on human and animal health.



Through an automated process, updating 24/7/365, the system monitors, organizes, integrates, filters, visualizes and disseminates online information about emerging diseases in nine languages, facilitating early detection of global public health threats

Recent success story: Ebola outbreak identification and tracking



http://www.healthmap.org/ebola/#timeline

PLOS | CURRENTS OUTBREAKS

2014 Ebola Outbreak: Media Events Track Changes in Observed Reproductive Number

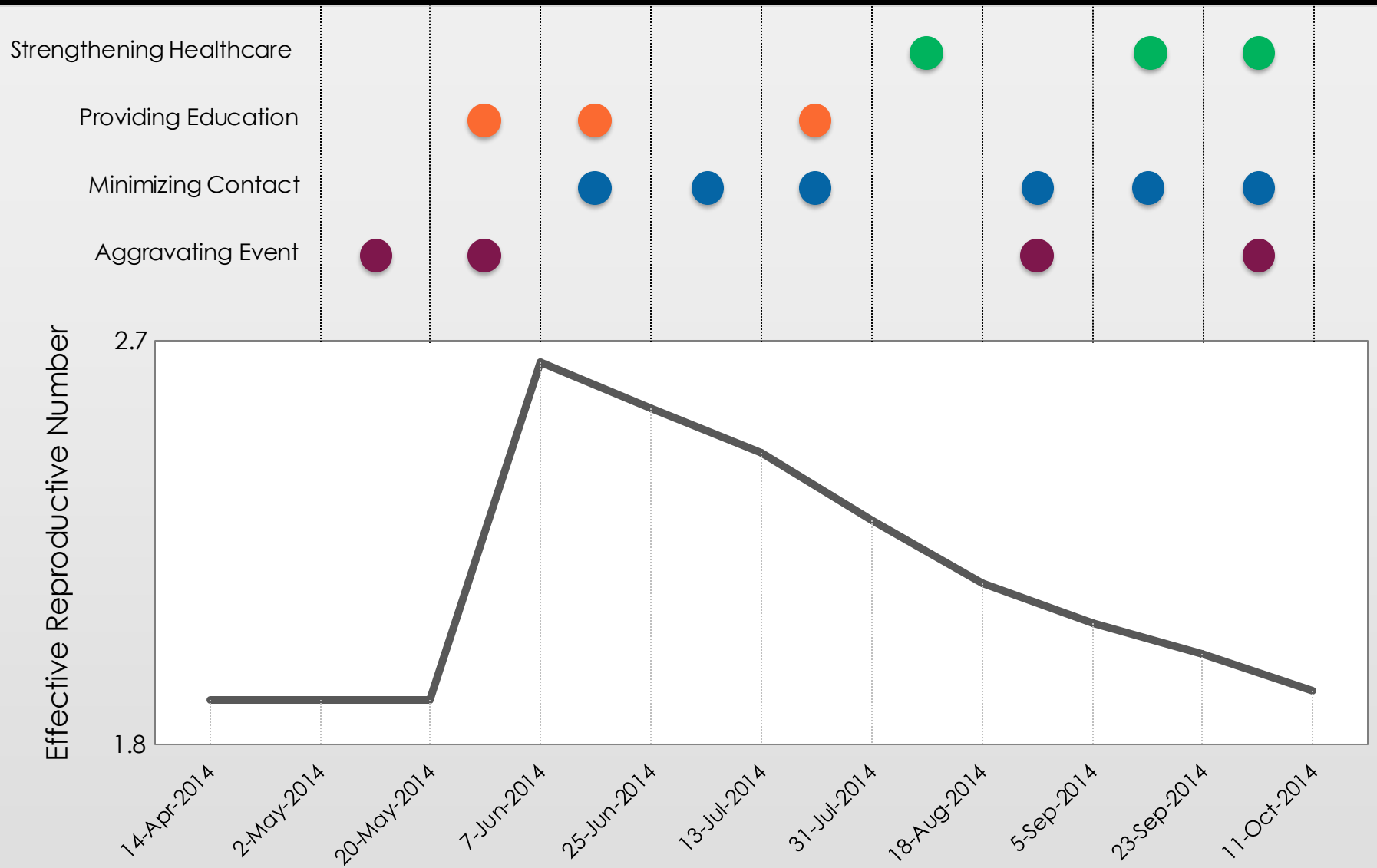
APRIL 28, 2015 · COMMENTARY

AUTHORS

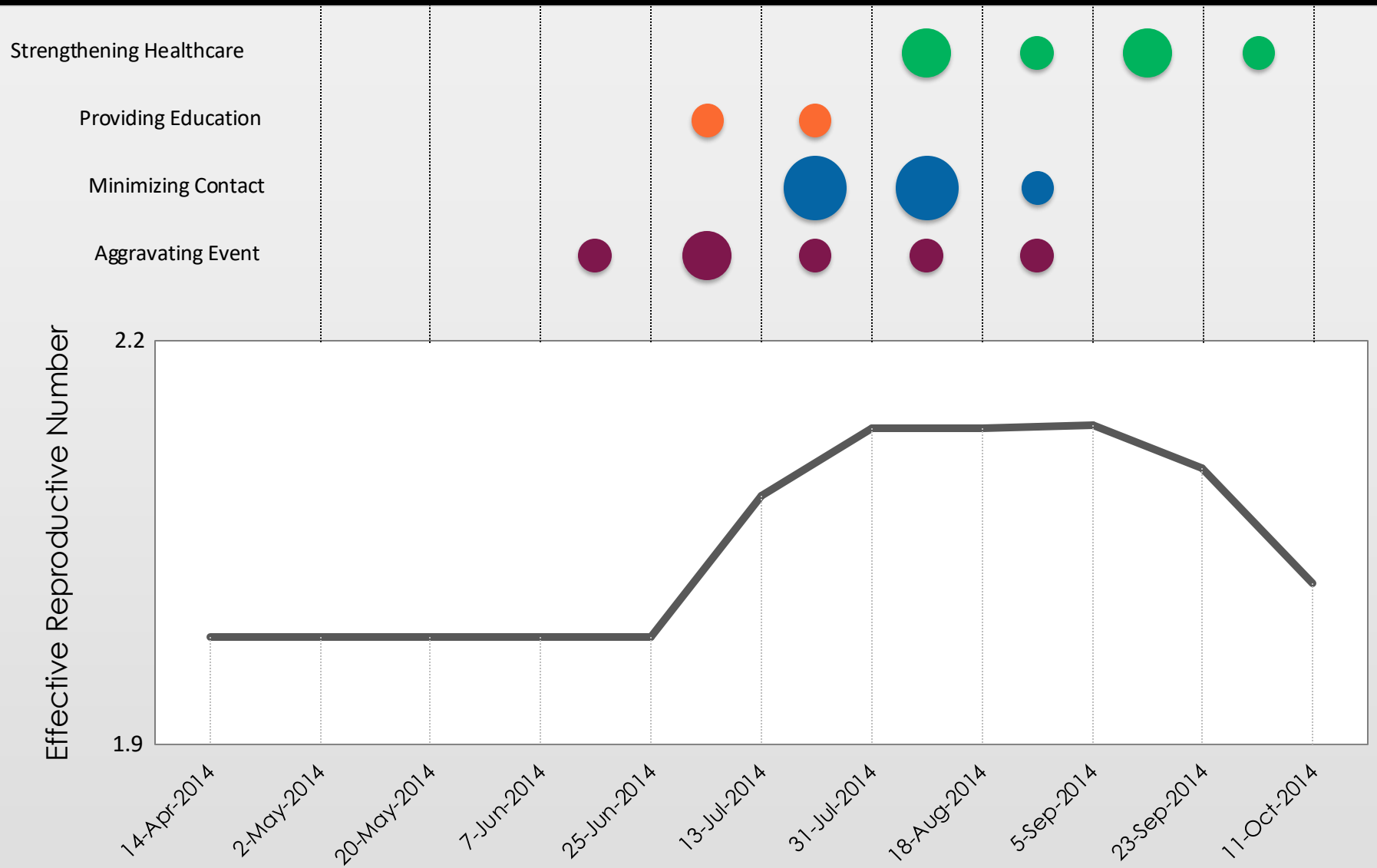
Maimuna S. Majumder Sheryl Kluberg Mauricio Santillana Sumiko Mekaru John S. Brownstein

ABSTRACT

In this commentary, we consider the relationship between early outbreak changes in the observed reproductive number of Ebola in West Africa and various media reported interventions and aggravating events. We find that media reports of interventions that provided education, minimized contact, or strengthened healthcare were typically followed by sustained transmission reductions in both Sierra Leone and Liberia. Meanwhile, media reports of aggravating events generally preceded temporary transmission increases in both countries. Given these preliminary findings, we conclude that media reported events could potentially be incorporated into future epidemic modeling efforts to improve mid-outbreak case projections.



Sierra Leone


Strengthening Healthcare
Providing Education
Minimizing Contact
Aggravating Event
Effective Reproductive Number
2.2
1.9
14-Apr-2014
2-May-2014
20-May-2014
7-Jun-2014
25-Jun-2014
13-Jul-2014
31-Jul-2014
18-Aug-2014
5-Sep-2014
23-Sep-2014
11-Oct-2014

Liberia

Can other Internet-based data sourced be used to monitor **emerging disease outbreaks** in **real time** in **Africa**?

Yes, news alerts related to the 2014 Ebola outbreak in Western Africa foreshadowed changes in the local reproductive number

PLOS | CURRENTS OUTBREAKS

Home Aims & Scope Review Board Authors ↓ Resources About

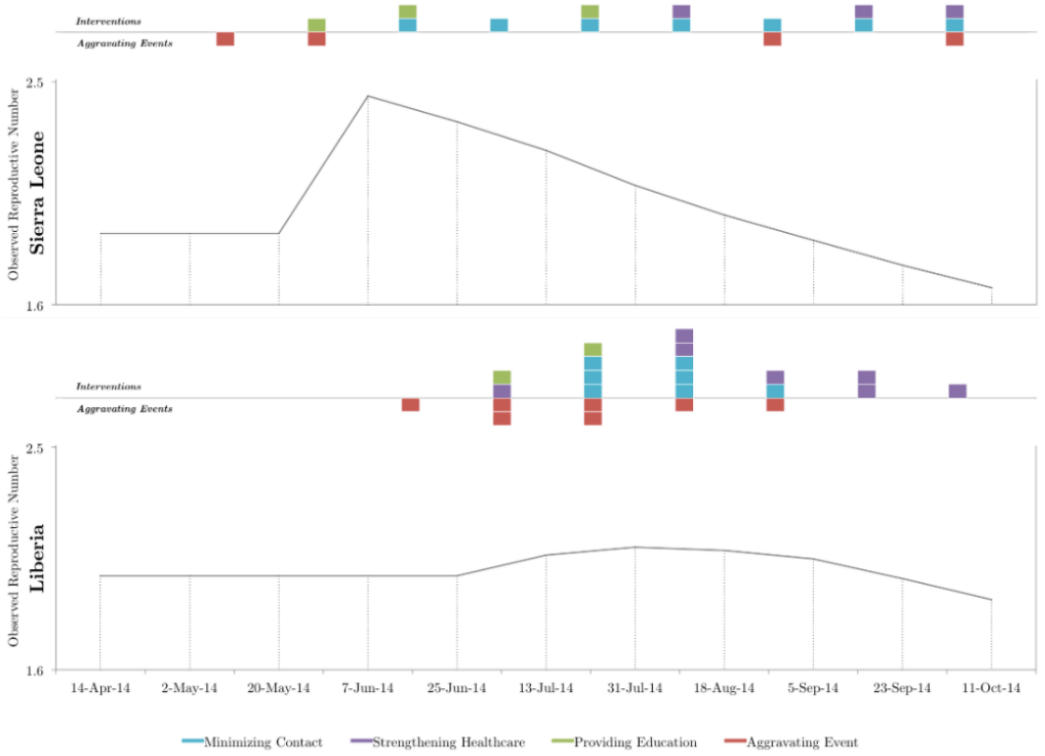
2014 Ebola Outbreak: Media Events Track Changes in Observed Reproductive Number

APRIL 28, 2015 · DISCUSSION

Print or Save PDF Citation XML Email

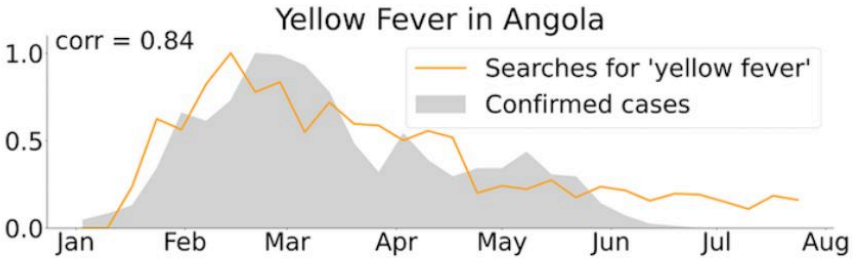
AUTHORS

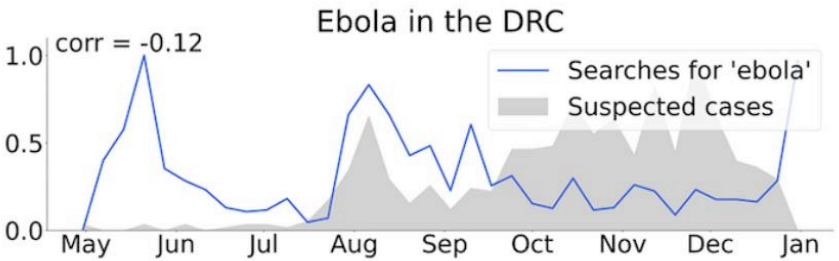
Maimuna S. Majumder Sheryl Kluberg Mauricio Santillana Sumiko Mekaru John S. Brownstein

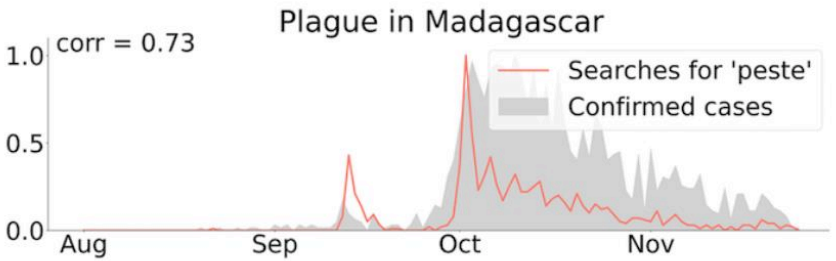


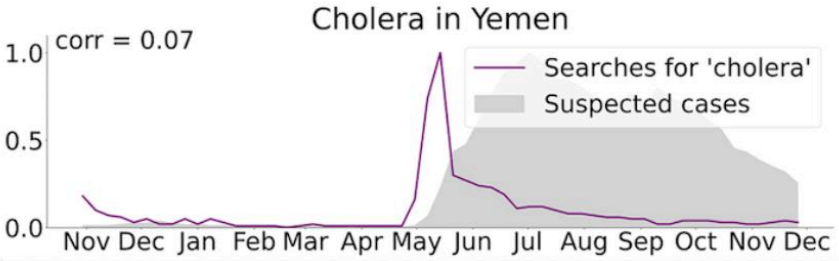
https://currents.plos.org/outbreaks/index.html%3Fp=50634.html

Do these methods work for **Emerging Disease Outbreaks** in in the developing world?
Yes, with certain limitations









https://journals.plos.org/ploscompbiol/article/authors?id=10.1371/journal.pcbi.1008117

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Real-time estimation of disease activity in emerging outbreaks using internet search information

Emily L. Aiken, Sarah F. McGough, Maimuna S. Majumder, Gal Wachtel, Andre T. Nguyen, Cecile Viboud, Mauricio Santillana

Climate and Health

JOURNAL OF THE ROYAL SOCIETY INTERFACE

Section

Research articles

A dynamic, ensemble learning approach to forecast dengue fever epidemic years in Brazil using weather and population susceptibility cycles

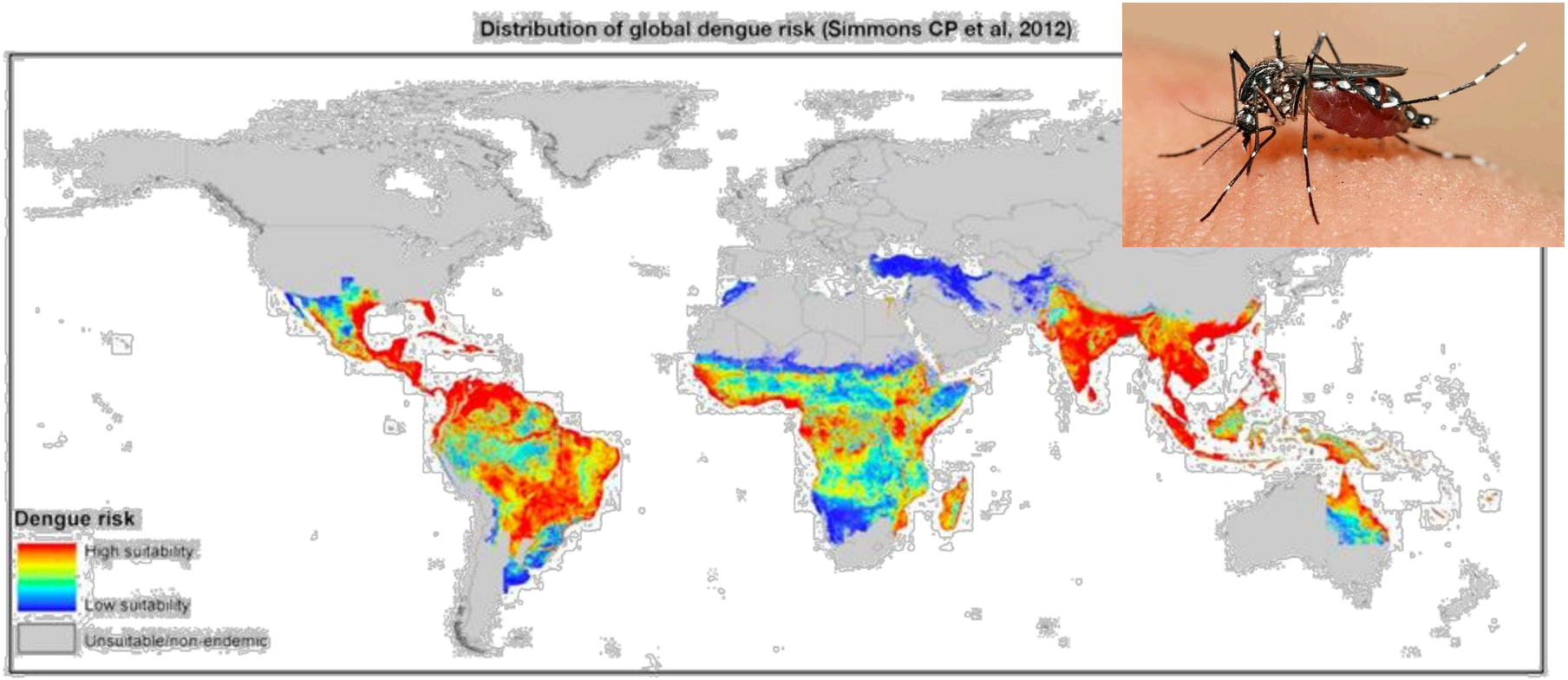
Sarah F. McGough, Leonardo Clemente, J. Nathan Kutz and Mauricio Santillana

Published: 16 June 2021 https://doi.org/10.1098/rsif.2020.1006

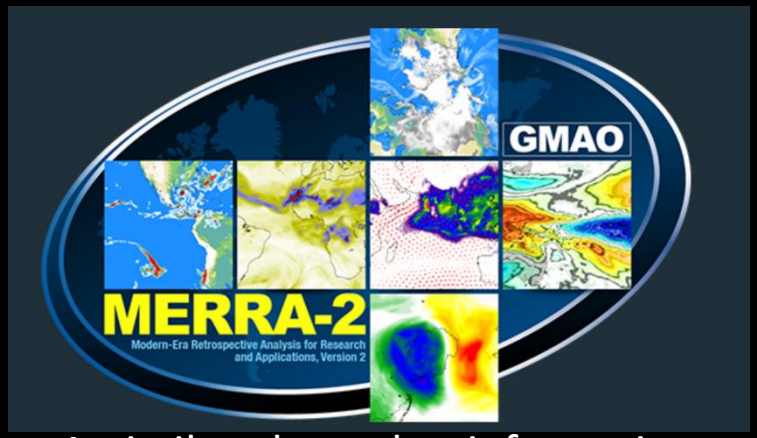
Abstract

Transmission of dengue fever depends on a complex interplay of human, climate and mosquito dynamics, which often change in time and space. It is well known that its disease dynamics are highly influenced by multiple factors including population susceptibility to infection as well as by microclimates: small-area climatic conditions which create environments favourable for the breeding and survival of mosquitoes. Here, we present a novel machine learning dengue forecasting approach, which, dynamically in time and space, identifies local patterns in weather and population susceptibility to make epidemic predictions at the city level in Brazil, months ahead of the occurrence of disease outbreaks. Weather-based predictions are improved when information on population susceptibility is incorporated, indicating that immunity is an important predictor neglected by most dengue forecast models. Given the generalizability of our methodology to any location or input data, it may prove valuable for public health decision-making aimed at mitigating the effects of seasonal dengue outbreaks in locations globally.

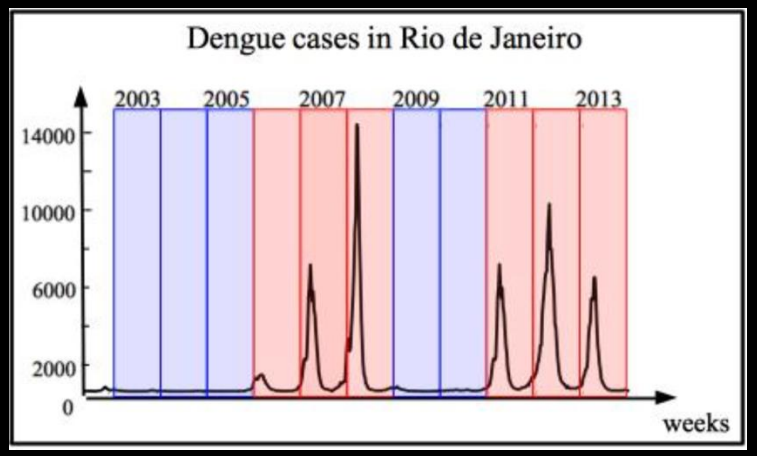
Predicting Dengue epidemic years in Brazil months before they happen



Can we leverage available weather information and susceptibility to predict an epidemic year– in a wide range of locations?



Assimilated weather information
(available for every location worldwide)



Data driven identification of 3-4 year susceptibility depletion cycles

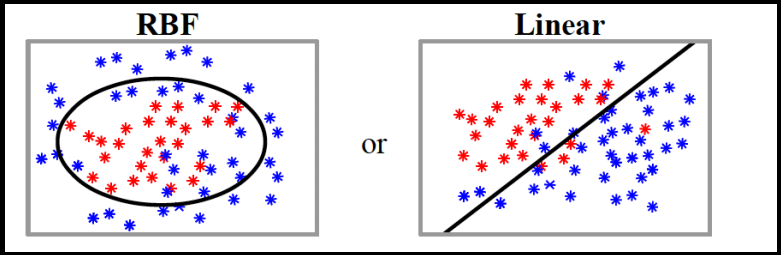
Our contribution:

- Used assimilated weather data
- 20 cities with 17 years of data
- Improved results with new weather data
- Out-of-sample predictions for 4-6 years
- Incorporated susceptibility data
- Adaptive and dynamically calibrated

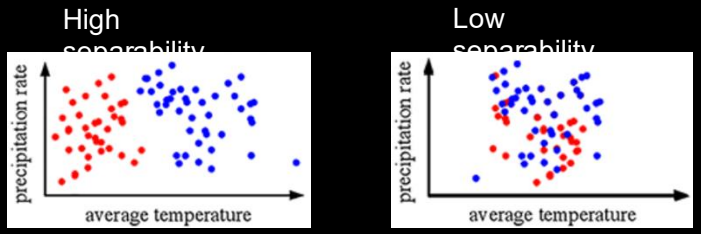
Team: Sarah McGough, Nathan Kutz, Mauricio Santillana

SVM methodology for classification

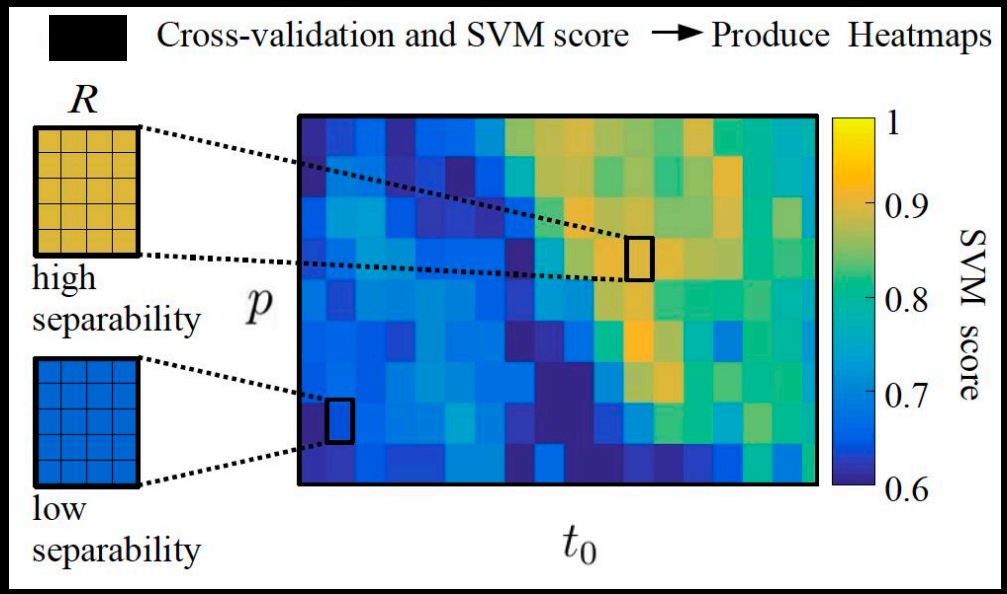
1. Choose kernel for SVM training



2. Calculate SVM score – % of correctly classified test points after an 80/20 split of the training data and re-sampling 100 times



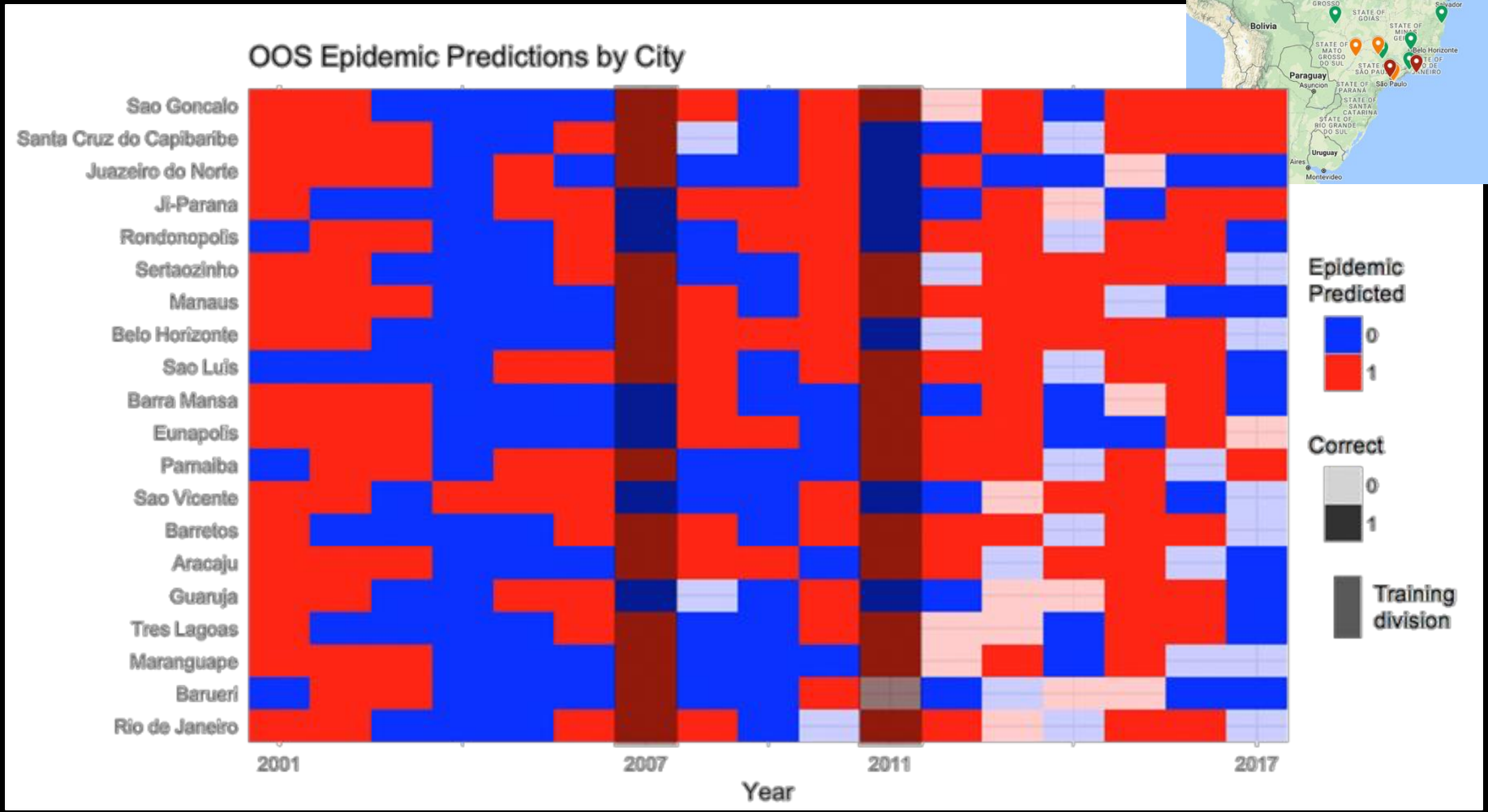
3. Repeat for different values of p and t0



High SVM score → clear separation in the trends between epidemic and non-epidemic years

Can we leverage available weather information and susceptibility to predict an epidemic year– in a wide range of locations?

(Results)



Letter | Published: 21 May 2018

Antibiotic resistance increases with local temperature

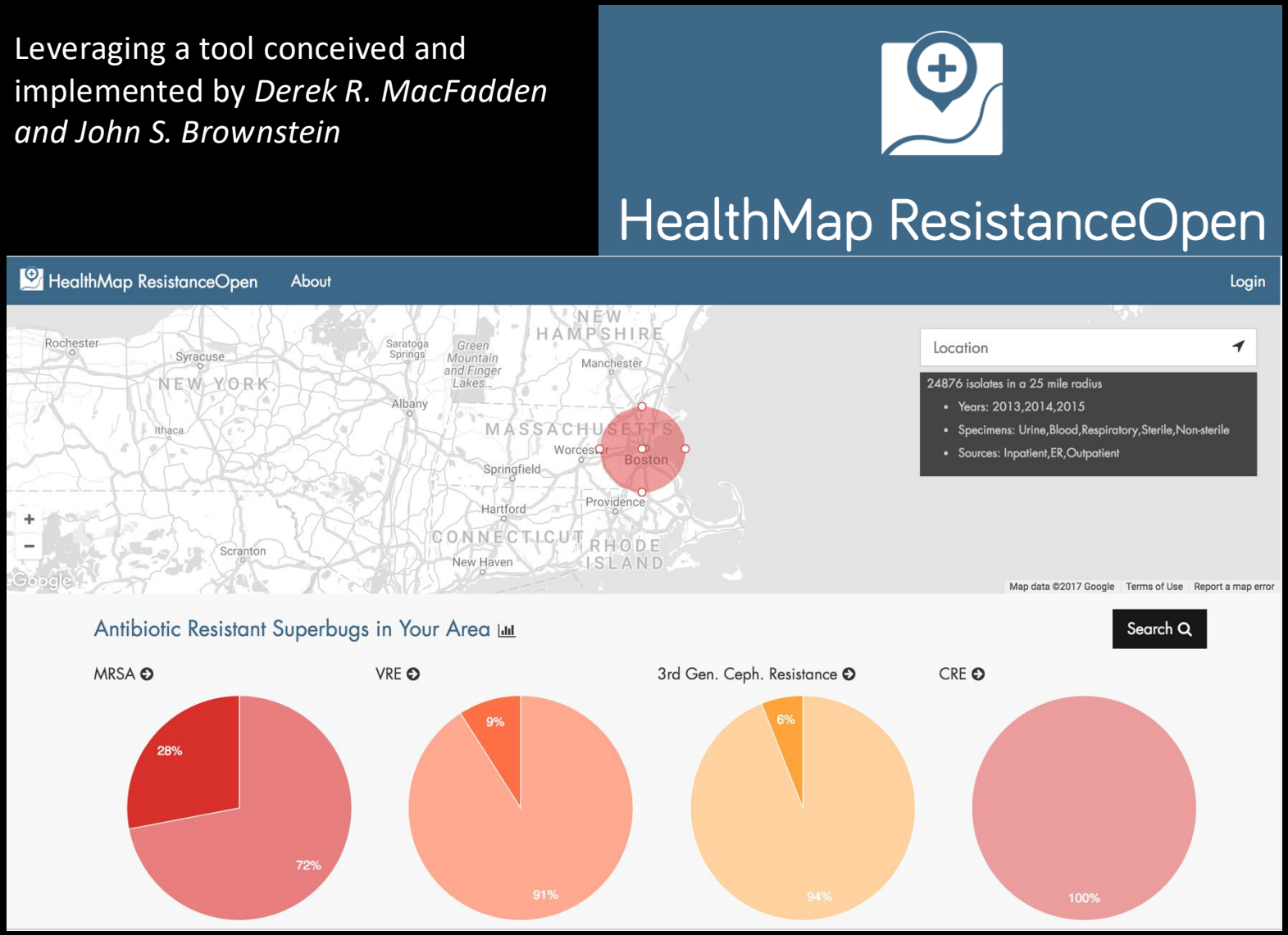
Derek R. MacFadden, Sarah F. McGough, David Fisman, Mauricio Santillana & John S. Brownstein

Nature Climate Change **8**, 510–514 (2018) | Download Citation

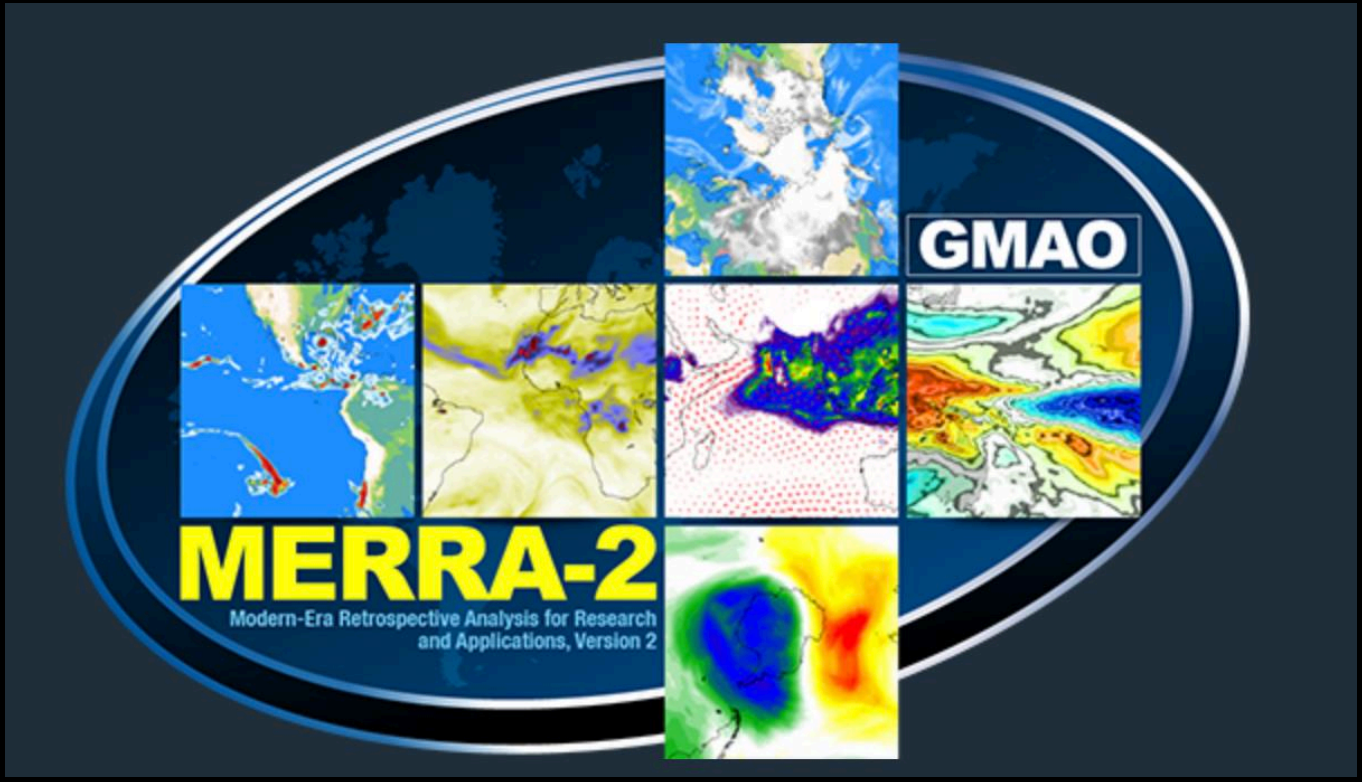
Abstract

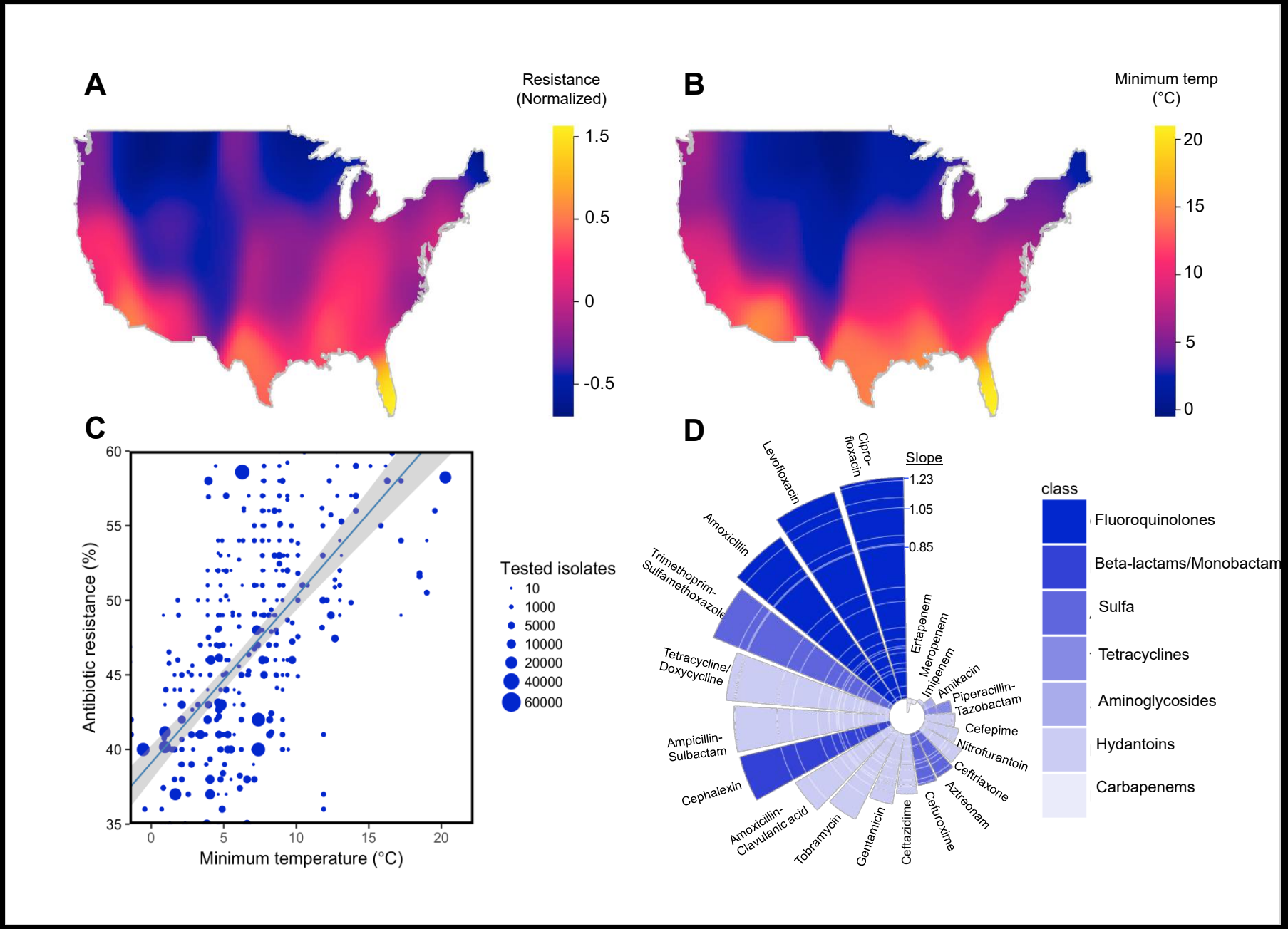
Bacteria that cause infections in humans can develop or acquire resistance to antibiotics commonly used against them[1,2]. Antimicrobial resistance (in bacteria and other microbes) causes significant morbidity

Leveraging a tool conceived and implemented by *Derek R. MacFadden and John S. Brownstein*



And leveraging assimilated weather information (available for every location worldwide)




A
Resistance
(Normalized)
1.5
0.5
0
-0.5
B
Minimum temp
(°C)
20
15
10
5
0
C
Antibiotic resistance (%)
60
55
50
45
40
35
0
5
10
15
20
Minimum temperature (°C)
Tested isolates
10
1000
5000
10000
20000
40000
60000
D
Slope
1.23
1.05
0.85
Levofloxacin
Cipro-
floxacin
Amoxicillin
Trimethoprim-
Sulfamethoxazole
Tetracycline/
Doxycycline
Ampicillin-
Sulbactam
Cephalexin
Amoxicillin-
Clavulanic acid
Tobramycin
Gentamicin
Ceftazidime
Cefuroxime
Aztreonam
Ceftriaxone
Nitrofurantoin
Cefepime
Piperacillin-
Tazobactam
Amikacin
Imipenem
Meropenem
Ertapenem
class
Fluoroquinolones
Beta-lactams/Monobactam
Sulfa
Tetracyclines
Aminoglycosides
Hydantoins
Carbapenems

Eurosurveillance Europe's journal on infectious disease surveillance, epidemiolo

Research

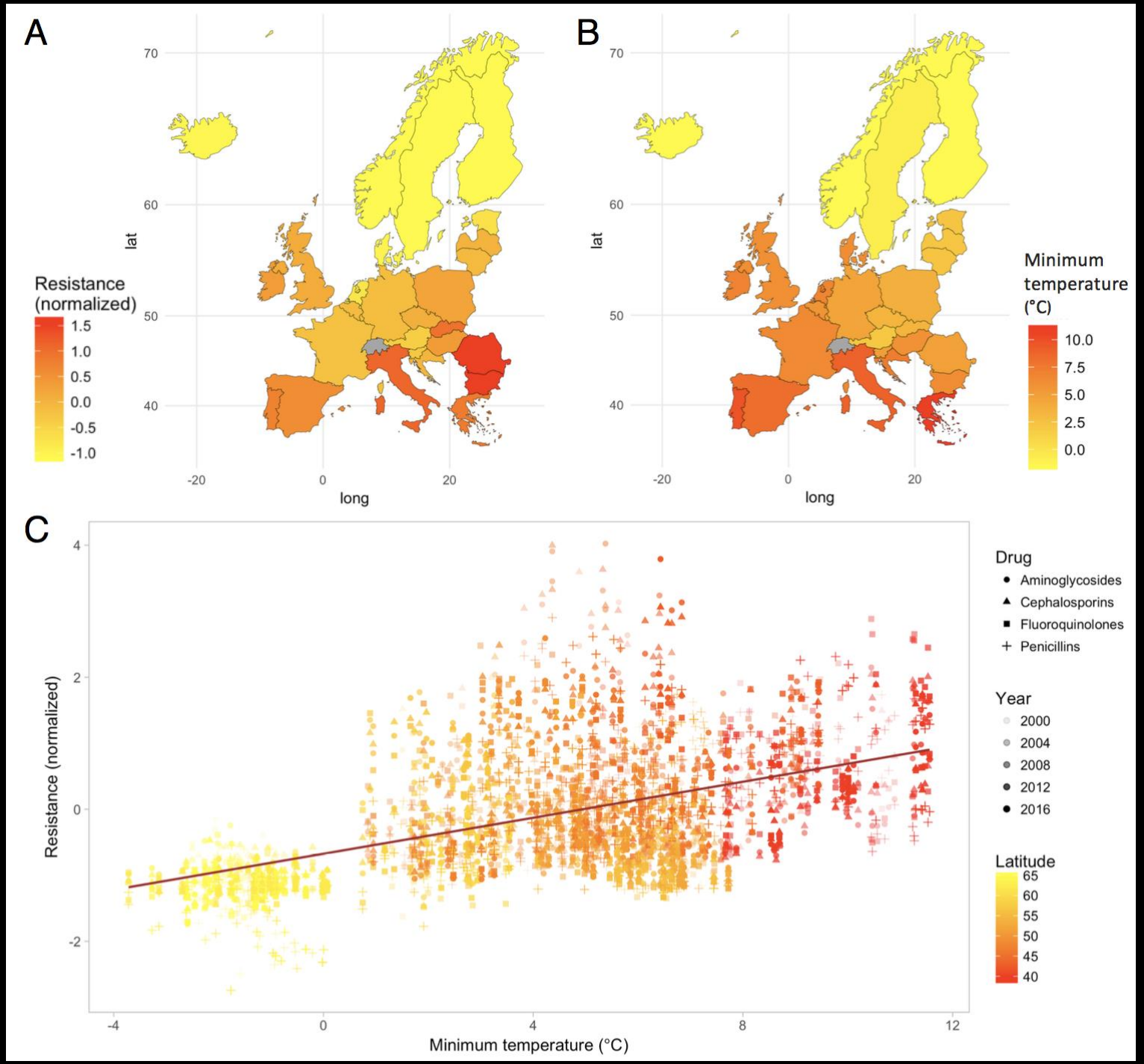
Open Access

Rates of increase of antibiotic resistance and ambient temperature in Europe: a cross-national analysis of 28 countries between 2000 and 2016

Sarah F McGough[1,2], Derek R MacFadden[1,3], Mohammad W Hattab[4], Kåre Mølbak[5,6], Mauricio Santillana[1,2,7]

View Affiliations

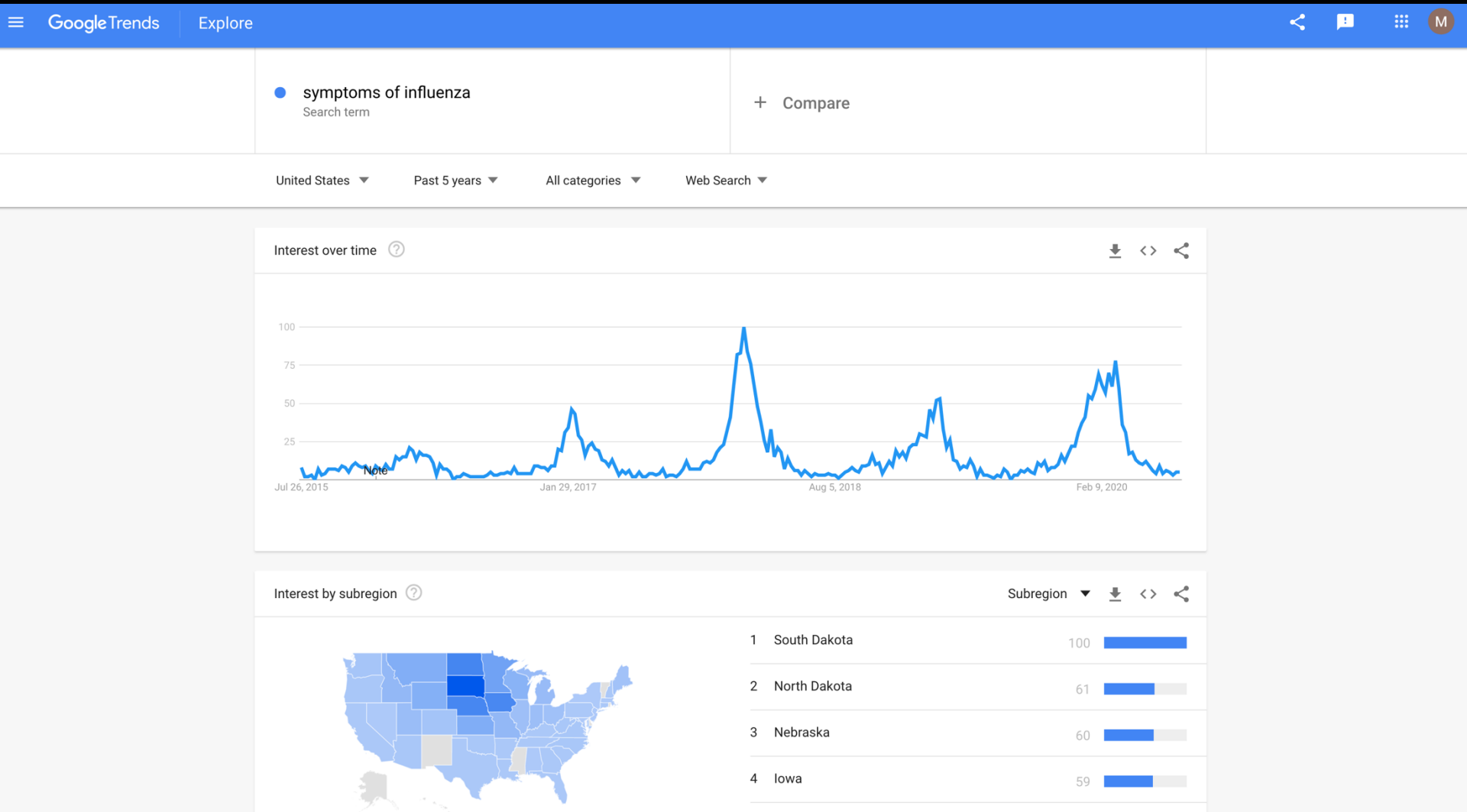
View Citation


A
B
C
Resistance (normalized)
1.5
1.0
0.5
0.0
-0.5
-1.0
lat
long
70
60
50
40
-20
0
20
Minimum temperature (°C)
10.0
7.5
5.0
2.5
0.0
Resistance (normalized)
Minimum temperature (°C)
4
2
0
-2
-4
0
4
8
12
Drug
Aminoglycosides
Cephalosporins
Fluoroquinolones
Penicillins
Year
2000
2004
2008
2012
2016
Latitude
65
60
55
50
45
40

What are these novel data sources? How do we access them?

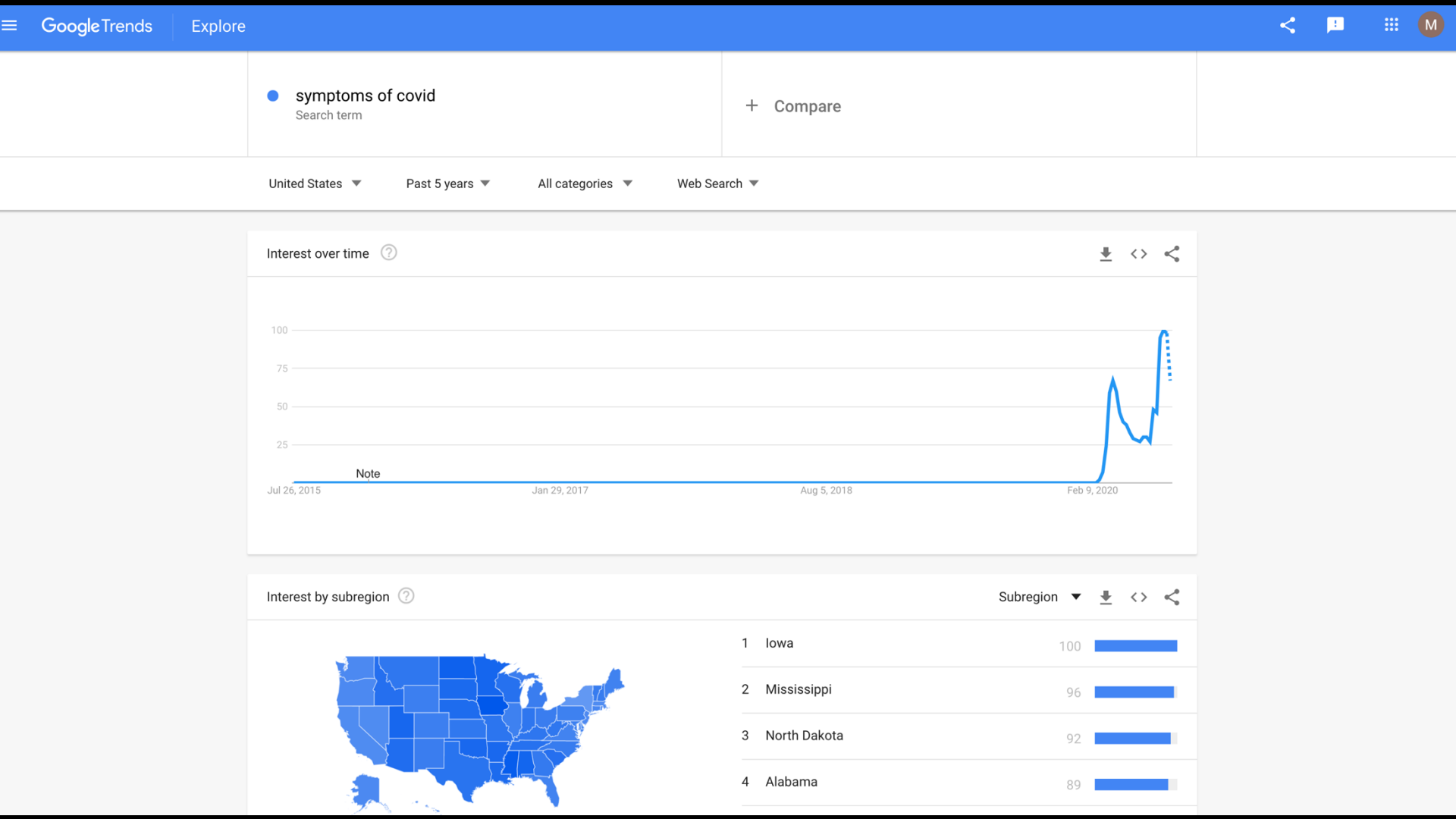
Accessing Google Trends for Specific Terms

“Influenza”

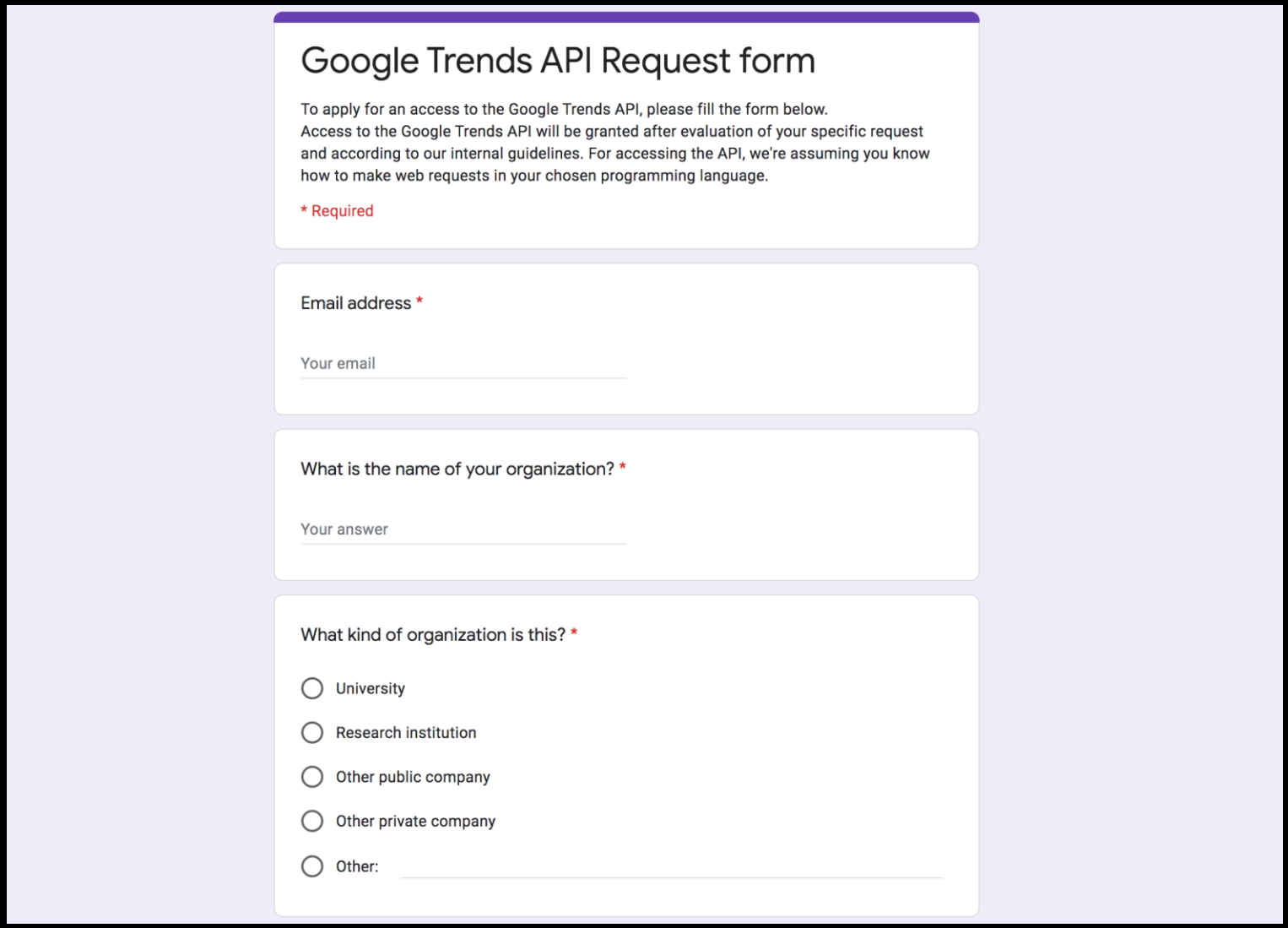


Accessing Google Trends for Specific Terms

“Covid”



How can you gain access to the Google Trends API?



Google Trends API Request form

To apply for an access to the Google Trends API, please fill the form below.
Access to the Google Trends API will be granted after evaluation of your specific request and according to our internal guidelines. For accessing the API, we're assuming you know how to make web requests in your chosen programming language.

* Required

Email address *

Your email

What is the name of your organization? *

Your answer

What kind of organization is this? *

- University
- Research institution
- Other public company
- Other private company
- Other:

https://docs.google.com/forms/d/e/1FAIpQLSenHdGiGl1YF-7rVDDmmuIN8R-ra9MnGLLs7gIIaAX9VHPdPg/viewform

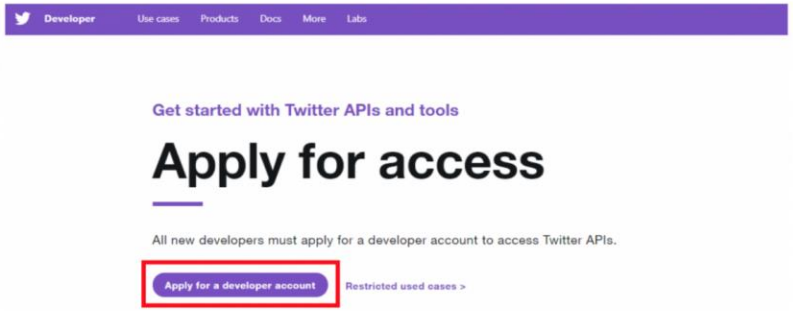
Explore Tweets Related to Flu in Boston

How to collect tweets from the Twitter Streaming API using Python

November 10, 2019 Laura South How to

Twitter provides a comprehensive streaming API that developers can use to download data about tweets in real-time, if they can figure out how to use it effectively. In this tutorial, we're going to retrace the steps I took to set up a server to collect tweets about hate speech as they occur to create a dataset we can use to learn more about patterns in hate speech.

The first step is to apply for a Twitter Developer account. You will have to link to your Twitter account and write a brief summary of what you plan to do with the data you extract using the Twitter API. You should receive approval quickly, but unfortunately you can't proceed with setup until your application is approved.

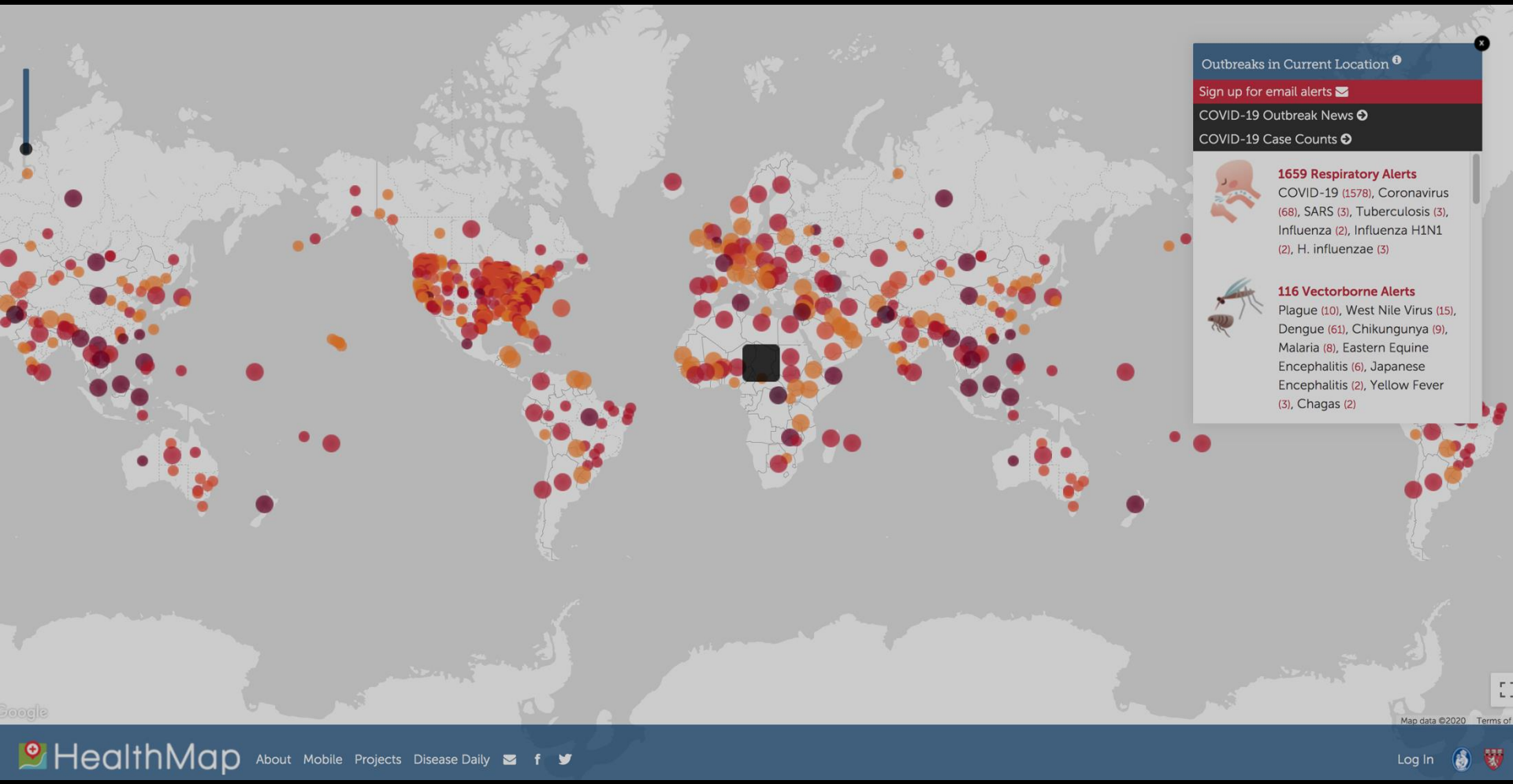


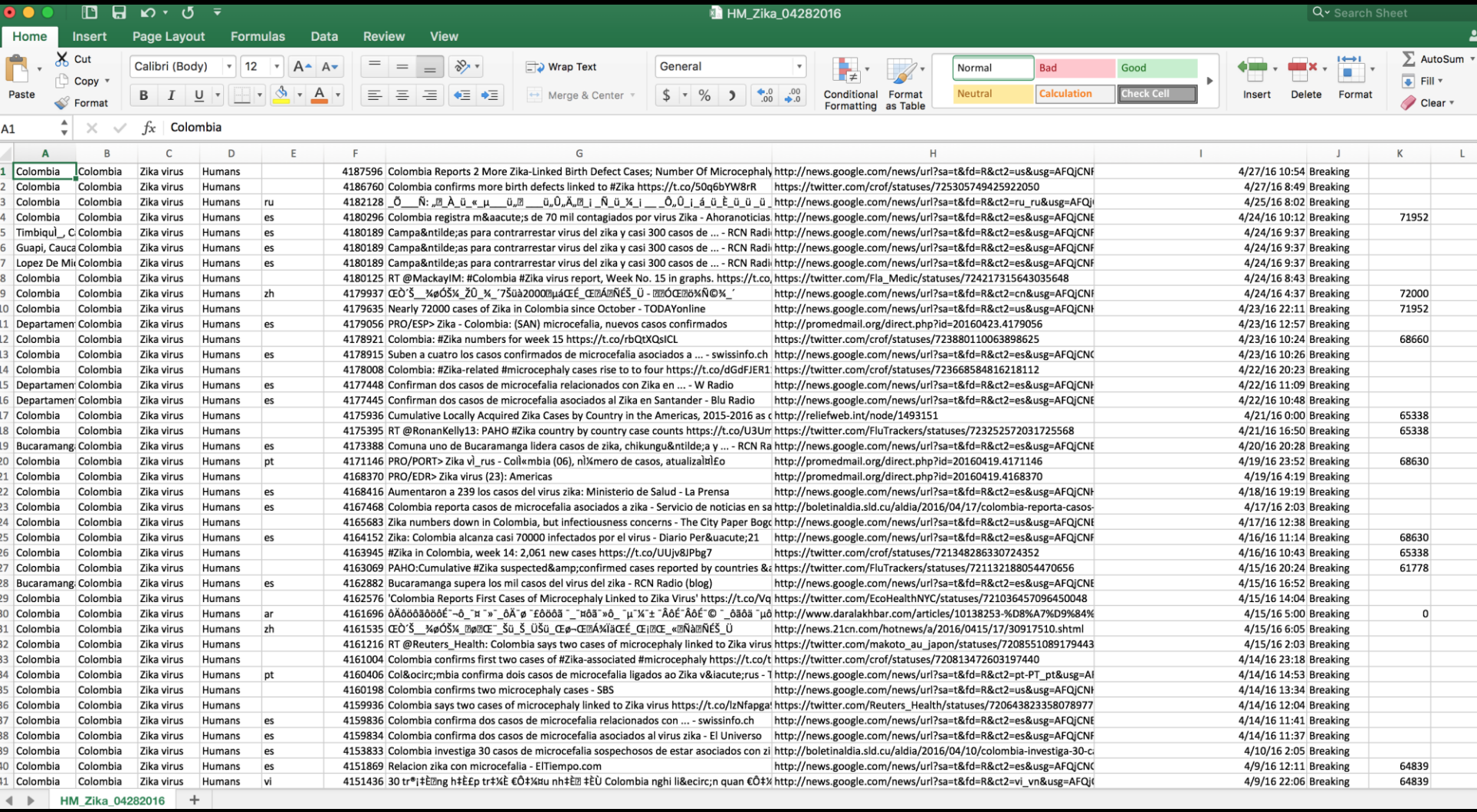
After you receive approval, you need to register your app. Within the Twitter Developer website, go to Apps and click "Create an app". The page will ask you for more information about what your app will do and where it will be hosted.

Now that we have all of the administrative stuff taken care of, we can get to the code. For this example, I'm using the Tweepy library to handle some of the streaming logistics.

Step 1: In Python, import Tweepy and set up your authentication and stream listener with API keys. Here we're using the authentication information Twitter provided when we registered our application.

Explore News Alerts for the Zika Outbreak in Colombia





HM_Zika_04282016

Search Sheet

Home Insert Page Layout Formulas Data Review View

Cut Copy Format Paste

Calibri (Body) 12

Wrap Text Merge & Center

General

Normal Bad Good Neutral Calculation Check Cell

Conditional Formatting Format as Table

Insert Delete Format

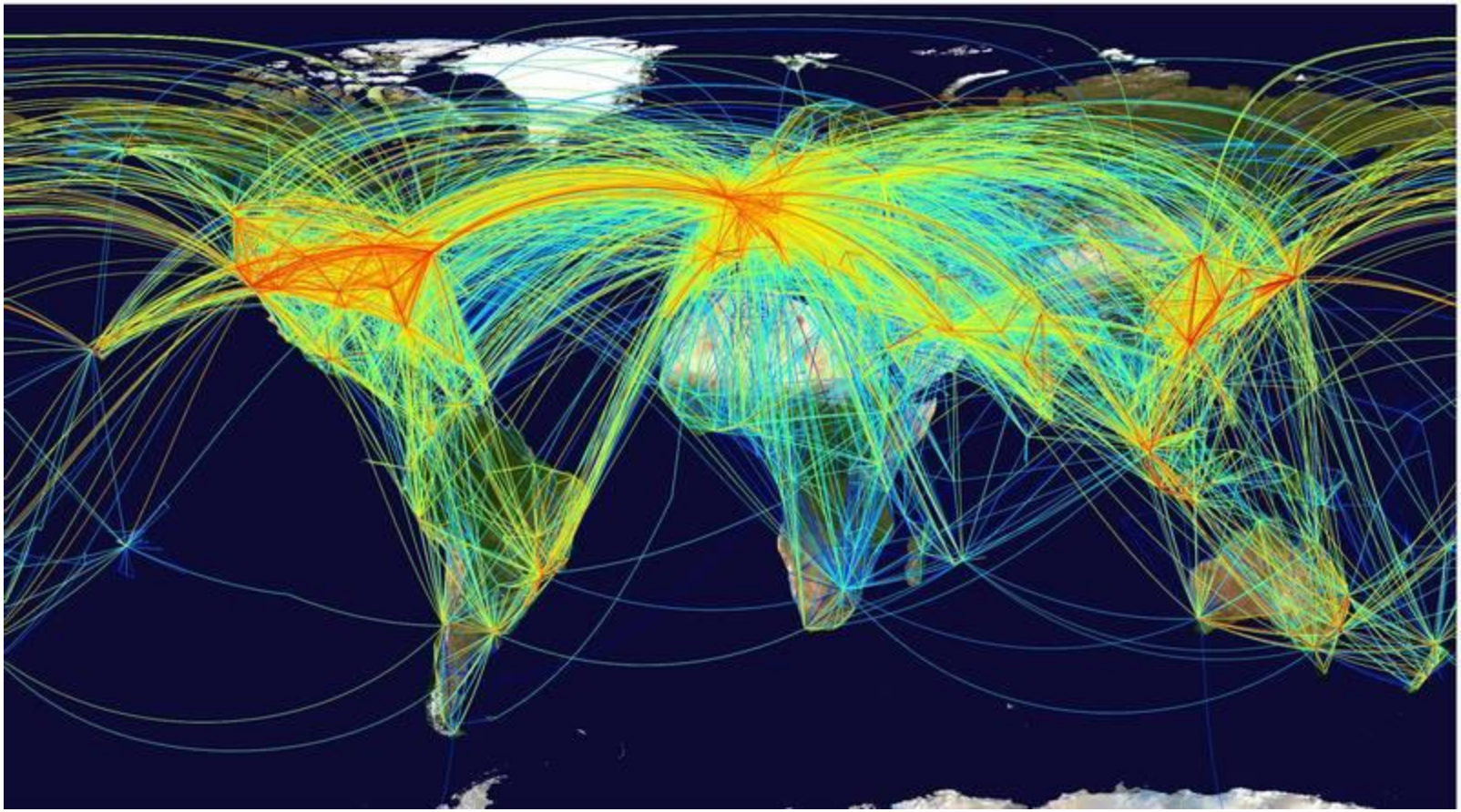
AutoSum Fill Clear

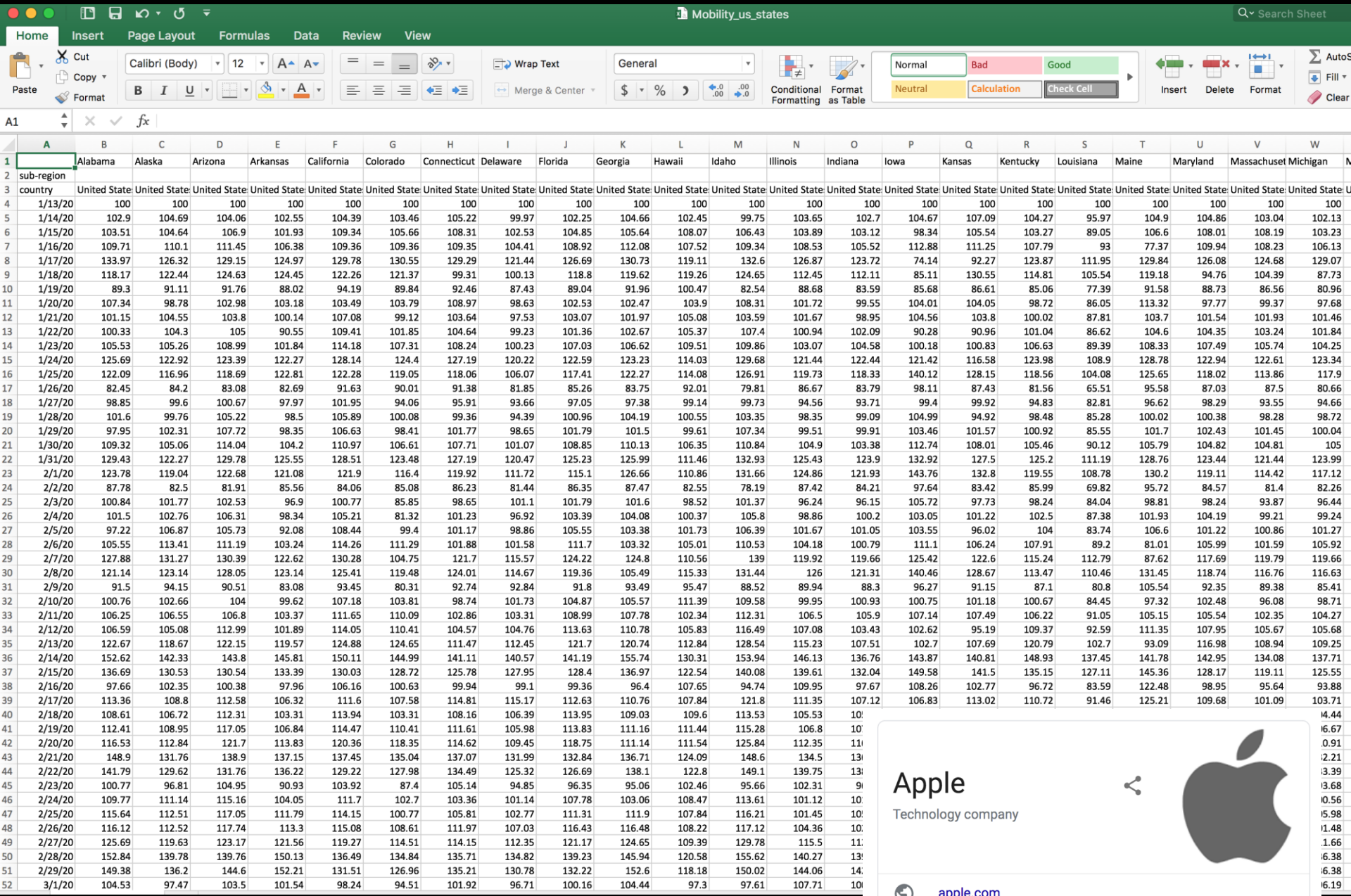
A1 Colombia

	A	B	C	D	E	F	G	H	I	J	K	L
1	Colombia	Colombia	Zika virus	Humans		4187596	Colombia Reports 2 More Zika-Linked Birth Defect Cases; Number Of Microcephaly	http://news.google.com/news/url?sa=t&fd=R&ct2=us&usg=AFQjCNF	4/27/16 10:54	Breaking		
2	Colombia	Colombia	Zika virus	Humans		4186760	Colombia confirms more birth defects linked to #Zika https://t.co/50q6bYW8rR	https://twitter.com/crof/statuses/725305749425922050	4/27/16 8:49	Breaking		
3	Colombia	Colombia	Zika virus	Humans	ru	4182128	_Õ___Ñ: „ß_À_ü_«_µ___ü„ß___ü„Õ„Ã„ß_i _Ñ_ü_¼_i __ _Õ„Û_i_á_ü_È_ü_ü _ü_	http://news.google.com/news/url?sa=t&fd=R&ct2=ru_ru&usg=AFQj	4/25/16 8:02	Breaking		
4	Colombia	Colombia	Zika virus	Humans	es	4180296	Colombia registra más de 70 mil contagiados por virus Zika - Ahoranoticias.	http://news.google.com/news/url?sa=t&fd=R&ct2=es&usg=AFQjCNE	4/24/16 10:12	Breaking	71952	
5	Timbiquì_, C	Colombia	Zika virus	Humans	es	4180189	Campañas para contrarrestar virus del zika y casi 300 casos de ... - RCN Radi	http://news.google.com/news/url?sa=t&fd=R&ct2=es&usg=AFQjCNF	4/24/16 9:37	Breaking		
6	Guapi, Cauca	Colombia	Zika virus	Humans	es	4180189	Campañas para contrarrestar virus del zika y casi 300 casos de ... - RCN Radi	http://news.google.com/news/url?sa=t&fd=R&ct2=es&usg=AFQjCNF	4/24/16 9:37	Breaking		
7	Lopez De Mi	Colombia	Zika virus	Humans	es	4180189	Campañas para contrarrestar virus del zika y casi 300 casos de ... - RCN Radi	http://news.google.com/news/url?sa=t&fd=R&ct2=es&usg=AFQjCNF	4/24/16 9:37	Breaking		
8	Colombia	Colombia	Zika virus	Humans		4180125	RT @MackayIM: #Colombia #Zika virus report, Week No. 15 in graphs. https://t.co,	https://twitter.com/Fla_Medic/statuses/724217315643035648	4/24/16 8:43	Breaking		
9	Colombia	Colombia	Zika virus	Humans	zh	4179937	ŒÒ´Š__¾øÓŠ¼_ŽÛ_¾_´7Šüà2000µáŒÉ_ŒÄÑÉŠ_Ü - ÓŒ©¾Ñ©¾_´	http://news.google.com/news/url?sa=t&fd=R&ct2=cn&usg=AFQjCNF	4/24/16 4:37	Breaking	72000	
10	Colombia	Colombia	Zika virus	Humans		4179635	Nearly 72000 cases of Zika in Colombia since October - TODAYonline	http://news.google.com/news/url?sa=t&fd=R&ct2=us&usg=AFQjCNH	4/23/16 22:11	Breaking	71952	
11	Departamen	Colombia	Zika virus	Humans	es	4179056	PRO/ESP> Zika - Colombia: (SAN) microcefalia, nuevos casos confirmados	http://promedmail.org/direct.php?id=20160423.4179056	4/23/16 12:57	Breaking		
12	Colombia	Colombia	Zika virus	Humans		4178921	Colombia: #Zika numbers for week 15 https://t.co/rbQtXQsICL	https://twitter.com/crof/statuses/723880110063898625	4/23/16 10:24	Breaking	68660	
13	Colombia	Colombia	Zika virus	Humans	es	4178915	Suben a cuatro los casos confirmados de microcefalia asociados a ... - swissinfo.ch	http://news.google.com/news/url?sa=t&fd=R&ct2=es&usg=AFQjCNC	4/23/16 10:26	Breaking		
14	Colombia	Colombia	Zika virus	Humans		4178008	Colombia: #Zika-related #microcephaly cases rise to to four https://t.co/dGdFJER1	https://twitter.com/crof/statuses/723668584816218112	4/22/16 20:23	Breaking		
15	Departamen	Colombia	Zika virus	Humans	es	4177448	Confirman dos casos de microcefalia relacionados con Zika en ... - W Radio	http://news.google.com/news/url?sa=t&fd=R&ct2=es&usg=AFQjCNH	4/22/16 11:09	Breaking		
16	Departamen	Colombia	Zika virus	Humans	es	4177445	Confirman dos casos de microcefalia asociados al Zika en Santander - Blu Radio	http://news.google.com/news/url?sa=t&fd=R&ct2=es&usg=AFQjCNE	4/22/16 10:48	Breaking		
17	Colombia	Colombia	Zika virus	Humans		4175936	Cumulative Locally Acquired Zika Cases by Country in the Americas, 2015-2016 as c	http://reliefweb.int/node/1493151	4/21/16 0:00	Breaking	65338	
18	Colombia	Colombia	Zika virus	Humans		4175395	RT @RonanKelly13: PAHO #Zika country by country case counts https://t.co/U3Um	https://twitter.com/FluTrackers/statuses/723252572031725568	4/21/16 16:50	Breaking	65338	
19	Bucaramang	Colombia	Zika virus	Humans	es	4173388	Comuna uno de Bucaramanga lidera casos de zika, chikunguña y ... - RCN Ra	http://news.google.com/news/url?sa=t&fd=R&ct2=es&usg=AFQjCNE	4/20/16 20:28	Breaking		
20	Colombia	Colombia	Zika virus	Humans	pt	4171146	PRO/PORT> Zika vì_rus - Coll«mbia (06), nì¼mero de casos, atualizaìtÌ£o	http://promedmail.org/direct.php?id=20160419.4171146	4/19/16 23:52	Breaking	68630	
21	Colombia	Colombia	Zika virus	Humans		4168370	PRO/EDR> Zika virus (23): Americas	http://promedmail.org/direct.php?id=20160419.4168370	4/19/16 4:19	Breaking		
22	Colombia	Colombia	Zika virus	Humans	es	4168416	Aumentaron a 239 los casos del virus zika: Ministerio de Salud - La Prensa	http://news.google.com/news/url?sa=t&fd=R&ct2=es&usg=AFQjCNH	4/18/16 19:19	Breaking		
23	Colombia	Colombia	Zika virus	Humans	es	4167468	Colombia reporta casos de microcefalia asociados a zika - Servicio de noticias en sa	http://boletinaldia.sld.cu/aldia/2016/04/17/colombia-reporta-casos-	4/17/16 2:03	Breaking		
24	Colombia	Colombia	Zika virus	Humans		4165683	Zika numbers down in Colombia, but infectiousness concerns - The City Paper Bogo	http://news.google.com/news/url?sa=t&fd=R&ct2=us&usg=AFQjCNE	4/17/16 12:38	Breaking		
25	Colombia	Colombia	Zika virus	Humans	es	4164152	Zika: Colombia alcanza casi 70000 infectados por el virus - Diario Perú21	http://news.google.com/news/url?sa=t&fd=R&ct2=es&usg=AFQjCNF	4/16/16 11:14	Breaking	68630	
26	Colombia	Colombia	Zika virus	Humans		4163945	#Zika in Colombia, week 14: 2,061 new cases https://t.co/UUjv8JPbg7	https://twitter.com/crof/statuses/721348286330724352	4/16/16 10:43	Breaking	65338	
27	Colombia	Colombia	Zika virus	Humans		4163069	PAHO:Cumulative #Zika suspected&confirmed cases reported by countries &a	https://twitter.com/FluTrackers/statuses/721132188054470656	4/15/16 20:24	Breaking	61778	
28	Bucaramang	Colombia	Zika virus	Humans	es	4162882	Bucaramanga supera los mil casos del virus del zika - RCN Radio (blog)	http://news.google.com/news/url?sa=t&fd=R&ct2=es&usg=AFQjCNE	4/15/16 16:52	Breaking		
29	Colombia	Colombia	Zika virus	Humans		4162576	'Colombia Reports First Cases of Microcephaly Linked to Zika Virus' https://t.co/Vq	https://twitter.com/EcoHealthNYC/statuses/721036457096450048	4/15/16 14:04	Breaking		
30	Colombia	Colombia	Zika virus	Humans	ar	4161696	ðÃôöõãõöõÉˆ¬ô_˜¤ ˜»˜_ôÃ˜ø ˆéôöõã ˜_˜¤ôã˜»ô_˜µ˜¼˜± ˆÃôÉˆÃôÉˆ© ˜_ôãõä ˜µô	http://www.daralakhbar.com/articles/10138253-%D8%A7%D9%84%	4/15/16 5:00	Breaking	0	
31	Colombia	Colombia	Zika virus	Humans	zh	4161535	ŒÒ´Š__¾øÓŠ¼_ß¤ß_Œø¬Œß¨_Šü_Š_ÜŠü_Œø¬Œß¤Á¾ÍãŒÉ_Œ¡ß_«ß_Ñàß_ÑÉŠ_Ü	http://news.21cn.com/hotnews/a/2016/0415/17/30917510.shtml	4/15/16 6:05	Breaking		
32	Colombia	Colombia	Zika virus	Humans		4161216	RT @Reuters_Health: Colombia says two cases of microcephaly linked to Zika virus	https://twitter.com/makoto_au_japon/statuses/720855108917944	4/15/16 2:03	Breaking		
33	Colombia	Colombia	Zika virus	Humans		4161004	Colombia confirms first two cases of #Zika-associated #microcephaly https://t.co/t	https://twitter.com/crof/statuses/720813472603197440	4/14/16 23:18	Breaking		
34	Colombia	Colombia	Zika virus	Humans	pt	4160406	Colômbia confirma dois casos de microcefalia ligados ao Zika vírus - T	http://news.google.com/news/url?sa=t&fd=R&ct2=pt-PT_pt&usg=AF	4/14/16 14:53	Breaking		
35	Colombia	Colombia	Zika virus	Humans		4160198	Colombia confirms two microcephaly cases - SBS	http://news.google.com/news/url?sa=t&fd=R&ct2=us&usg=AFQjCNH	4/14/16 13:34	Breaking		
36	Colombia	Colombia	Zika virus	Humans		4159936	Colombia says two cases of microcephaly linked to Zika virus https://t.co/lzNfapga	https://twitter.com/Reuters_Health/statuses/720643823358078977	4/14/16 12:04	Breaking		
37	Colombia	Colombia	Zika virus	Humans	es	4159836	Colombia confirma dos casos de microcefalia relacionados con ... - swissinfo.ch	http://news.google.com/news/url?sa=t&fd=R&ct2=es&usg=AFQjCNE	4/14/16 11:41	Breaking		
38	Colombia	Colombia	Zika virus	Humans	es	4159834	Colombia confirma dos casos de microcefalia asociados al virus zika - El Universo	http://news.google.com/news/url?sa=t&fd=R&ct2=es&usg=AFQjCNF	4/14/16 11:37	Breaking		
39	Colombia	Colombia	Zika virus	Humans	es	4153833	Colombia investiga 30 casos de microcefalia sospechosos de estar asociados con zi	http://boletinaldia.sld.cu/aldia/2016/04/10/colombia-investiga-30-c	4/10/16 2:05	Breaking		
40	Colombia	Colombia	Zika virus	Humans	es	4151869	Relacion zika con microcefalia - ElTiempo.com	http://news.google.com/news/url?sa=t&fd=R&ct2=es&usg=AFQjCNC	4/9/16 12:11	Breaking	64839	
41	Colombia	Colombia	Zika virus	Humans	vi	4151436	30 tr®¡‡È ng h‡È£p tr‡¼È €Ô‡¼¤u nh‡È ‡ÈÙ Colombia nghi liên quan €Ô‡¼	http://news.google.com/news/url?sa=t&fd=R&ct2=vi_vn&usg=AFQj	4/9/16 22:06	Breaking	64839	

HM_Zika_04282016

Explore Human Mobility in the US





	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1		Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	Florida	Georgia	Hawaii	Idaho	Illinois	Indiana	Iowa	Kansas	Kentucky	Louisiana	Maine	Maryland	Massachuset	Michigan
2	sub-region																						
3	country	United State	United State	United State	United State	United State	United State	United State	United State	United State	United State	United State	United State	United State	United State	United State	United State	United State	United State	United State	United State	United State	United State
4	1/13/20	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
5	1/14/20	102.9	104.69	104.06	102.55	104.39	103.46	105.22	99.97	102.25	104.66	102.45	99.75	103.65	102.7	104.67	107.09	104.27	95.97	104.9	104.86	103.04	102.13
6	1/15/20	103.51	104.64	106.9	101.93	109.34	105.66	108.31	102.53	104.85	105.64	108.07	106.43	103.89	103.12	98.34	105.54	103.27	89.05	106.6	108.01	108.19	103.23
7	1/16/20	109.71	110.1	111.45	106.38	109.36	109.36	109.35	104.41	108.92	112.08	107.52	109.34	108.53	105.52	112.88	111.25	107.79	93	77.37	109.94	108.23	106.13
8	1/17/20	133.97	126.32	129.15	124.97	129.78	130.55	129.29	121.44	126.69	130.73	119.11	132.6	126.87	123.72	74.14	92.27	123.87	111.95	129.84	126.08	124.68	129.07
9	1/18/20	118.17	122.44	124.63	124.45	122.26	121.37	99.31	100.13	118.8	119.62	119.26	124.65	112.45	112.11	85.11	130.55	114.81	105.54	119.18	94.76	104.39	87.73
10	1/19/20	89.3	91.11	91.76	88.02	94.19	89.84	92.46	87.43	89.04	91.96	100.47	82.54	88.68	83.59	85.68	86.61	85.06	77.39	91.58	88.73	86.56	80.96
11	1/20/20	107.34	98.78	102.98	103.18	103.49	103.79	108.97	98.63	102.53	102.47	103.9	108.31	101.72	99.55	104.01	104.05	98.72	86.05	113.32	97.77	99.37	97.68
12	1/21/20	101.15	104.55	103.8	100.14	107.08	99.12	103.64	97.53	103.07	101.97	105.08	103.59	101.67	98.95	104.56	103.8	100.02	87.81	103.7	101.54	101.93	101.46
13	1/22/20	100.33	104.3	105	90.55	109.41	101.85	104.64	99.23	101.36	102.67	105.37	107.4	100.94	102.09	90.28	90.96	101.04	86.62	104.6	104.35	103.24	101.84
14	1/23/20	105.53	105.26	108.99	101.84	114.18	107.31	108.24	100.23	107.03	106.62	109.51	109.86	103.07	104.58	100.18	100.83	106.63	89.39	108.33	107.49	105.74	104.25
15	1/24/20	125.69	122.92	123.39	122.27	128.14	124.4	127.19	120.22	122.59	123.23	114.03	129.68	121.44	122.44	121.42	116.58	123.98	108.9	128.78	122.94	122.61	123.34
16	1/25/20	122.09	116.96	118.69	122.81	122.28	119.05	118.06	106.07	117.41	122.27	114.08	126.91	119.73	118.33	140.12	128.15	118.56	104.08	125.65	118.02	113.86	117.9
17	1/26/20	82.45	84.2	83.08	82.69	91.63	90.01	91.38	81.85	85.26	83.75	92.01	79.81	86.67	83.79	98.11	87.43	81.56	65.51	95.58	87.03	87.5	80.66
18	1/27/20	98.85	99.6	100.67	97.97	101.95	94.06	95.91	93.66	97.05	97.38	99.14	99.73	94.56	93.71	99.4	99.92	94.83	82.81	96.62	98.29	93.55	94.66
19	1/28/20	101.6	99.76	105.22	98.5	105.89	100.08	99.36	94.39	100.96	104.19	100.55	103.35	98.35	99.09	104.99	94.92	98.48	85.28	100.02	100.38	98.28	98.72
20	1/29/20	97.95	102.31	107.72	98.35	106.63	98.41	101.77	98.65	101.79	101.5	99.61	107.34	99.51	99.91	103.46	101.57	100.92	85.55	101.7	102.43	101.45	100.04
21	1/30/20	109.32	105.06	114.04	104.2	110.97	106.61	107.71	101.07	108.85	110.13	106.35	110.84	104.9	103.38	112.74	108.01	105.46	90.12	105.79	104.82	104.81	105
22	1/31/20	129.43	122.27	129.78	125.55	128.51	123.48	127.19	120.47	125.23	125.99	111.46	132.93	125.43	123.9	132.92	127.5	125.2	111.19	128.76	123.44	121.44	123.99
23	2/1/20	123.78	119.04	122.68	121.08	121.9	116.4	119.92	111.72	115.1	126.66	110.86	131.66	124.86	121.93	143.76	132.8	119.55	108.78	130.2	119.11	114.42	117.12
24	2/2/20	87.78	82.5	81.91	85.56	84.06	85.08	86.23	81.44	86.35	87.47	82.55	78.19	87.42	84.21	97.64	83.42	85.99	69.82	95.72	84.57	81.4	82.26
25	2/3/20	100.84	101.77	102.53	96.9	100.77	85.85	98.65	101.1	101.79	101.6	98.52	101.37	96.24	96.15	105.72	97.73	98.24	84.04	98.81	98.24	93.87	96.44
26	2/4/20	101.5	102.76	106.31	98.34	105.21	81.32	101.23	96.92	103.39	104.08	100.37	105.8	98.86	100.2	103.05	101.22	102.5	87.38	101.93	104.19	99.21	99.24
27	2/5/20	97.22	106.87	105.73	92.08	108.44	99.4	101.17	98.86	105.55	103.38	101.73	106.39	101.67	101.05	103.55	96.02	104	83.74	106.6	101.22	100.86	101.27
28	2/6/20	105.55	113.41	111.19	103.24	114.26	111.29	101.88	101.58	111.7	103.32	105.01	110.53	104.18	100.79	111.1	106.24	107.91	89.2	81.01	105.99	101.59	105.92
29	2/7/20	127.88	131.27	130.39	122.62	130.28	104.75	121.7	115.57	124.22	124.8	110.56	139	119.92	119.66	125.42	122.6	115.24	112.79	87.62	117.69	119.79	119.66
30	2/8/20	121.14	123.14	128.05	123.14	125.41	119.48	124.01	114.67	119.36	105.49	115.33	131.44	126	121.31	140.46	128.67	113.47	110.46	131.45	118.74	116.76	116.63
31	2/9/20	91.5	94.15	90.51	83.08	93.45	80.31	92.74	92.84	91.8	93.49	95.47	88.52	89.94	88.3	96.27	91.15	87.1	80.8	105.54	92.35	89.38	85.41
32	2/10/20	100.76	102.66	104	99.62	107.18	103.81	98.74	101.73	104.87	105.57	111.39	109.58	99.95	100.93	100.75	101.18	100.67	84.45	97.32	102.48	96.08	98.71
33	2/11/20	106.25	106.55	106.8	103.37	111.65	110.09	102.86	103.31	108.99	107.78	102.34	112.31	106.5	105.9	107.14	107.49	106.22	91.05	105.15	105.54	102.35	104.27
34	2/12/20	106.59	105.08	112.99	101.89	114.05	110.41	104.57	104.76	113.63	110.78	105.83	116.49	107.08	103.43	102.62	95.19	109.37	92.59	111.35	107.95	105.67	105.68
35	2/13/20	122.67	118.67	122.15	119.57	124.88	124.65	111.47	112.45	121.7	120.74	112.84	128.54	115.23	107.51	102.7	107.69	120.79	102.7	93.09	116.98	108.94	109.25
36	2/14/20	152.62	142.33	143.8	145.81	150.11	144.99	141.11	140.57	141.19	155.74	130.31	153.94	146.13	136.76	143.87	140.81	148.93	137.45	141.78	142.95	134.08	137.71
37	2/15/20	136.69	130.53	130.54	133.39	130.03	128.72	125.78	127.95	128.4	136.97	122.54	140.08	139.61	132.04	149.58	141.5	135.15	127.11	145.36	128.17	119.11	125.55
38	2/16/20	97.66	102.35	100.38	97.96	106.16	100.63	99.94	99.1	99.36	96.4	107.65	94.74	109.95	97.67	108.26	102.77	96.72	83.59	122.48	98.95	95.64	93.88
39	2/17/20	113.36	108.8	112.58	106.32	111.6	107.58	114.81	115.17	112.63	110.76	107.84	121.8	111.35	107.12	106.83	113.02	110.72	91.46	125.21	109.68	101.09	103.71
40	2/18/20	108.61	106.72	112.31	103.31	113.94	103.31	108.16	106.39	113.95	109.03	109.6	113.53	105.53	10								4.44
41	2/19/20	112.41	108.95	117.05	106.84	114.47	110.41	111.61	105.98	113.83	111.16	111.44	115.28	106.8	10								6.67
42	2/20/20	116.53	112.84	121.7	113.83	120.36	118.35	114.62	109.45	118.75	111.14	111.54	125.84	112.35	11								0.91
43	2/21/20	148.9	131.76	138.9	137.15	137.45	135.04	137.07	131.99	132.84	136.71	124.09	148.6	134.5	13								2.21
44	2/22/20	141.79	129.62	131.76	136.22	129.22	127.98	134.49	125.32	126.69	138.1	122.8	149.1	139.75	13								3.39
45	2/23/20	100.77	96.81	104.95	90.93	103.92	87.4	105.14	94.85	96.35	95.06	102.46	95.66	102.31	9								3.68
46	2/24/20	109.77	111.14	115.16	104.05	111.7	102.7	103.36	101.14	107.78	103.06	108.47	113.61	101.12	10								0.56
47	2/25/20	115.64	112.51	117.05	111.79	114.15	100.77	105.81	102.77	111.31	111.9	107.84	116.21	101.45	10								5.98
48	2/26/20	116.12	112.52	117.74	113.3	115.08	108.61	111.97	107.03	116.43	116.48	108.22	117.12	104.36	10								1.48
49	2/27/20	125.69	119.63	123.17	121.56	119.27	114.51	114.15	112.35	121.17	124.65	109.39	129.78	115.5	11								1.66
50	2/28/20	152.84	139.78	139.76	150.13	136.49	134.84	135.71	134.82	139.23	145.94	120.58	155.62	140.27	13								6.38
51	2/29/20	149.38	136.2	144.6	152.21	131.51	126.96	135.21	130.78	132.22	152.6	118.18	150.02	144.06	14								6.38
52	3/1/20	104.53	97.47	103.5	101.54	98.24	94.51	101.92	96.71	100.16	104.44	97.3	97.61	107.71	10								6.19

Google COVID-19 Community Mobility Reports

See how your community is moving around differently due to COVID-19

As global communities respond to COVID-19, we've heard from public health officials that the same type of aggregated, anonymized insights we use in products such as Google Maps could be helpful as they make critical decisions to combat COVID-19.

These Community Mobility Reports aim to provide insights into what has changed in response to policies aimed at combating COVID-19. The reports chart movement trends over time by geography, across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential.

Google

COVID-19 Community Mobility Report

Washington July 19, 2020

Mobility changes

This dataset is intended to help remediate the impact of COVID-19. It shouldn't be used for medical diagnostic, prognostic, or treatment purposes. It also isn't intended to be used for guidance on personal travel plans.

Each Community Mobility Report dataset is presented by location and highlights the percent change in visits to places like grocery stores and parks within a geographic area. How to use this report.

Location accuracy and the understanding of categorized places varies from region to region, so we don't recommend using this data to compare changes between countries, or between regions with different characteristics (e.g. rural versus urban areas).

We'll leave a region out of the report if we don't have statistically significant levels of data. To learn how we calculate these trends and preserve privacy, read About this data.

Retail & recreation

-9%

compared to baseline

Mobility trends for places like restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters.

Grocery & pharmacy

+1%

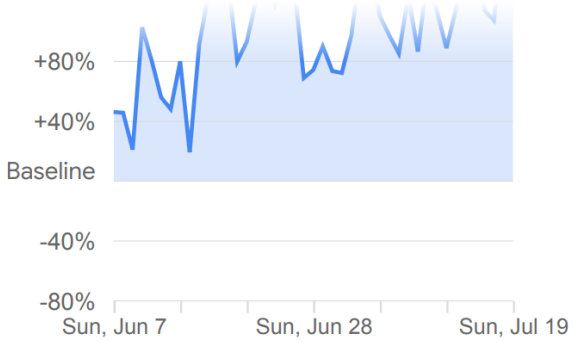
compared to baseline

Mobility trends for places like grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies.

Parks

+138%

compared to baseline



Mobility trends for places like national parks, public beaches, marinas, dog parks, plazas, and public gardens.

Transit stations

-17%

compared to baseline

Mobility trends for places like public transport hubs such as subway, bus, and train stations.

Workplaces

-16%

compared to baseline

Mobility trends for places of work.

Residential

+1%

compared to baseline

Mobility trends for places of residence.

Weather data

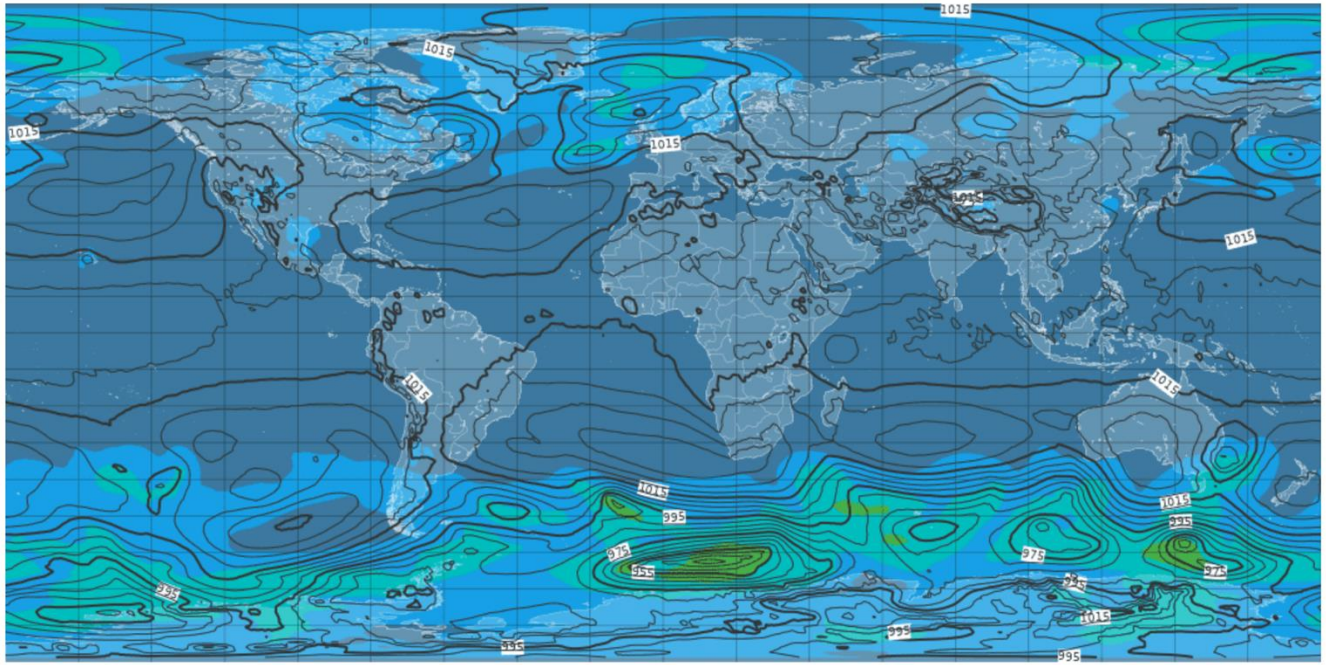
You can obtain data from the European Centre for Medium-Range Weather Forecasts (ECMWF):

ECMWF

Search site... Contact Log in

Home About Forecasts Computing Research Learning Publications

Advancing global NWP through international collaboration



High resolution mean sea level pressure and ensemble spread

Thursday 23 July, 00 UTC T+96 Valid: Monday 27 July, 00 UTC

Ensemble forecasts explained

One 'ensemble forecast' consists of 51 separate forecasts made by the same computer model, all activated from the same starting time. The starting conditions for each member of the ensemble are slightly different, and physical parameter values used also differ slightly. The differences between these ensemble members tend to grow as the forecasts progress, that is as the forecast lead time increases.

View all charts ›

Climate reanalysis

Coupled Earth-system reanalysis

Reanalysis for climate monitoring

Ocean reanalysis

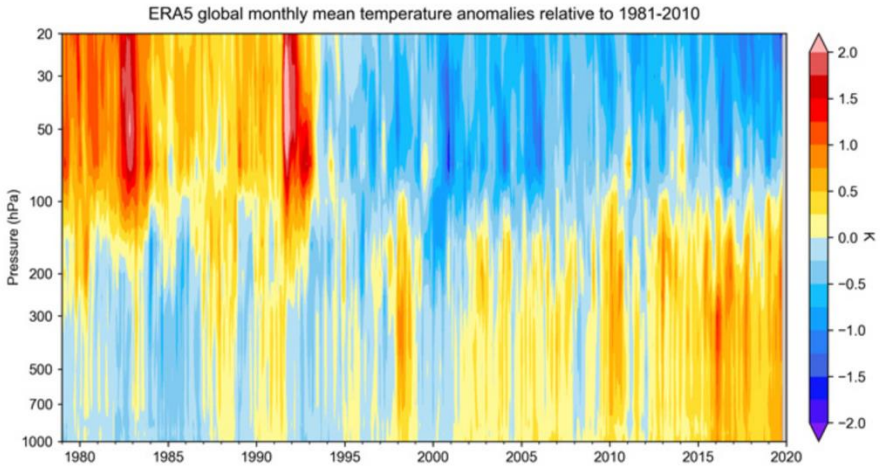
CAMS Reanalysis

ECMWF periodically uses its forecast models and data assimilation systems to 'reanalyse' archived observations, creating global data sets describing the recent history of the atmosphere, land surface, and oceans. The essence of this process, which provides consistent and convenient 'maps without gaps', is explained in this animation.

Reanalysis is very popular and is used for monitoring climate change, for research and education, and for commercial applications.

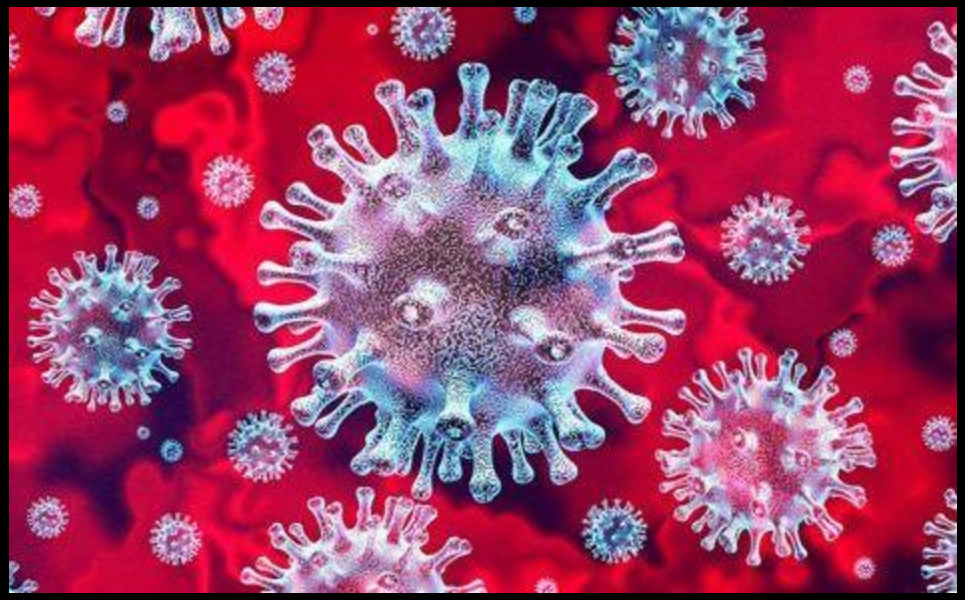
The latest ECMWF reanalysis is ERA5, which is being produced by C3S. ERA5 provides a snapshot of the atmosphere, land surface and ocean waves for each hour from 1979 onwards (and eventually from 1950). It includes an uncertainty estimate which highlights the considerable evolution of the observing system, on which reanalysis products rely.

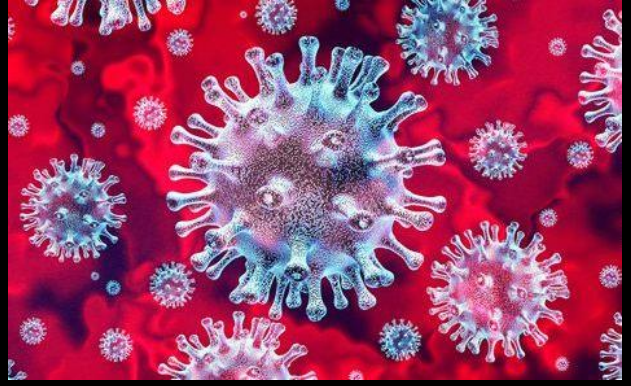
Browse the reanalysis datasets >>



Fourth session started here

COVID-19





Coronavirus Disease (COVID - 19) Year of detection **2019**

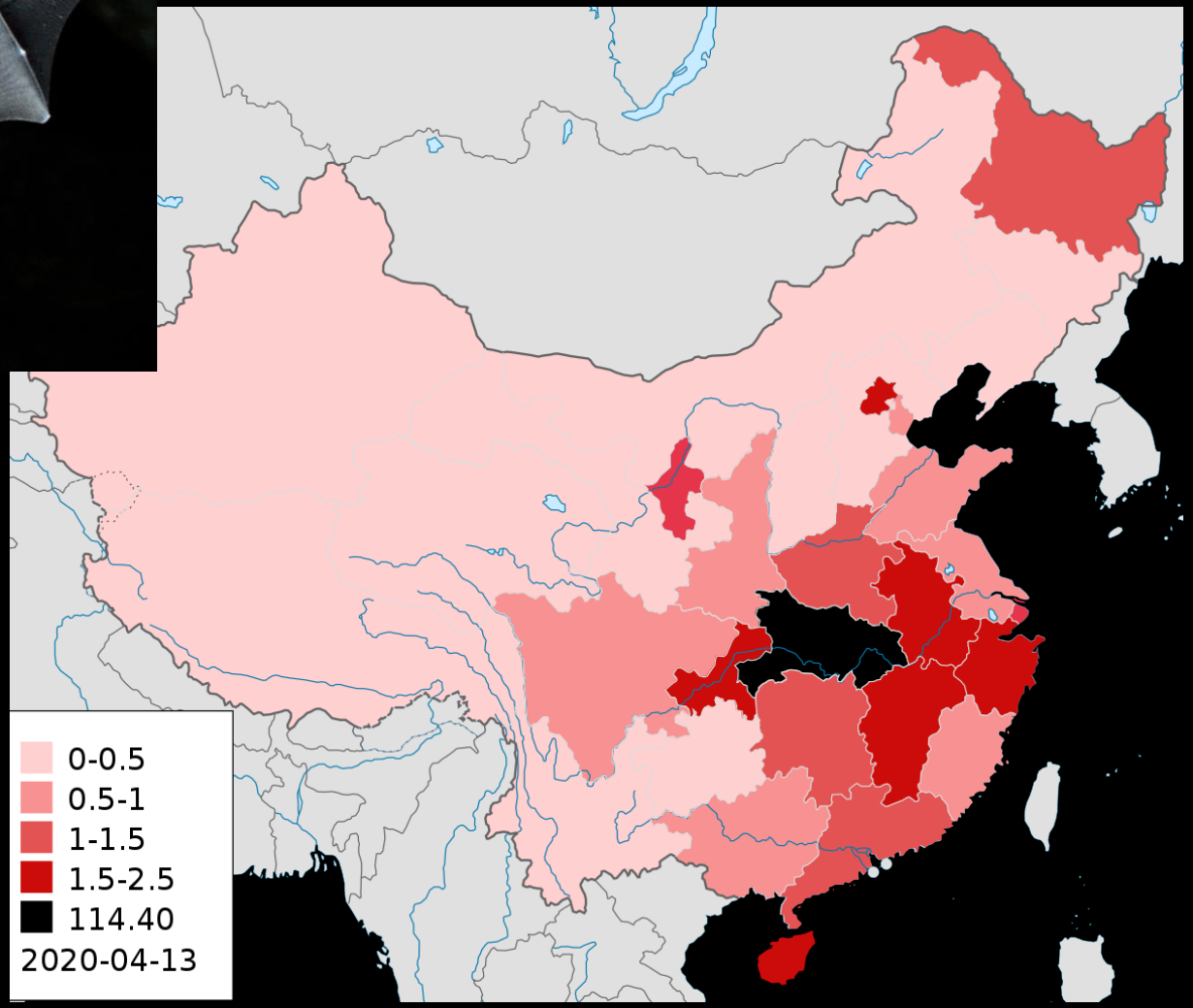
Caused by: severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)

It was first identified in December 2019 in Wuhan, China, and has resulted in an ongoing pandemic.[10][11] The first case may be traced back to 17 November 2019.[12] As of 17 June 2020, more than **8.18 million** cases have been reported across 188 countries and territories, resulting in more than **443,000 deaths**.

Common symptoms include fever, cough, fatigue, shortness of breath, and loss of smell and taste.

Source: Wikipedia

First detected in **China**



Source:
Wikipedia

Published on 17.8.2020 in Vol 22, No 8 (2020): August

Preprints (earlier versions) of this paper are available at https://preprints.jmir.org/preprint/20285, first published May 14, 2020.

Real-Time Forecasting of the COVID-19 Outbreak in Chinese Provinces: Machine Learning Approach Using Novel Digital Data and Estimates From Mechanistic Models

Dianbo Liu [1,2]; Leonardo Clemente [1,2,3]; Canelle Poirier [1,2]; Xiyu Ding [1,4]; Matteo Chinazzi [5]; Jessica Davis [5]; Alessandro Vespignani [5,6]; Mauricio Santillana [1,2,4]

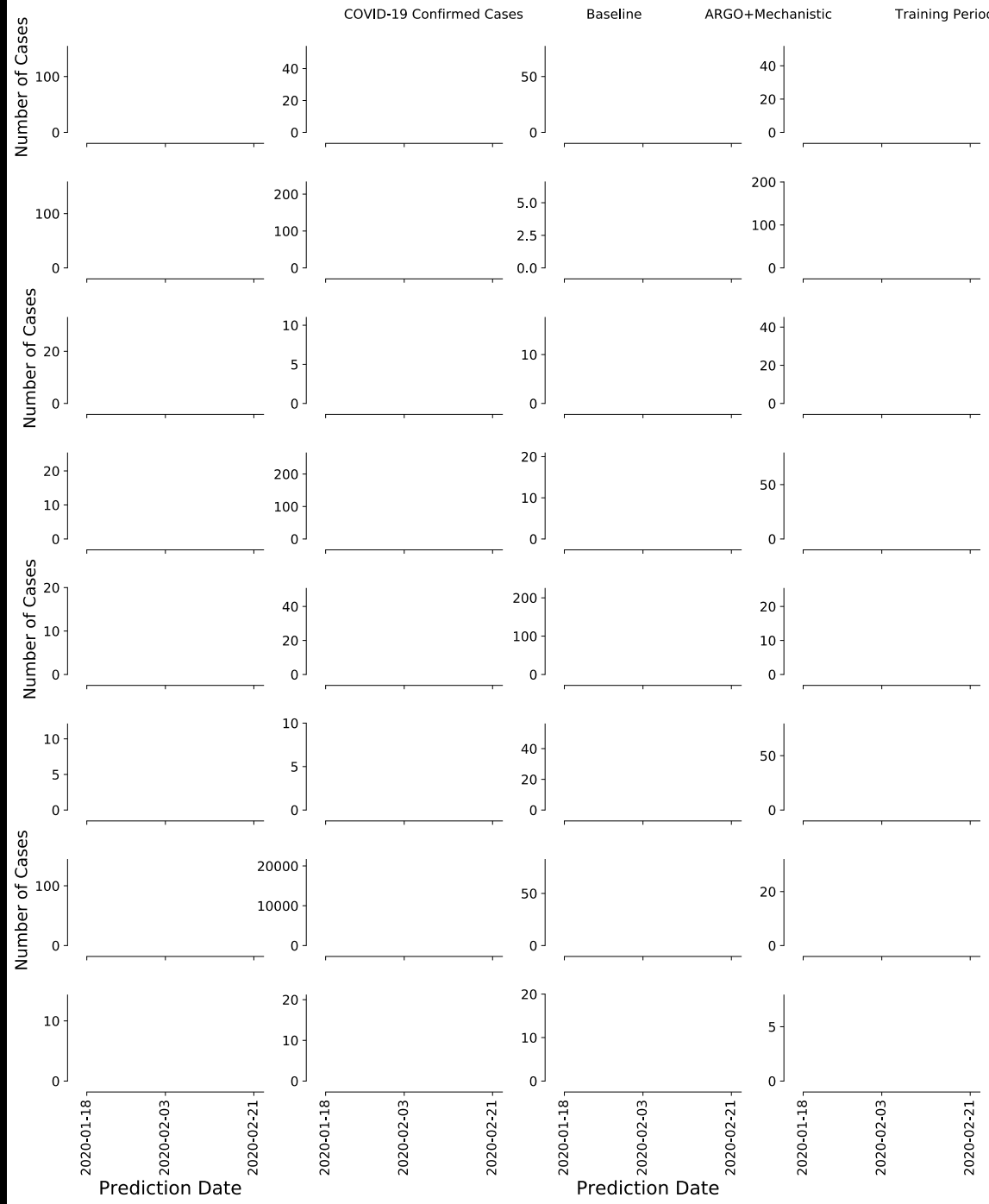
Related Article

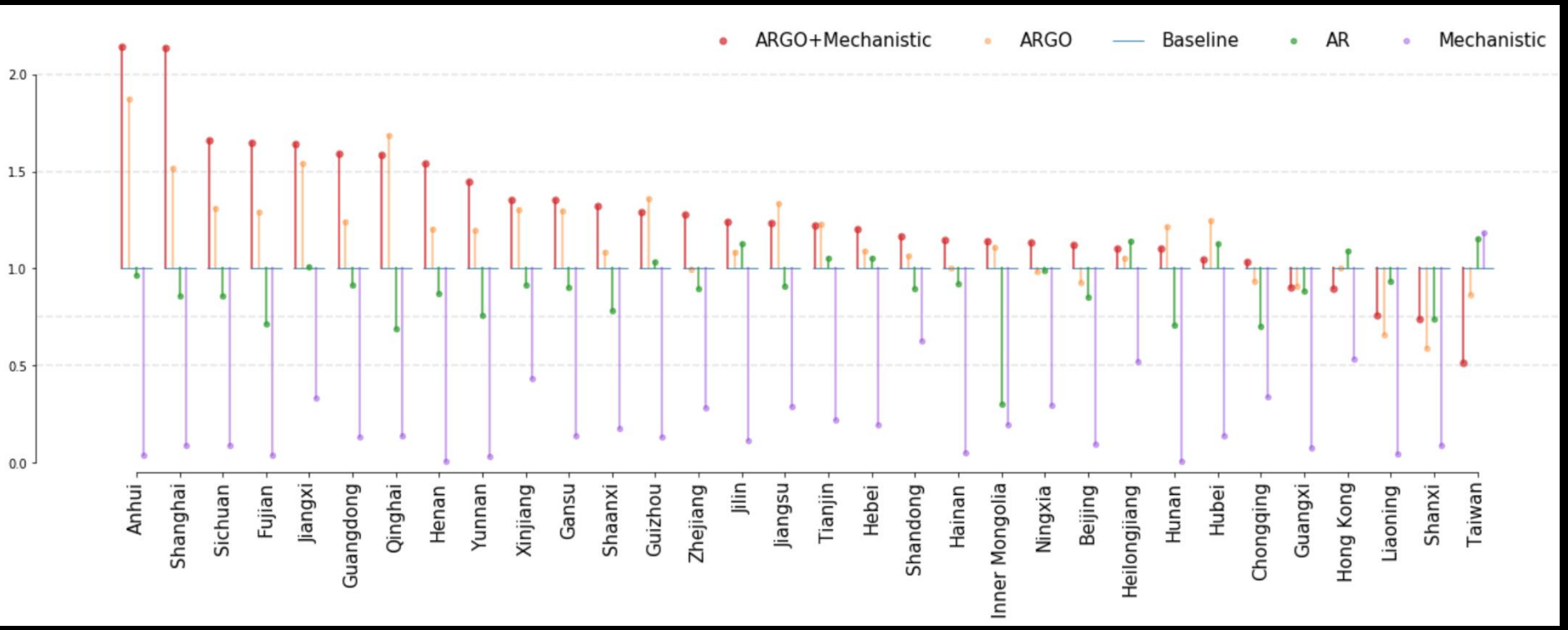
This is a corrected version. See correction statement in: https://www.jmir.org/2020/9/e23996/

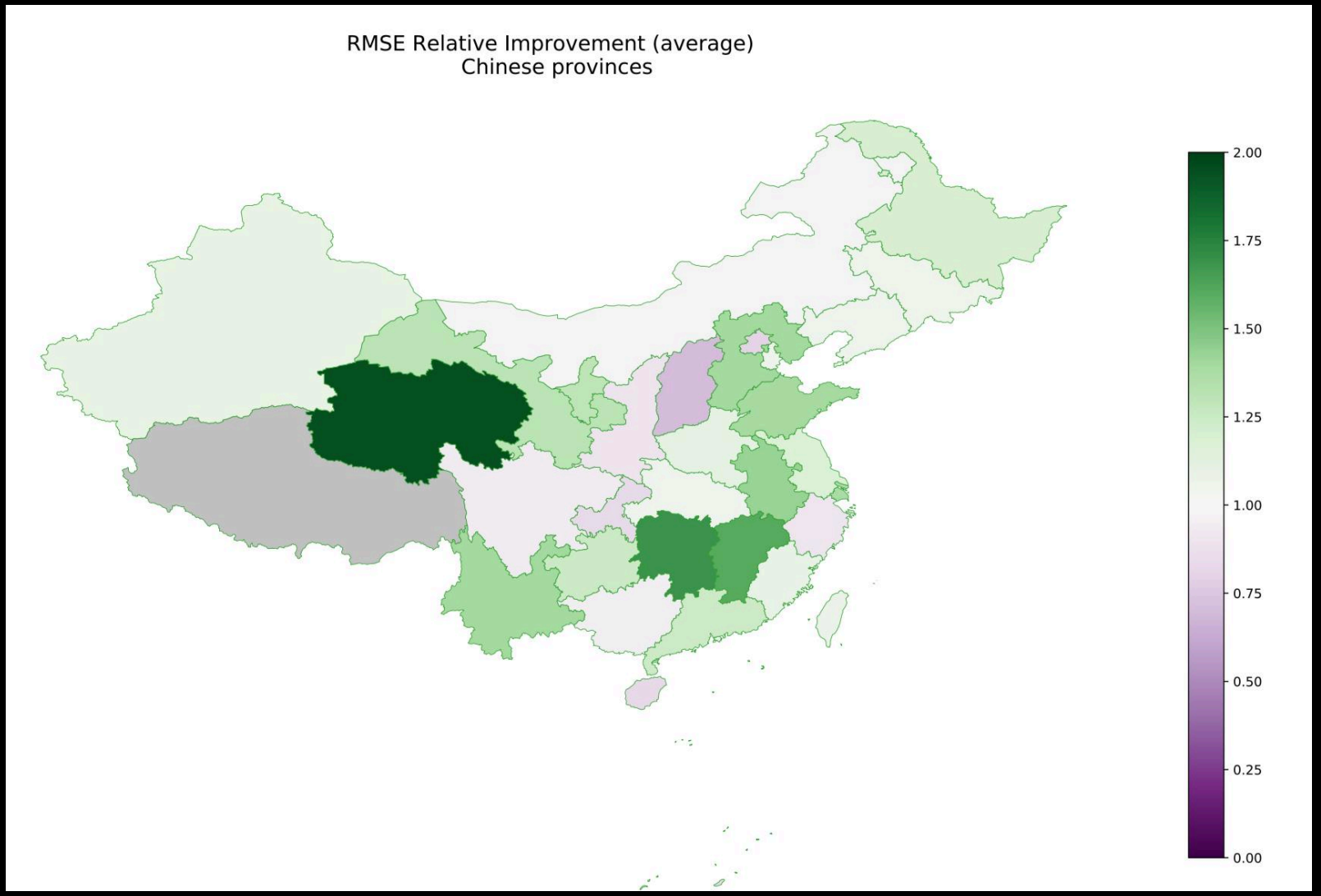
Abstract

Background:
The inherent difficulty of identifying and monitoring emerging outbreaks caused by novel pathogens can lead to their rapid spread; and if left unchecked, they may become major public health threats to the planet. The ongoing coronavirus disease (COVID-19) outbreak, which has infected over 2,300,000 individuals and caused over 150,000 deaths, is an example of one of these catastrophic events.

Objective:
We present a timely and novel methodology that combines disease estimates from mechanistic models and digital traces, via interpretable machine learning methodologies, to reliably forecast COVID-19 activity in Chinese provinces in real time.


COVID-19 Confirmed Cases
Baseline
ARGO+Mechanistic
Training Period
Number of Cases
Prediction Date
2020-01-18
2020-02-03
2020-02-21


ARGO+Mechanistic
ARGO
Baseline
AR
Mechanistic
2.0
1.5
1.0
0.5
0.0
Anhui
Shanghai
Sichuan
Fujian
Jiangxi
Guangdong
Qinghai
Henan
Yunnan
Xinjiang
Gansu
Shaanxi
Guizhou
Zhejiang
Jilin
Jiangsu
Tianjin
Hebei
Shandong
Hainan
Inner Mongolia
Ningxia
Beijing
Heilongjiang
Hunan
Hubei
Chongqing
Guangxi
Hong Kong
Liaoning
Shanxi
Taiwan


RMSE Relative Improvement (average)
Chinese provinces
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

Article | Published: 04 May 2020

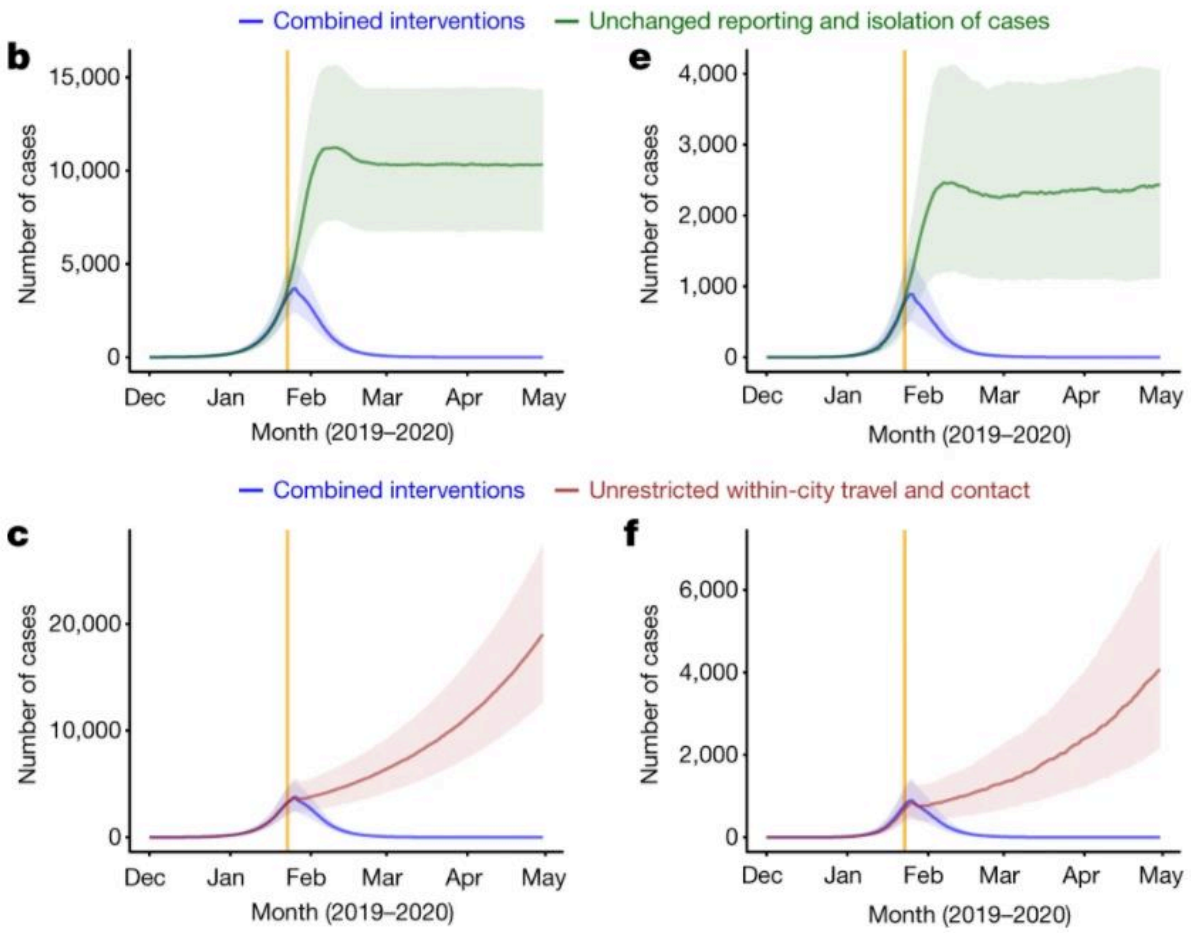
Effect of non-pharmaceutical interventions to contain COVID-19 in China

Shengjie Lai, Nick W. Ruktanonchai, Liangcai Zhou, Olivia Prosper, Wei Luo, Jessica R. Floyd, Amy Wesolowski, Mauricio Santillana, Chi Zhang, Xiangjun Du, Hongjie Yu & Andrew J. Tatem

Nature **585**, 410–413 (2020) | Cite this article

Abstract

On 11 March 2020, the World Health Organization (WHO) declared coronavirus disease 2019 (COVID-19) a pandemic[1]. The strategies based on non-pharmaceutical interventions that were used to contain the outbreak in China appear to be effective[2], but quantitative research is still needed to assess the efficacy of non-pharmaceutical interventions and their timings[3]. Here, using epidemiological data on COVID-19 and anonymized data on human movement[4,5], we develop a modelling framework that uses daily travel networks to simulate different outbreak and intervention scenarios across China. We estimate that there were a total of 114,325 cases of COVID-19 (interquartile range 76,776–164,576) in mainland China as of 29 February 2020. Without non-pharmaceutical interventions, we predict that the number of cases would have been 67-fold higher (interquartile range 44–94-fold) by 29 February 2020, and we find that the effectiveness of different interventions varied. We estimate that early detection and isolation of cases prevented more infections than did travel restrictions and contact reductions, but that a combination of non-pharmaceutical interventions achieved the strongest and most rapid effect. According to our model, the



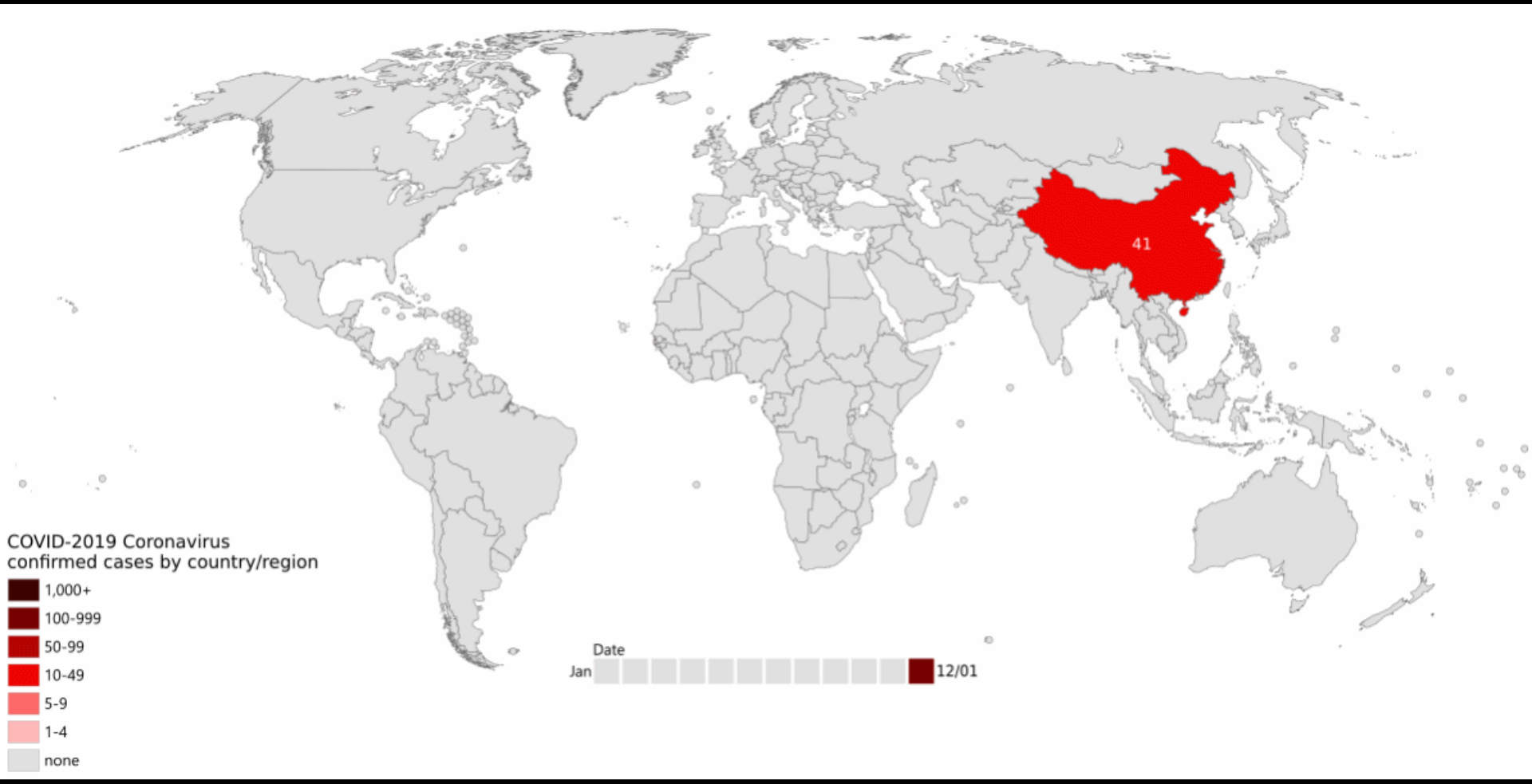
a–**c**, Estimates for the city of Wuhan. **d**–**f**, Estimates for cities outside of Hubei province in mainland China. The blue lines represent estimated transmission under combined NPIs, and the other coloured lines represent the scenario without one type of intervention. Data are presented as the median (solid line) and IQR (shading) of estimates (1,000 simulations). The orange vertical lines indicate the date on which the lockdown of Wuhan began (23 January 2020).

$$\frac{\mathrm{d}S}{\mathrm{d}t} = S - c\frac{SI}{N}$$

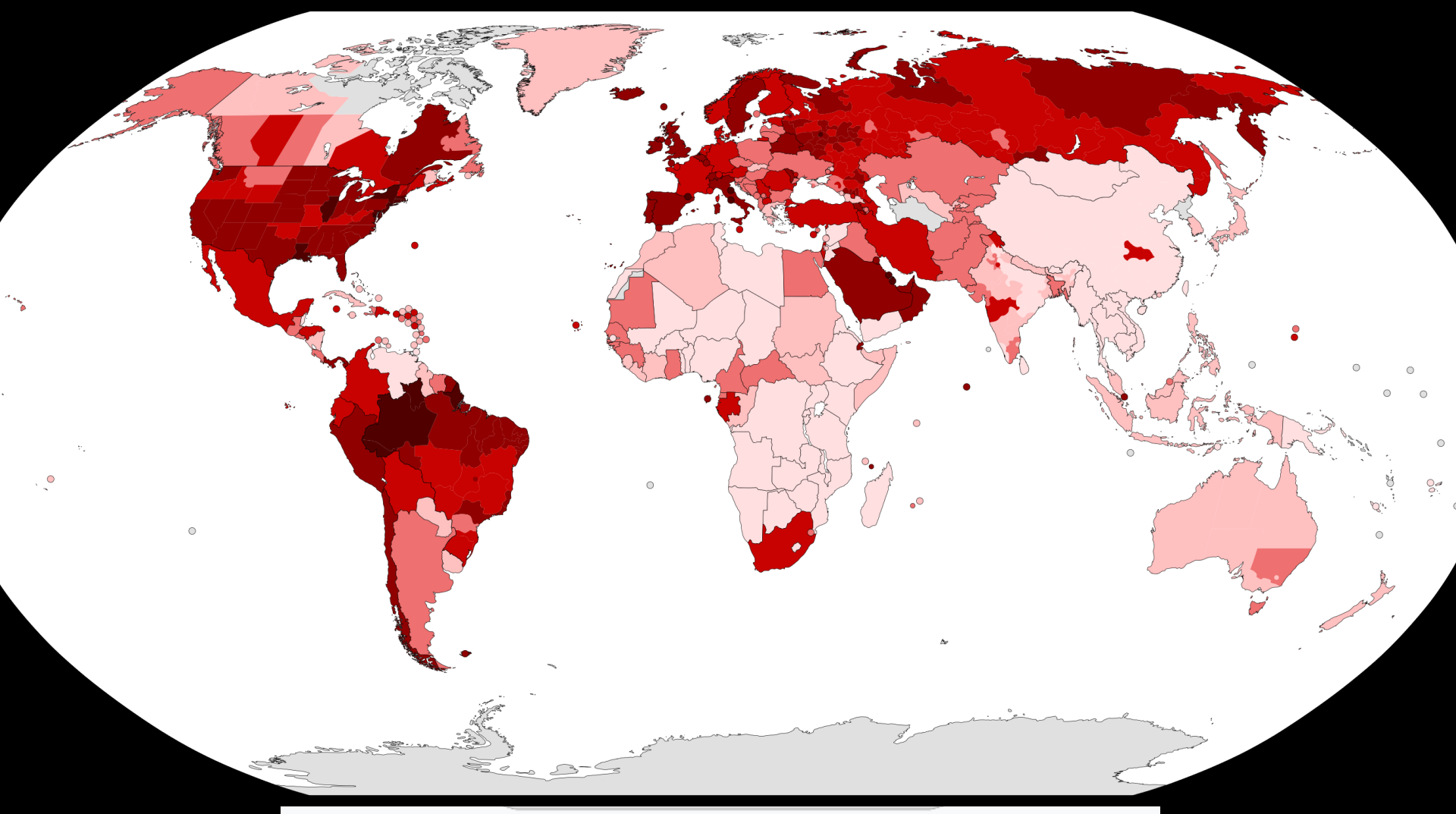
$$\frac{\mathrm{d}E}{\mathrm{d}t} = c\frac{SI}{N} - \varepsilon E$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \varepsilon E - rI$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = rI$$



Source:
Wikipedia



Source:
Wikipedia

scientific reports

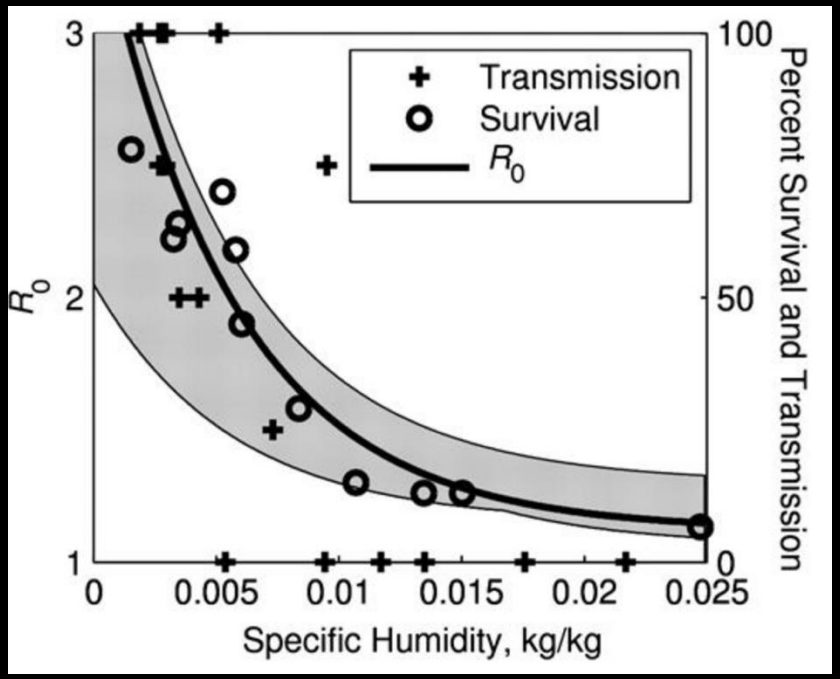
Article | Open Access | Published: 12 October 2020

The role of environmental factors on transmission rates of the COVID-19 outbreak: an initial assessment in two spatial scales

Canelle Poirier ✉, Wei Luo, Maimuna S. Majumder, Dianbo Liu, Kenneth D. Mandl, Todd A. Mooring & Mauricio Santillana ✉

Scientific Reports **10**, Article number: 17002 (2020) | Cite this article

Background



*"[…] the timing of **pandemic influenza** outbreaks is controlled by a combination of absolute humidity conditions, levels of susceptibility, and changes in population-mixing and contact rates."*

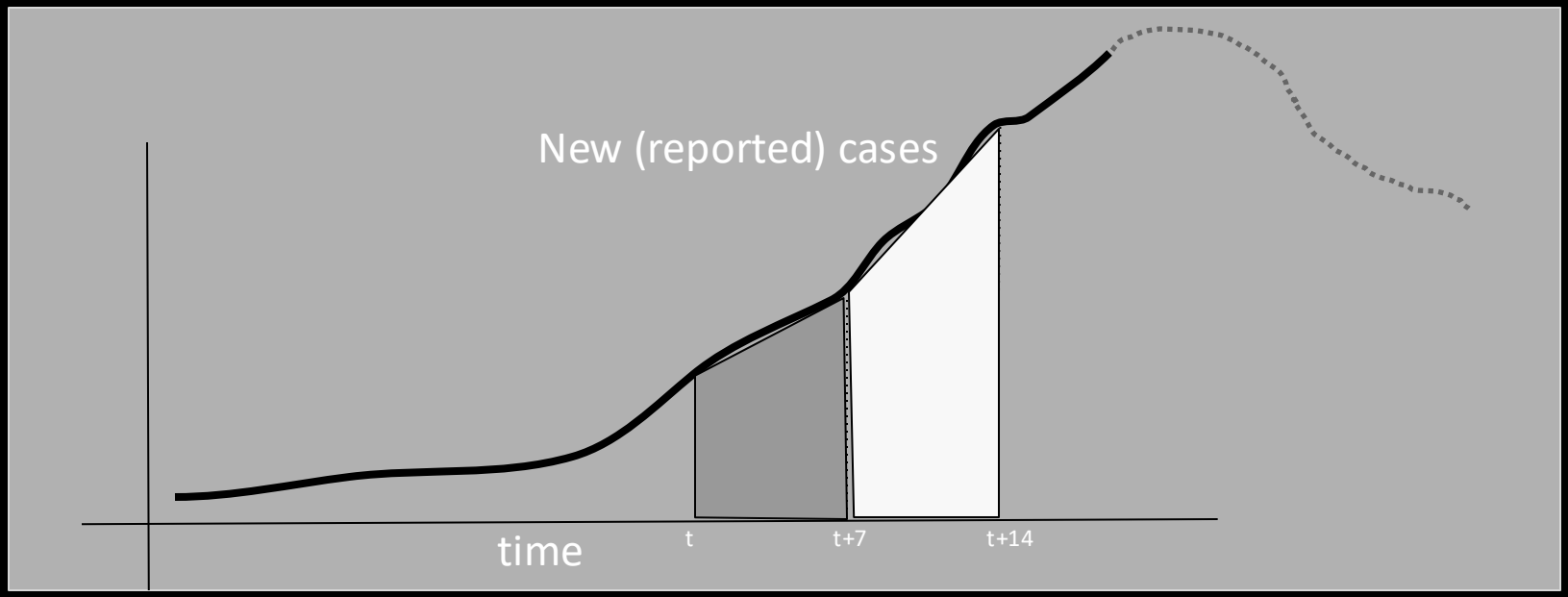
These observed relationships in influenza transmission have been assumed (without any evidence) for the ongoing COVID-19 outbreak. It has been stated that COVID-19 transmission will decrease as warmer temperatures (leading to higher absolute humidity conditions) are experienced in the upcoming spring months

From: J. Shaman, E. Goldstein, M. Lipsitch, Absolute Humidity and Pandemic Versus Epidemic Influenza, *American Journal of Epidemiology*, Volume 173, Issue 2, 15 January 2011, Pages 127–135,

Our findings :

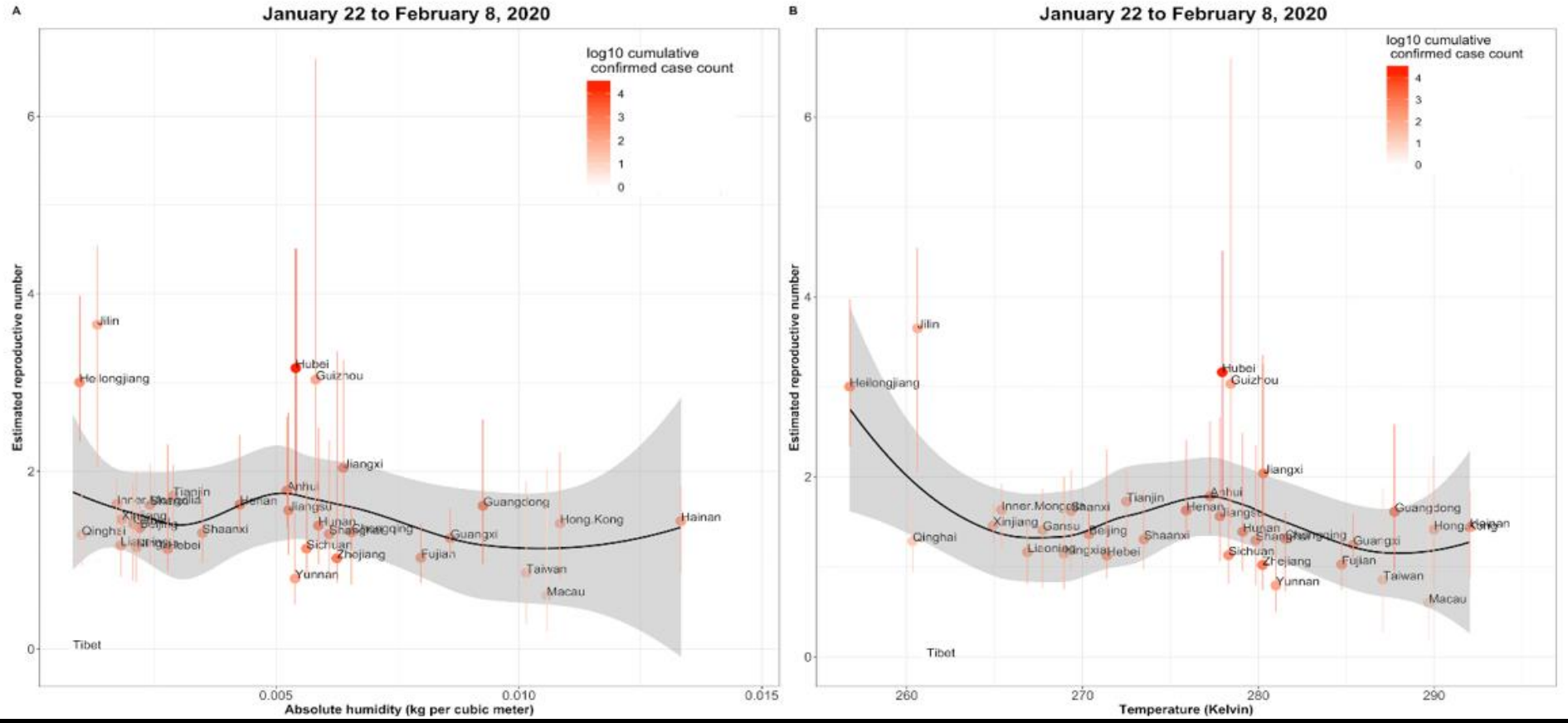
A novel coronavirus (SARS-CoV-2) was identified in Wuhan, Hubei Province, China, in December 2019 and has caused over 240,000 cases of COVID-19 worldwide as of March 19, 2020. Previous studies have supported an epidemiological hypothesis that cold and dry environments facilitate the survival and spread of droplet-mediated viral diseases, and warm and humid environments see attenuated viral transmission (e.g., influenza). However, the role of temperature and humidity in transmission of COVID-19 has not yet been established. Here, we examine the spatial variability of the basic reproductive numbers of COVID-19 across provinces and cities in China and show that environmental variables alone cannot explain this variability. Our findings suggest that changes in weather alone (i.e., increase of temperature and humidity as spring and summer months arrive in the Northern Hemisphere) will not necessarily lead to declines in case count without the implementation of extensive public health interventions.

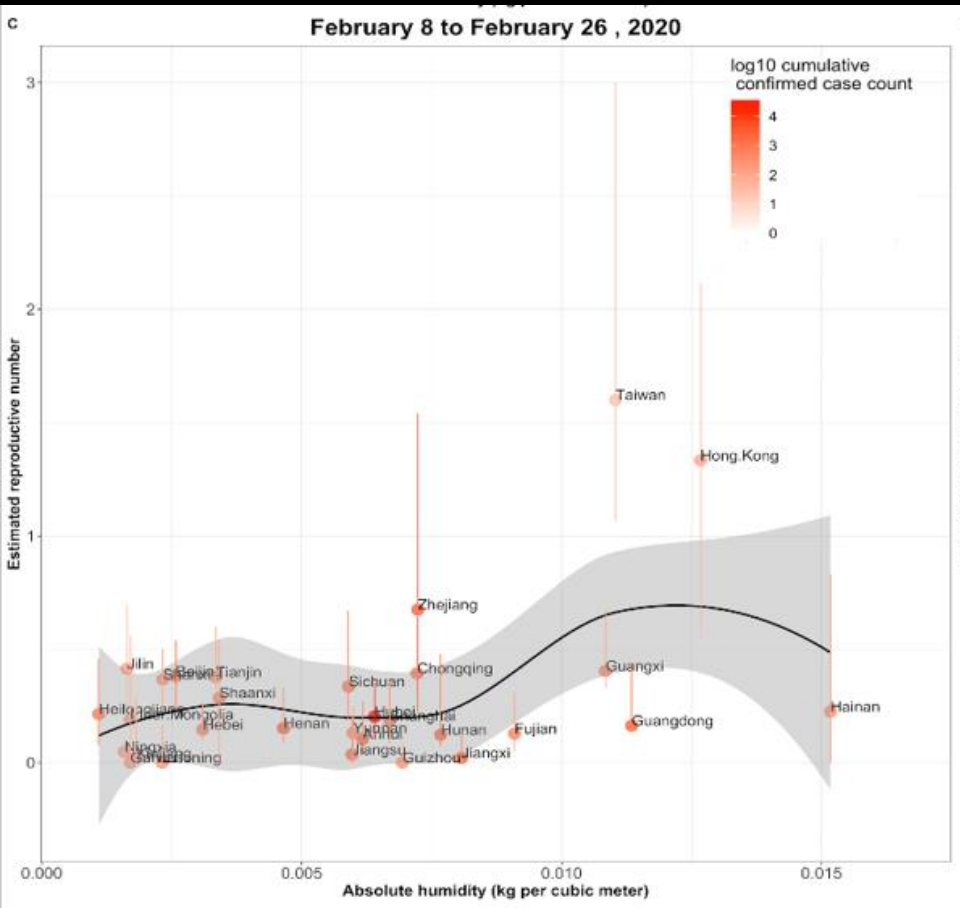
Calculating a proxy for R0

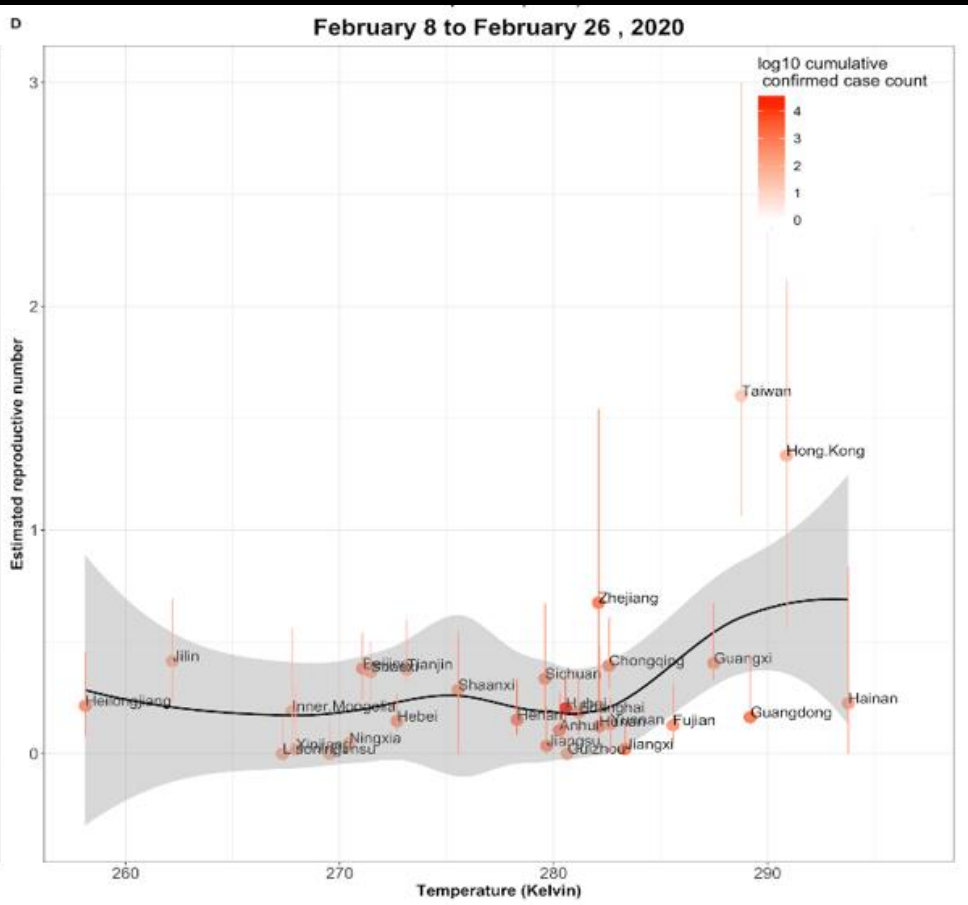


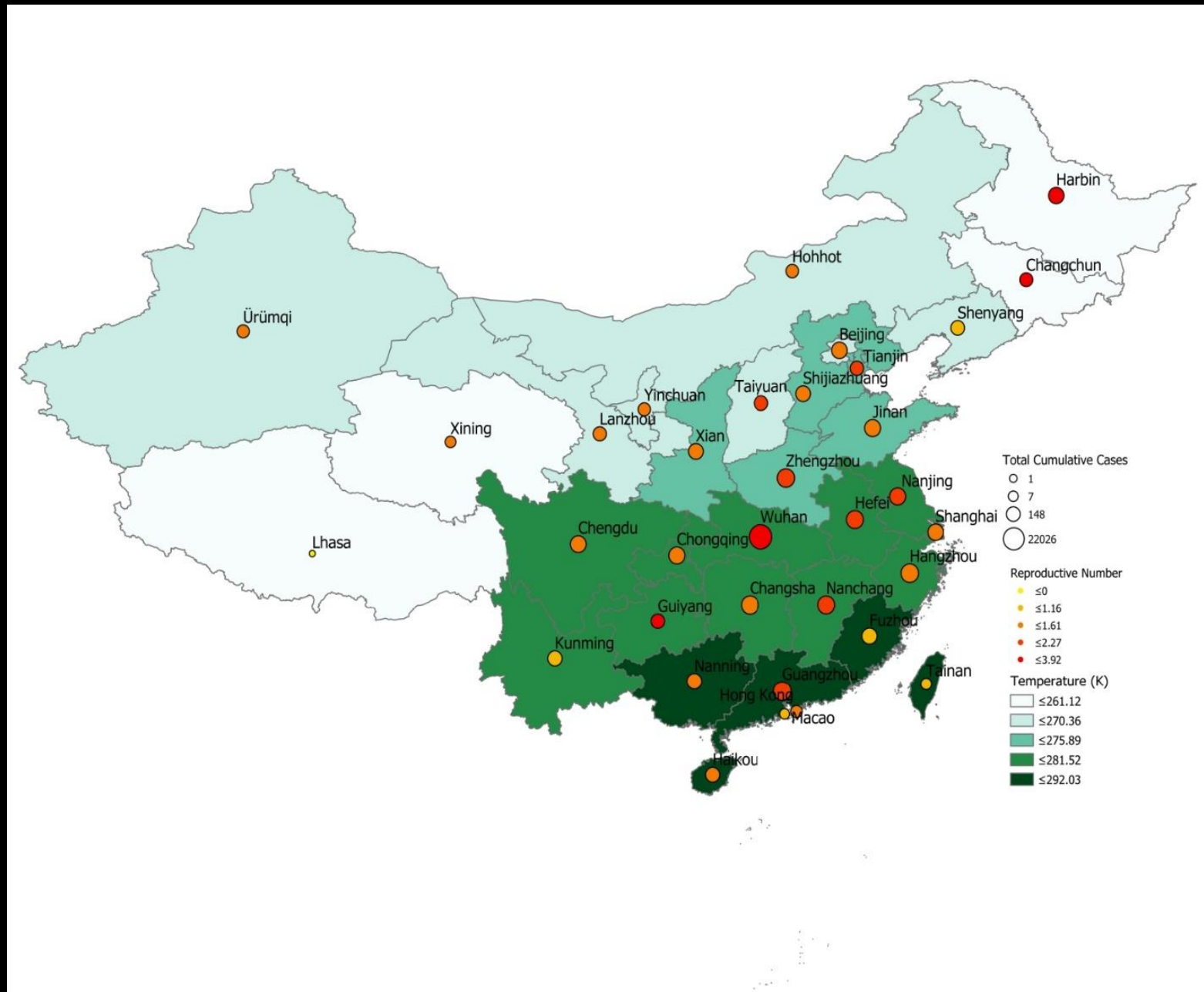
$$R_{proxy}(t,d) = \frac{C(t+2d) - C(t+d)}{C(t+d) - C(t)}$$

C(t) is the total cumulative cases at time t, and d = 5,6,7 is an estimate of the serial interval


A
January 22 to February 8, 2020
log10 cumulative confirmed case count
4
3
2
1
0
Estimated reproductive number
6
4
2
0
Jilin
Heilongjiang
Hubei
Guizhou
Jiangxi
Tianjin
Anhui
Henan
Jiangsu
Guangdong
Hong Kong
Hainan
Qinghai
Shaanxi
Sichuan
Zhejiang
Fujian
Guangxi
Yunnan
Taiwan
Macau
Tibet
0.005
0.010
0.015
Absolute humidity (kg per cubic meter)
B
January 22 to February 8, 2020
log10 cumulative confirmed case count
4
3
2
1
0
Estimated reproductive number
6
4
2
0
Jilin
Heilongjiang
Hubei
Guizhou
Jiangxi
Tianjin
Anhui
Henan
Guangdong
Hong Kong
Hainan
Qinghai
Xinjiang
Gansu
Shaanxi
Sichuan
Zhejiang
Fujian
Guangxi
Yunnan
Taiwan
Macau
Tibet
260
270
280
290
Temperature (Kelvin)


C
February 8 to February 26 , 2020
log10 cumulative confirmed case count
4
3
2
1
0
Estimated reproductive number
3
2
1
0
Taiwan
Hong.Kong
Zhejiang
Jilin
Tianjin
Shaanxi
Sichuan
Chongqing
Guangxi
Hainan
Hebei
Henan
Hunan
Fujian
Guangdong
Jiangsu
Guizhou
Jiangxi
0.000
0.005
0.010
0.015
Absolute humidity (kg per cubic meter)


D
February 8 to February 26 , 2020
log10 cumulative confirmed case count
4
3
2
1
0
Estimated reproductive number
3
2
1
0
Taiwan
Hong.Kong
Zhejiang
Jilin
Sichuan
Chongqing
Guangxi
Shaanxi
Hainan
Hebei
Fujian
Guangdong
Ningxia
Jiangxi
260
270
280
290
Temperature (Kelvin)


Harbin
Changchun
Shenyang
Hohhot
Ürümqi
Beijing
Tianjin
Shijiazhuang
Taiyuan
Yinchuan
Lanzhou
Xining
Jinan
Xian
Zhengzhou
Nanjing
Hefei
Shanghai
Wuhan
Chengdu
Chongqing
Lhasa
Hangzhou
Changsha
Nanchang
Guiyang
Fuzhou
Kunming
Nanning
Guangzhou
Tainan
Hong Kong
Macao
Haikou
Total Cumulative Cases
1
7
148
22026
Reproductive Number
≤0
≤1.16
≤1.61
≤2.27
≤3.92
Temperature (K)
≤261.12
≤270.36
≤275.89
≤281.52
≤292.03

Science Advances Contents News Careers Journals

An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time

Nicole E. Kogan[1,2,*,†], Leonardo Clemente[1,*,†], Parker Liautaud[3,*,†], Justin Kaashoek[1,4], Nicholas B. Link[1,5], …
+ See all authors and affiliations

Science Advances 05 Mar 2021:
Vol. 7, no. 10, eabd6989
DOI: 10.1126/sciadv.abd6989

Abstract

Given still-high levels of coronavirus disease 2019 (COVID-19) susceptibility and inconsistent transmission-containing strategies, outbreaks have continued to emerge across the United States. Until effective vaccines are widely deployed, curbing COVID-19 will require carefully timed nonpharmaceutical interventions (NPIs). A COVID-19 early warning system is vital for this. Here, we evaluate digital data streams as early indicators of state-level COVID-19 activity from 1 March to 30 September 2020. We observe that increases in digital data stream activity anticipate increases in confirmed cases and deaths by 2 to 3 weeks. Confirmed cases and deaths also decrease 2 to 4 weeks after NPI implementation, as measured by anonymized, phone-derived human mobility data. We propose a means of harmonizing these data streams to identify future COVID-19 outbreaks. Our results suggest that combining disparate health and behavioral data may help identify disease activity changes weeks before observation using traditional epidemiological monitoring.

Nicole Kogan

Leonardo Clemente

Parker Liautaud

Justin Kaashoek

Nick Link

Andre Nguyen

Fred Lu

Peter Huybers

Bernd Resch

Clemens Havas

Andy Petutschnig

Jessica Davis

Matteo Chinazzi

Backtosch Mustafa

Bill Hanage

Alex Vespignani

Mauricio Santillana

RESEARCH ARTICLE | CORONAVIRUS

An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time

Nicole E. Kogan[1,2,*,†], Leonardo Clemente[1,*,†], Parker Liautaud[3,*,†], Justin Kaashoek[1,4], Nicholas B. Link[1,5], ...
+ See all authors and affiliations

Science Advances 05 Mar 2021:
Vol. 7, no. 10, eabd6989
DOI: 10.1126/sciadv.abd6989

Abstract

Given still-high levels of coronavirus disease 2019 (COVID-19) susceptibility and inconsistent transmission-containing strategies, outbreaks have continued to emerge across the United States. Until effective vaccines are widely deployed, curbing COVID-19 will require carefully timed nonpharmaceutical interventions (NPIs). A COVID-19 early warning system is vital for this. Here, we evaluate digital data streams as early indicators of state-level COVID-19 activity from 1 March to 30 September 2020. We observe that increases in digital data stream activity anticipate increases in confirmed cases and deaths by 2 to 3 weeks. Confirmed cases and deaths also decrease 2 to 4 weeks after NPI implementation, as measured by anonymized, phone-derived human mobility data. We propose a means of harmonizing these data streams to identify future COVID-19 outbreaks. Our results suggest that combining disparate health and behavioral data may help identify disease activity changes weeks before observation using traditional epidemiological monitoring.

Data Sources

We were interested in comparing the performance of digital data sources (COVID-19 "proxies") to the performance of traditional COVID-19 measures ("gold standards") in forecasting sharp changes in epidemic activity

COVID-19 proxies

Google Trends

kinsa

cuebiq

GLEAM
GLOBAL EPIDEMIC AND MOBILITY MODEL

COVID-19 gold standards

The COVID Tracking Project

JOHNS HOPKINS UNIVERSITY & MEDICINE | CORONAVIRUS RESOURCE CENTER

CDC *

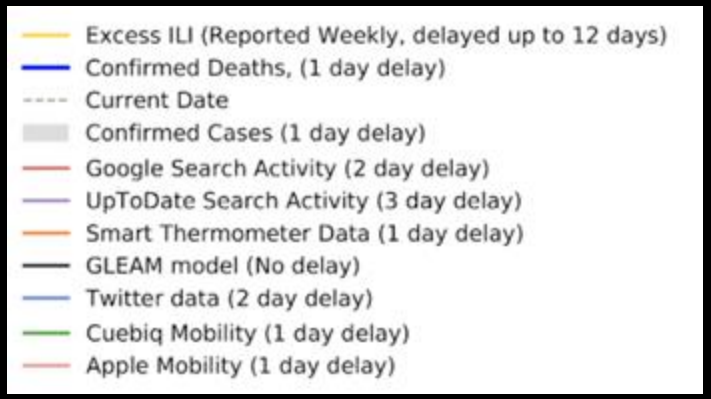
* ILINet

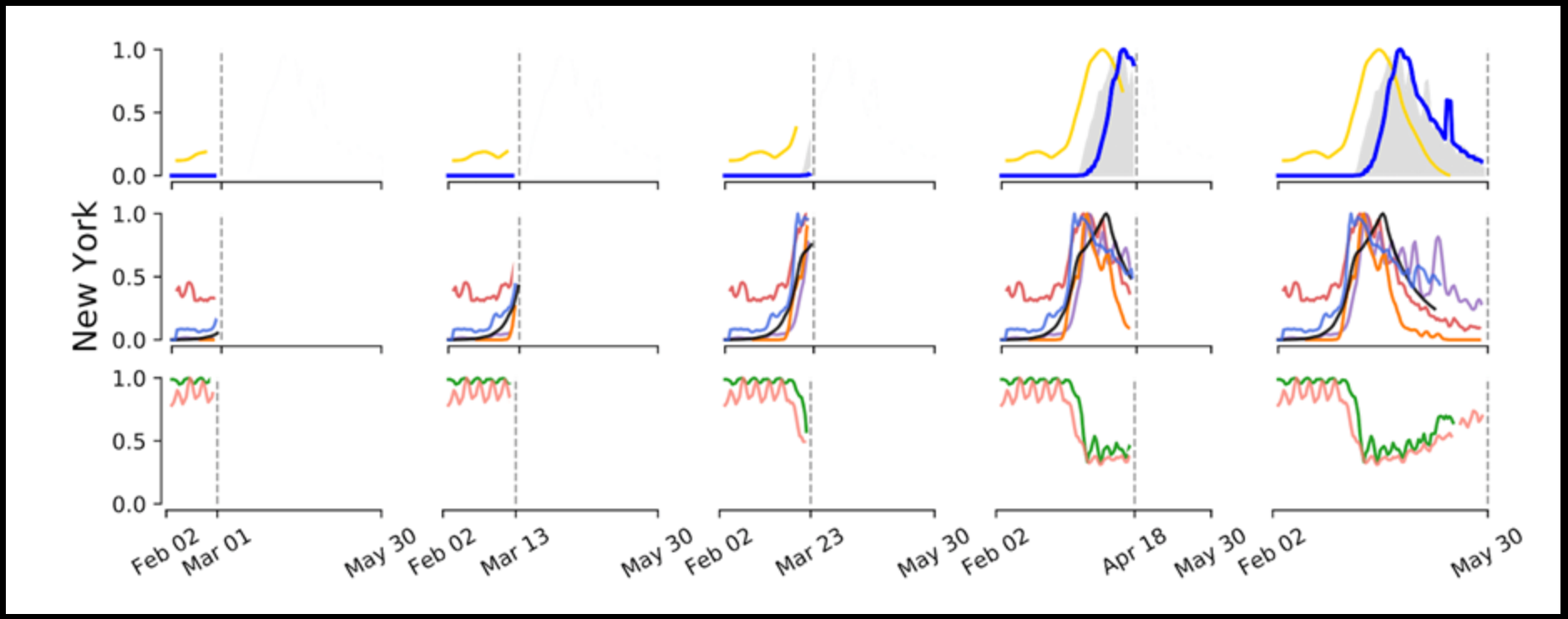
We started working on exercise 3 here

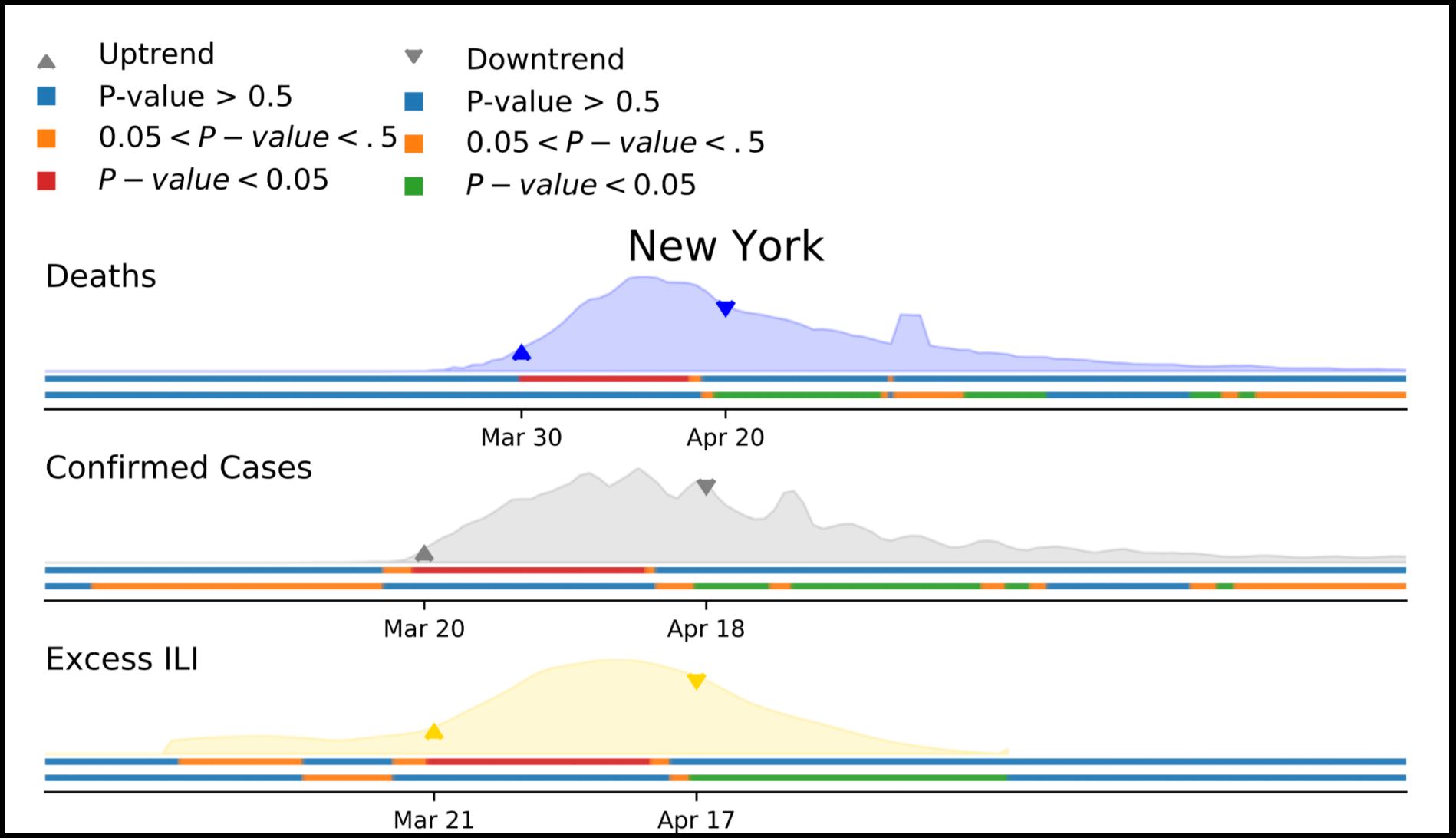
- Plotting the data sources
- Compound interest and exponential growth
- Anomaly detection approaches

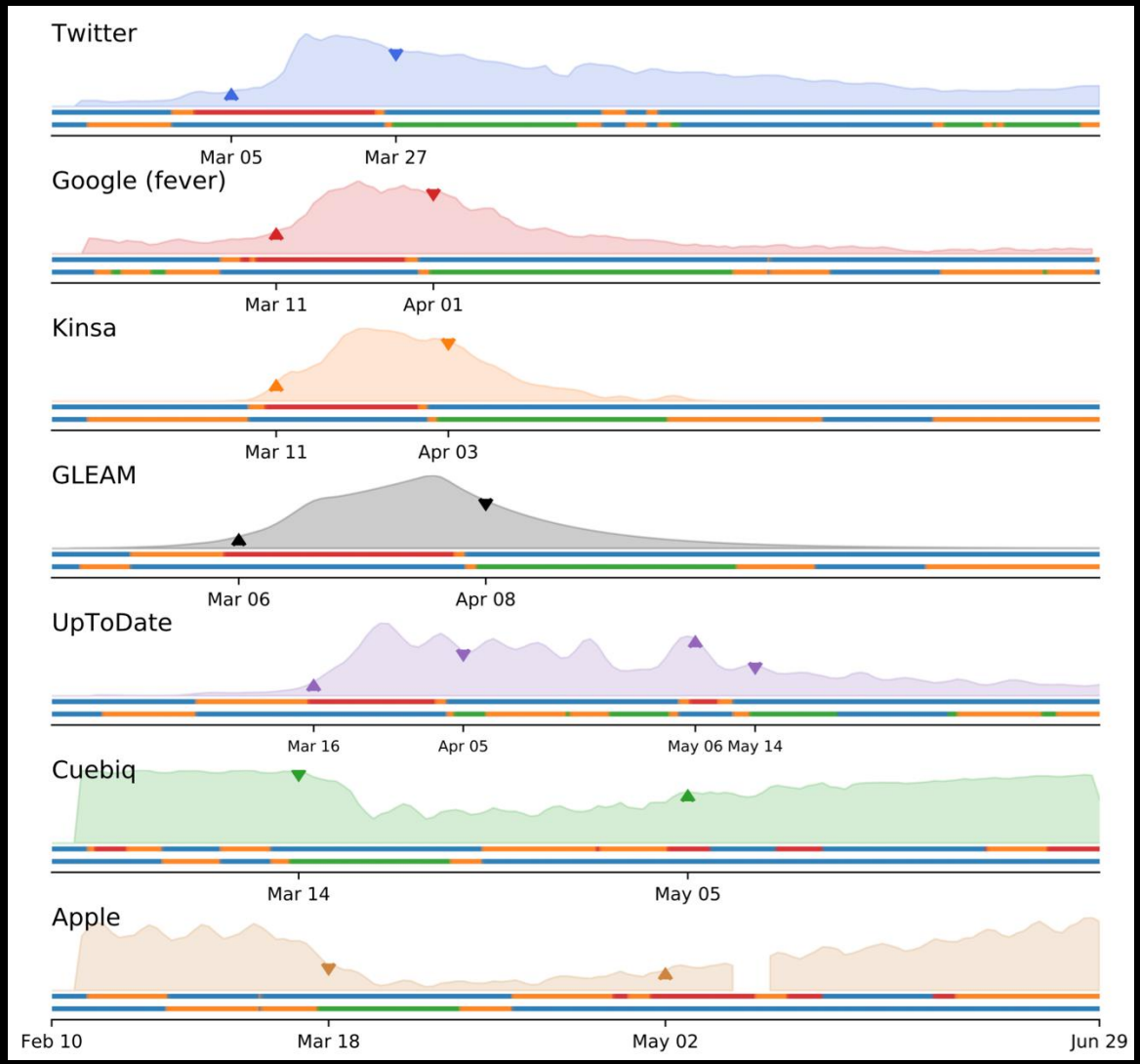
Fig. I - Time series for COVID-19 proxies and gold standards

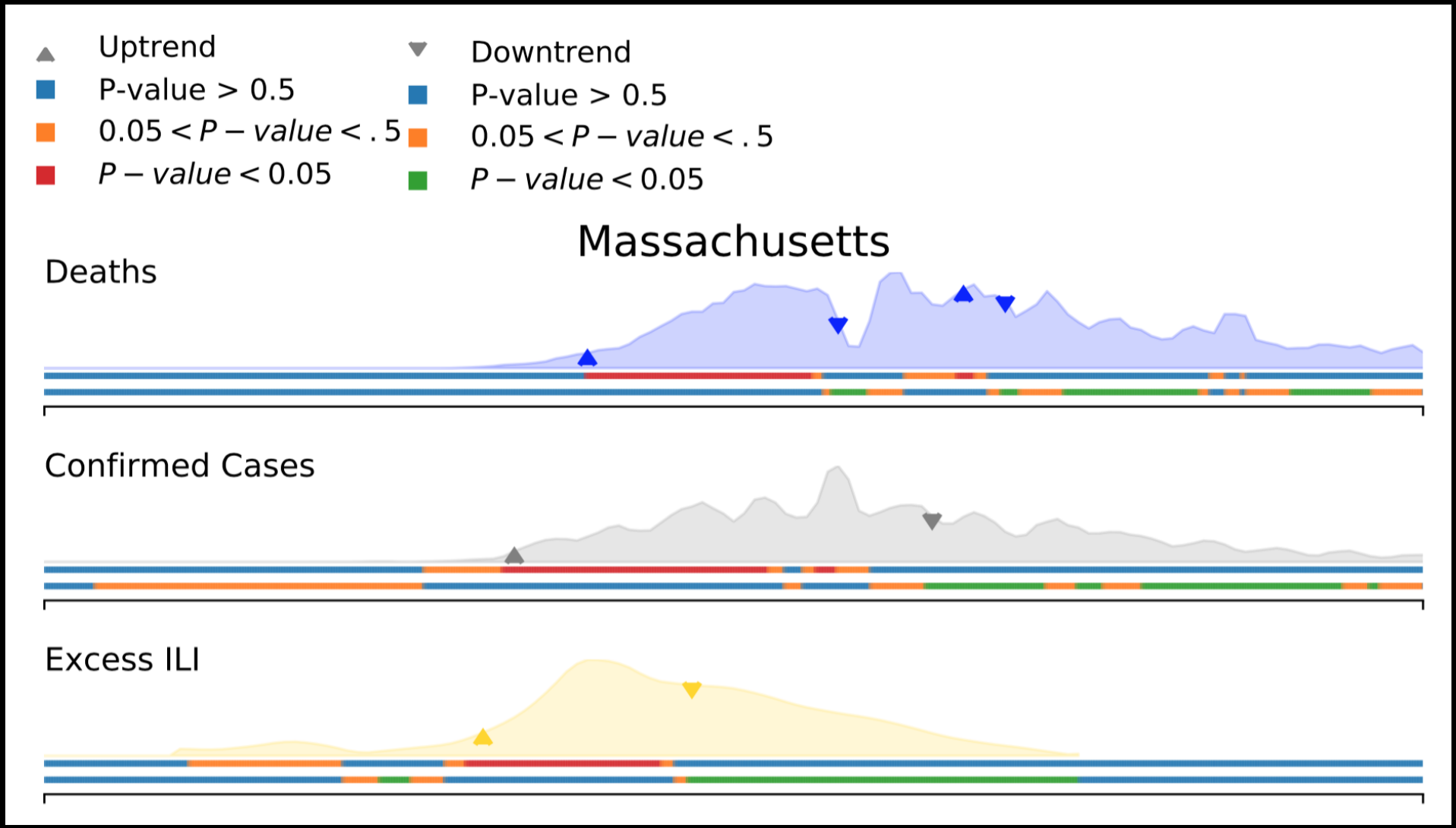
Legend (*delays represent lags in data availability*):






Uptrend
P-value > 0.5
0.05 < P − value < .5
P − value < 0.05
Downtrend
P-value > 0.5
0.05 < P − value < .5
P − value < 0.05
New York
Deaths
Mar 30
Apr 20
Confirmed Cases
Mar 20
Apr 18
Excess ILI
Mar 21
Apr 17


Twitter
Mar 05
Mar 27
Google (fever)
Mar 11
Apr 01
Kinsa
Mar 11
Apr 03
GLEAM
Mar 06
Apr 08
UpToDate
Mar 16
Apr 05
May 06
May 14
Cuebiq
Mar 14
May 05
Apple
Feb 10
Mar 18
May 02
Jun 29


Uptrend
P-value > 0.5
0.05 < P − value < .5
P − value < 0.05
Downtrend
P-value > 0.5
0.05 < P − value < .5
P − value < 0.05
Massachusetts
Deaths
Confirmed Cases
Excess ILI

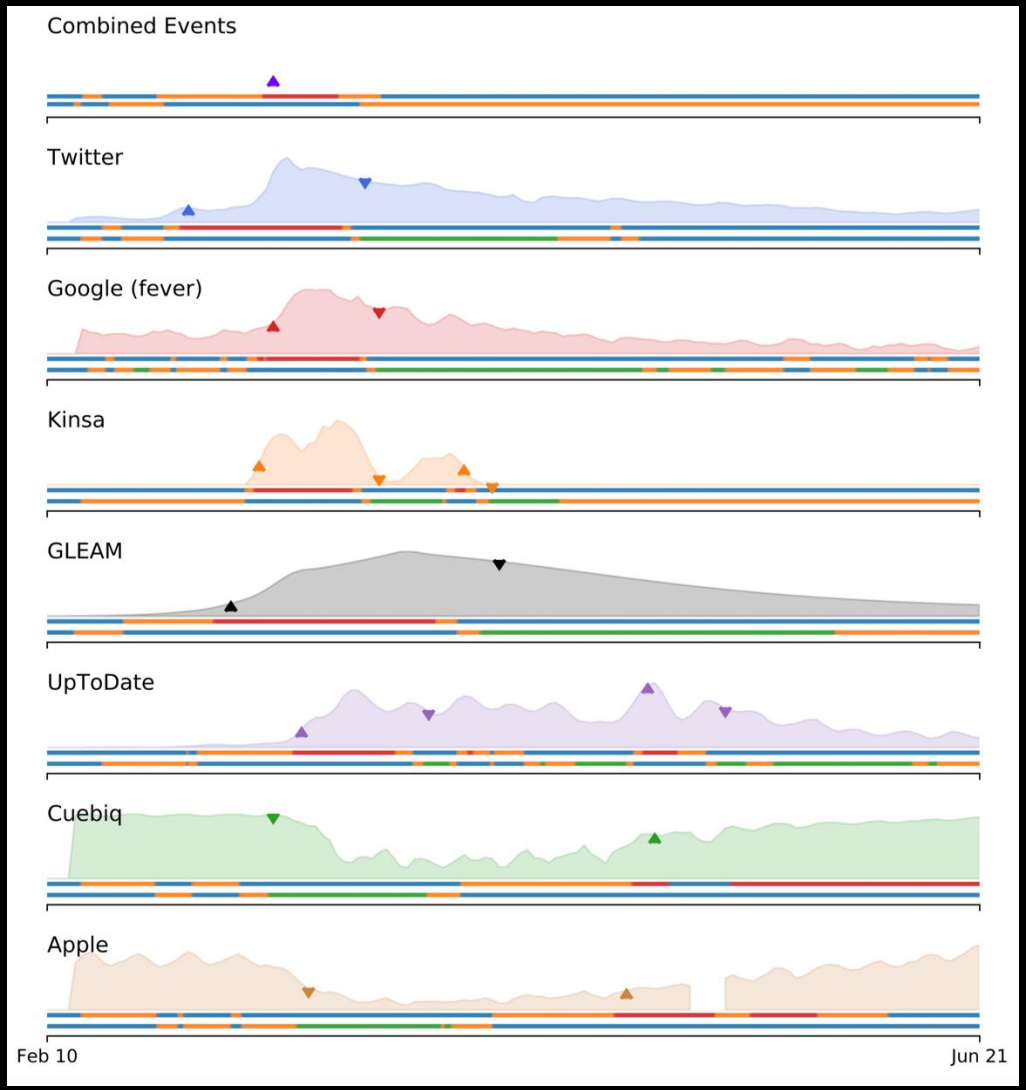

Combined Events
Twitter
Google (fever)
Kinsa
GLEAM
UpToDate
Cuebiq
Apple
Feb 10
Jun 21

Fig. III - Uptrends and downtrends are detected earliest for Twitter and Cuebiq, respectively, across the US

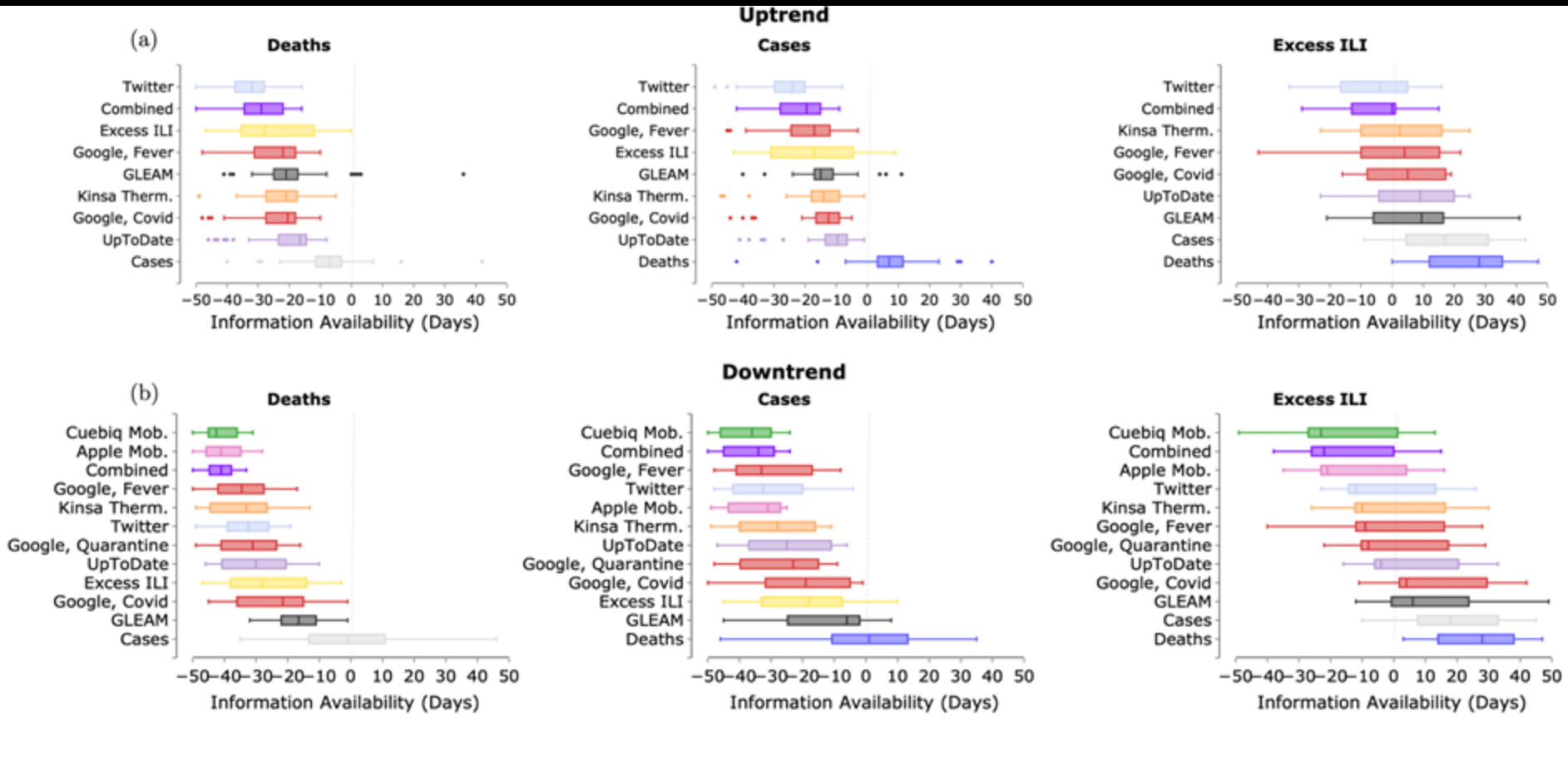
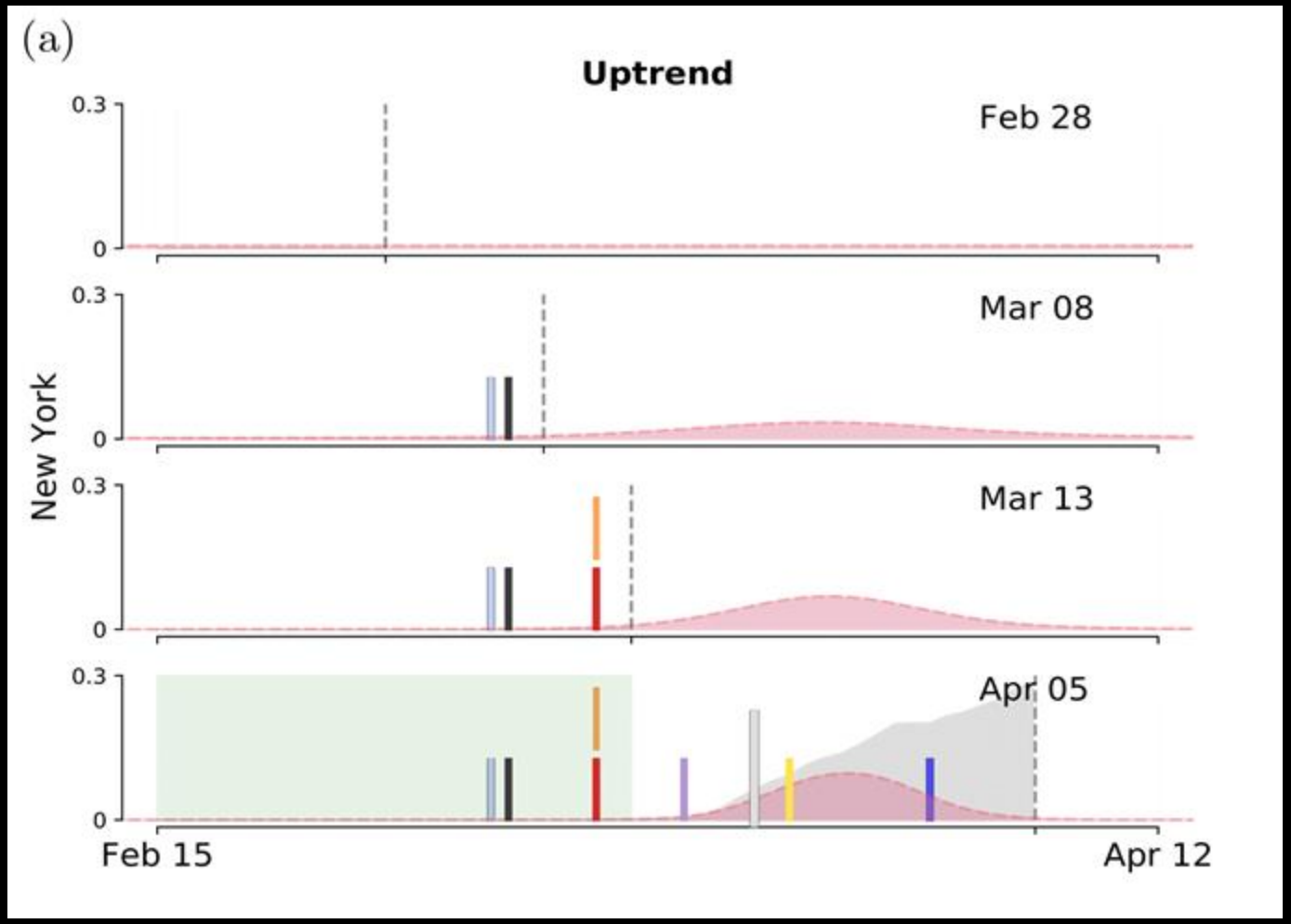


Fig. IV - Evolving posterior probability distribution for time-to-event estimation in New York



Legend:

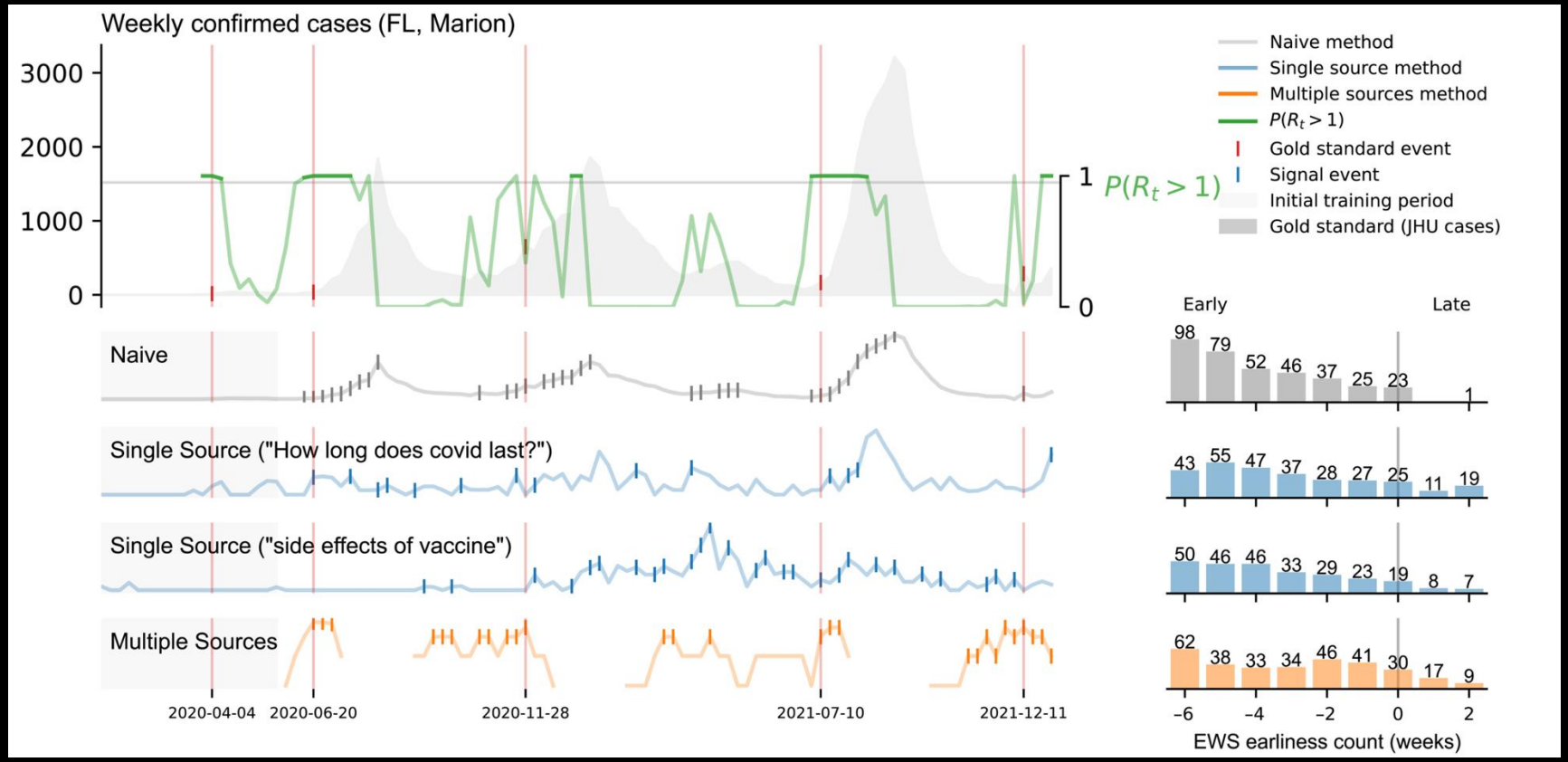
Confirmed Cases
Deaths
Google Trends (Fever)
UpToDate
Kinsa Thermometers
GLEAM
Excess ILI
Twitter
Cuebiq
Apple

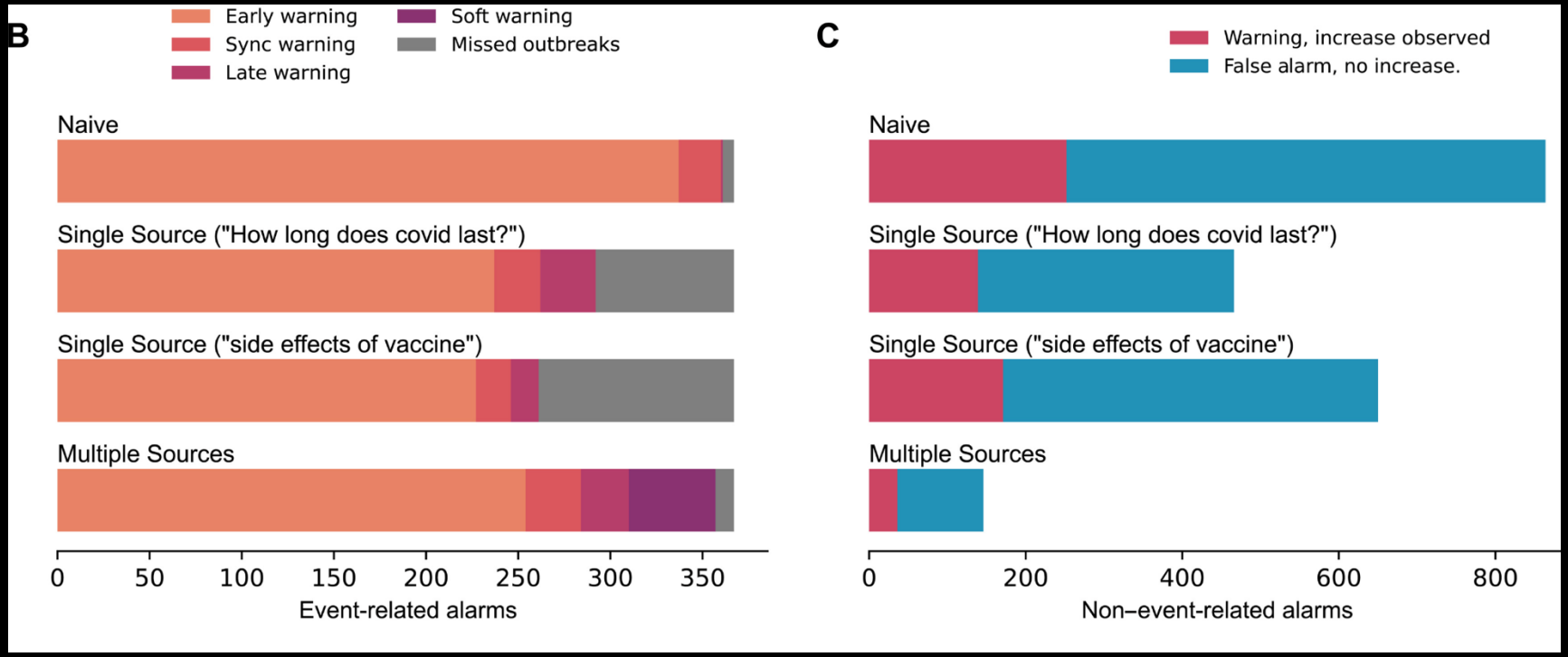
RESEARCH ARTICLE | CORONAVIRUS

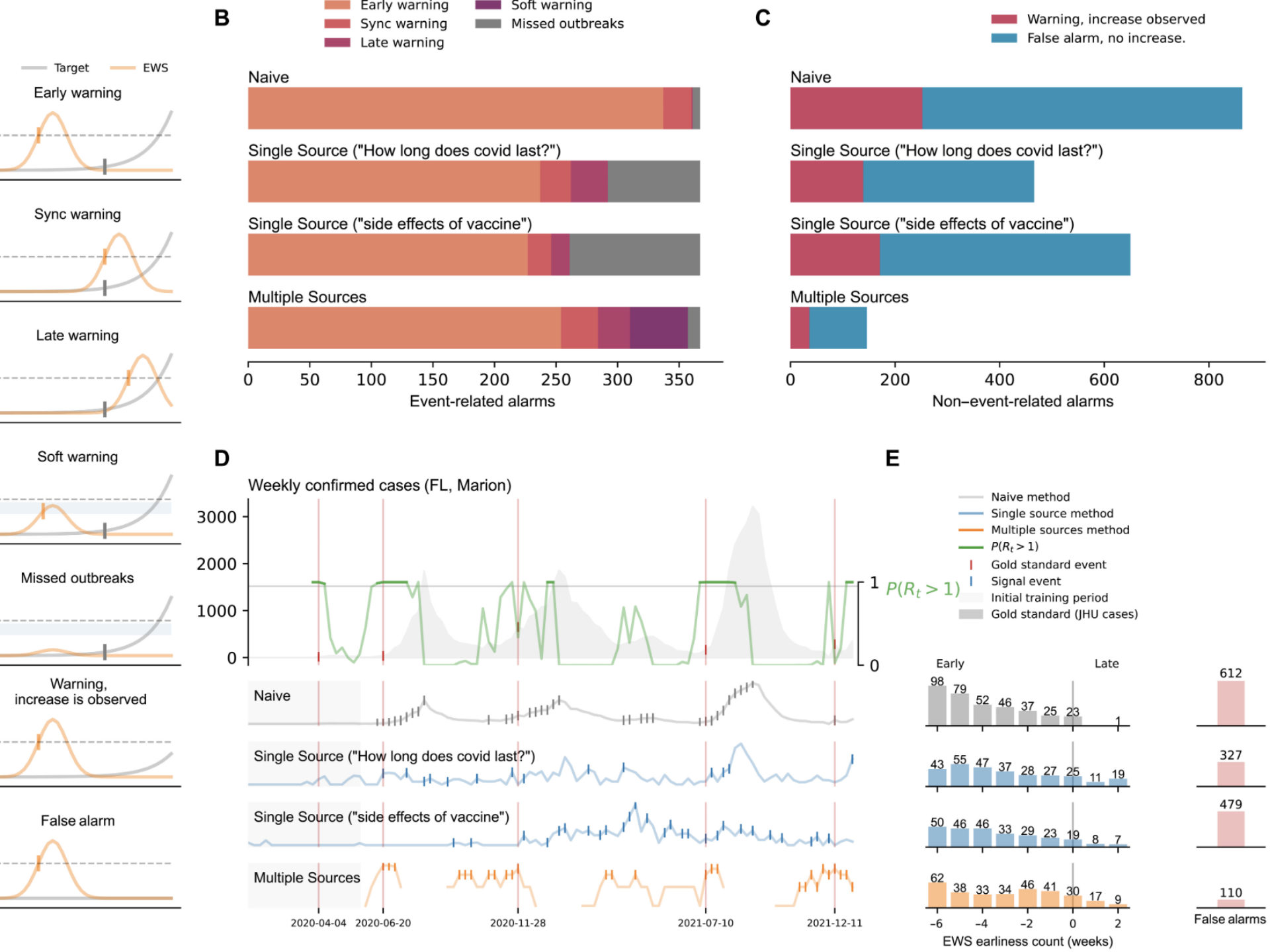
Using digital traces to build prospective and real-time county-level early warning systems to anticipate COVID-19 outbreaks in the United States

LUCAS M. STOLERMAN, LEONARDO CLEMENTE, CANELLE POIRIER, KRIS V. PARAG, ATREYEE MAJUMDER, SERGE MASYN, BERND RESCH, AND MAURICIO SANTILLANA Authors Info & Affiliations

SCIENCE ADVANCES • 18 Jan 2023 • Vol 9, Issue 3 • DOI: 10.1126/sciadv.abq0199


Weekly confirmed cases (FL, Marion)
3000
2000
1000
0
1
0
$P(R_t > 1)$
Naive method
Single source method
Multiple sources method
$P(R_t > 1)$
Gold standard event
Signal event
Initial training period
Gold standard (JHU cases)
Naive
Single Source ("How long does covid last?")
Single Source ("side effects of vaccine")
Multiple Sources
2020-04-04
2020-06-20
2020-11-28
2021-07-10
2021-12-11
Early
Late
98 79 52 46 37 25 23 1
43 55 47 37 28 27 25 11 19
50 46 46 33 29 23 19 8 7
62 38 33 34 46 41 30 17 9
−6 −4 −2 0 2
EWS earliness count (weeks)


B
Early warning
Sync warning
Late warning
Soft warning
Missed outbreaks
Naive
Single Source ("How long does covid last?")
Single Source ("side effects of vaccine")
Multiple Sources
0
50
100
150
200
250
300
350
Event-related alarms
C
Warning, increase observed
False alarm, no increase.
Naive
Single Source ("How long does covid last?")
Single Source ("side effects of vaccine")
Multiple Sources
0
200
400
600
800
Non–event-related alarms


Target
EWS
Early warning
Sync warning
Late warning
Soft warning
Missed outbreaks
Warning,
increase is observed
False alarm
B
Early warning
Sync warning
Late warning
Soft warning
Missed outbreaks
Naive
Single Source ("How long does covid last?")
Single Source ("side effects of vaccine")
Multiple Sources
0
50
100
150
200
250
300
350
Event-related alarms
C
Warning, increase observed
False alarm, no increase.
Naive
Single Source ("How long does covid last?")
Single Source ("side effects of vaccine")
Multiple Sources
0
200
400
600
800
Non–event-related alarms
D
Weekly confirmed cases (FL, Marion)
3000
2000
1000
0
1
0
$P(R_t > 1)$
Naive
Single Source ("How long does covid last?")
Single Source ("side effects of vaccine")
Multiple Sources
2020-04-04
2020-06-20
2020-11-28
2021-07-10
2021-12-11
E
Naive method
Single source method
Multiple sources method
$P(R_t > 1)$
Gold standard event
Signal event
Initial training period
Gold standard (JHU cases)
Early
Late
98 79 52 46 37 25 23 1
43 55 47 37 28 27 25 11 19
50 46 46 33 29 23 19 8 7
62 38 33 34 46 41 30 17 9
612
327
479
110
False alarms
−6
−4
−2
0
2
EWS earliness count (weeks)

Lecture 4 ended here

Switch to presentation of the EWS
Santillana_EWS_InsightNet_May_7_2025

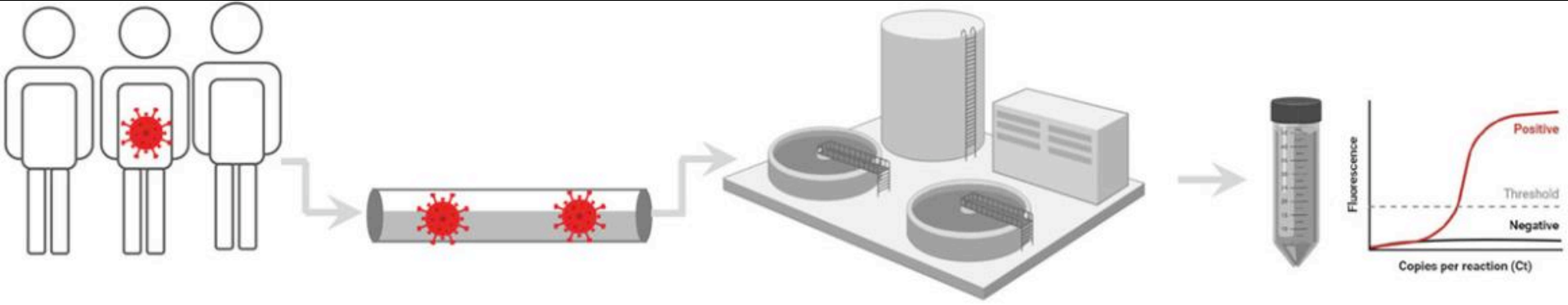
Wastewater monitoring

Science of The Total Environment

Volume 805, 20 January 2022, 150121

SARS-CoV-2 RNA concentrations in wastewater foreshadow dynamics and clinical presentation of new COVID-19 cases

Fuqing Wu [a, b, 1], Amy Xiao [a, b, 1], Jianbo Zhang [a, b, 1], Katya Moniz [a, b], Noriko Endo [c], Federica Armas [d, e], Richard Bonneau [f], Megan A. Brown [f], Mary Bushman [g], Peter R. Chai [h, i], Claire Duvallet [c], Timothy B. Erickson [h, j], Katelyn Foppe [c], Newsha Ghaeli [c], Xiaoqiong Gu [d, e], William P. Hanage [g], Katherine H. Huang [k], Wei Lin Lee [d, e], Mariana Matus [c], Kyle A. McElroy [c], Jonathan Nagler [f], Steven F. Rhode [l], Mauricio Santillana [g, m, n], Joshua A. Tucker [f], Stefan Wuertz [e, o, p], Shijie Zhao [a, b], Janelle Thompson [e, o, q], Eric J. Alm [a, b, d, e, k]



New infection cases $I(t)$

Mean viral shedding $S(t)$

Viral copies in wastewater

$W(t)$

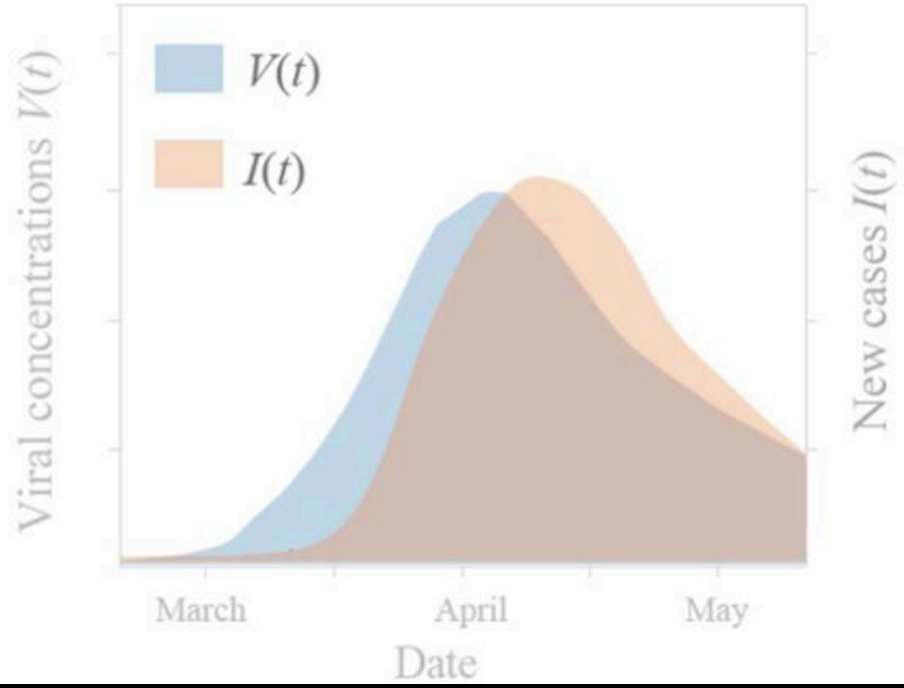
Viral concentrations

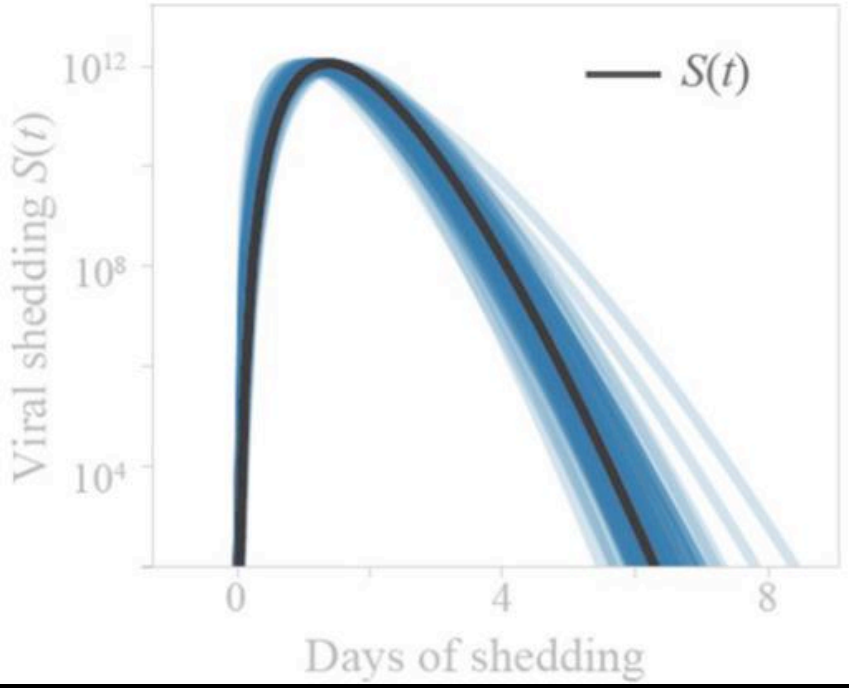
$V(t)$

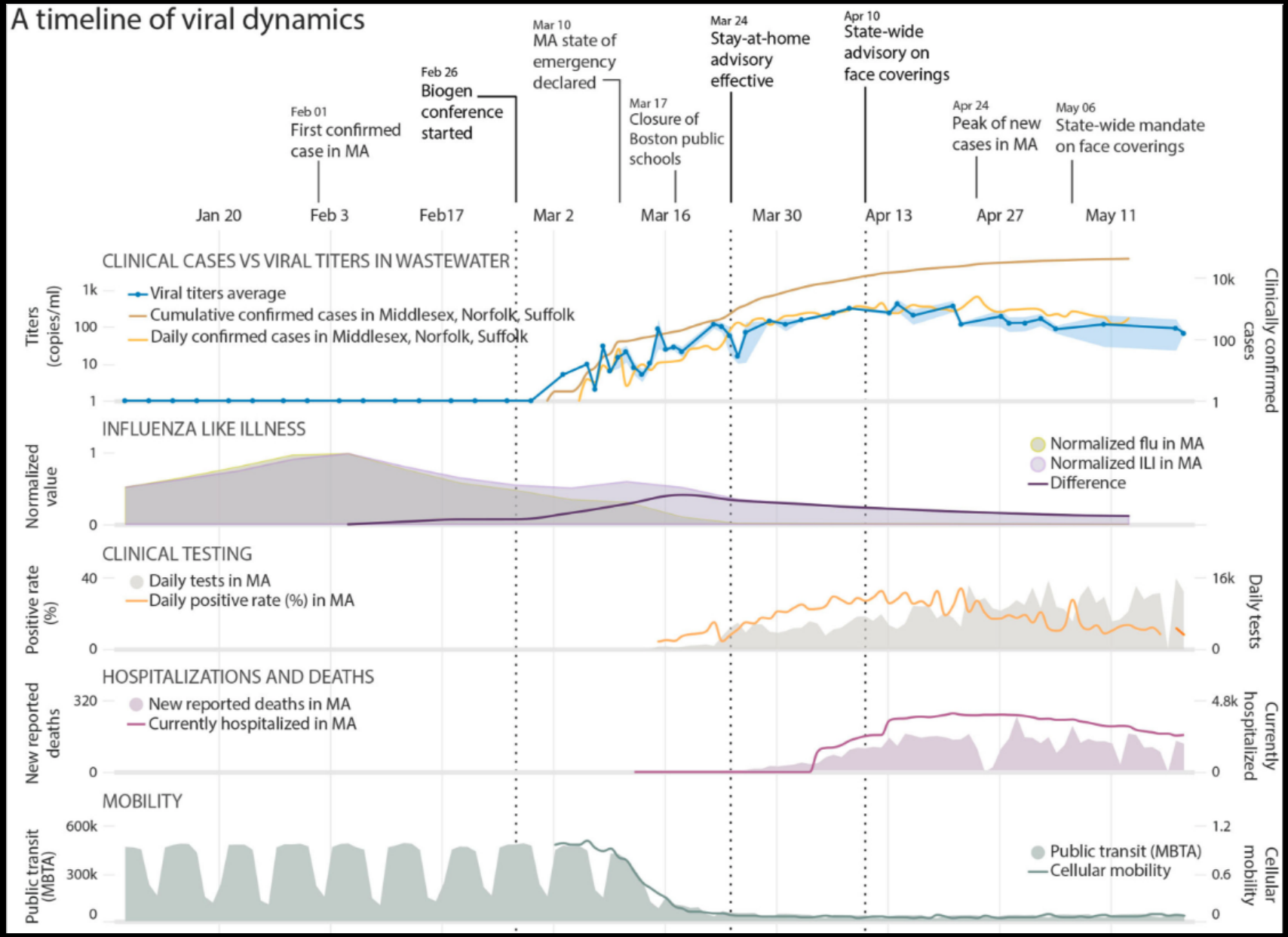
$$V(t) \propto I(t)$$

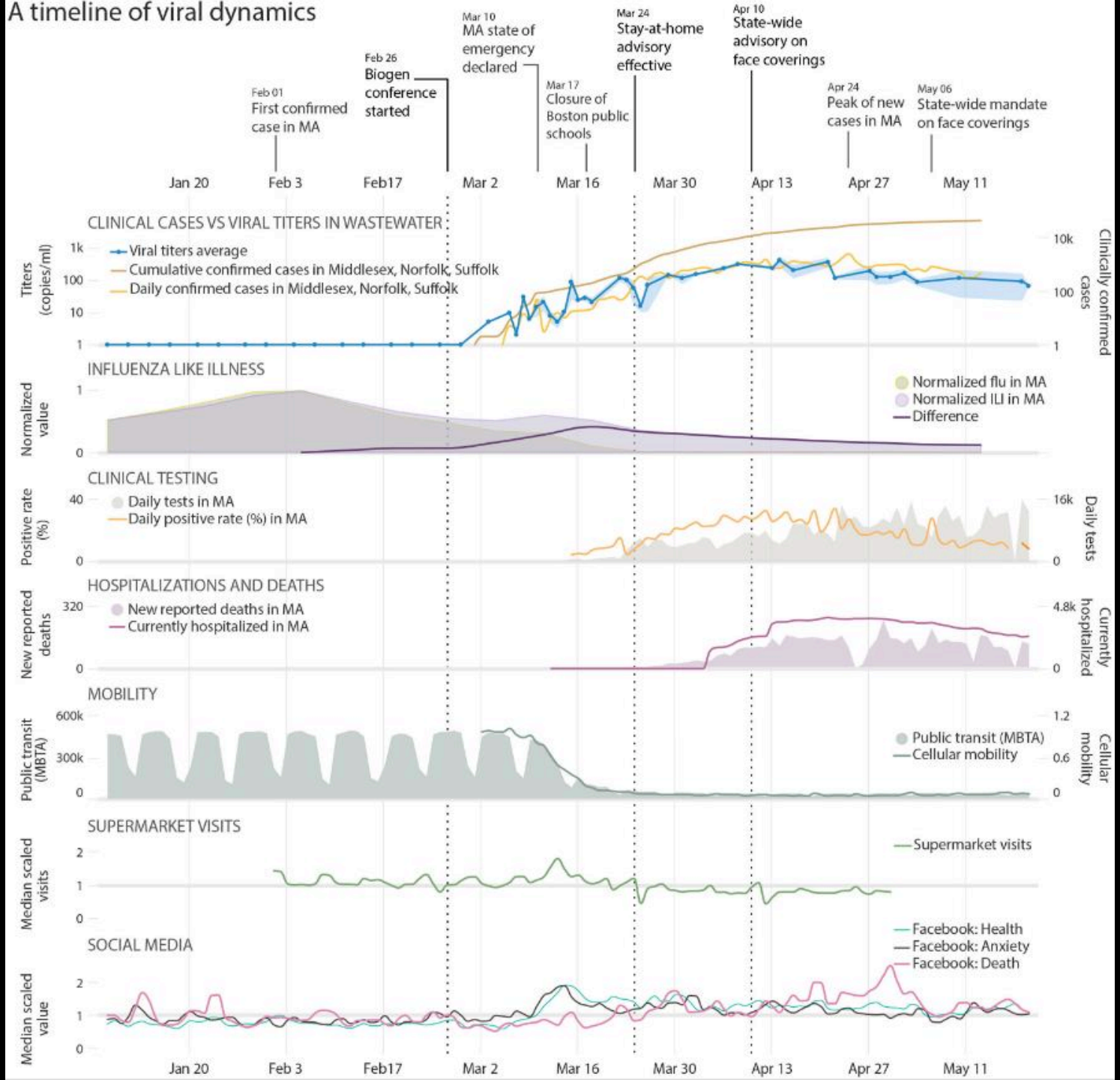
$$S(t) = Beta(t,\alpha,\beta)$$

$$W(t) = S(t) * I(t)$$






A timeline of viral dynamics
Feb 01
First confirmed case in MA
Feb 26
Biogen conference started
Mar 10
MA state of emergency declared
Mar 17
Closure of Boston public schools
Mar 24
Stay-at-home advisory effective
Apr 10
State-wide advisory on face coverings
Apr 24
Peak of new cases in MA
May 06
State-wide mandate on face coverings
Jan 20
Feb 3
Feb17
Mar 2
Mar 16
Mar 30
Apr 13
Apr 27
May 11
CLINICAL CASES VS VIRAL TITERS IN WASTEWATER
Titers (copies/ml)
1k
100
10
1
Viral titers average
Cumulative confirmed cases in Middlesex, Norfolk, Suffolk
Daily confirmed cases in Middlesex, Norfolk, Suffolk
10k
100
1
Clinically confirmed cases
INFLUENZA LIKE ILLNESS
Normalized value
1
0
Normalized flu in MA
Normalized ILI in MA
Difference
CLINICAL TESTING
Positive rate (%)
40
0
Daily tests in MA
Daily positive rate (%) in MA
16k
0
Daily tests
HOSPITALIZATIONS AND DEATHS
New reported deaths
320
0
New reported deaths in MA
Currently hospitalized in MA
4.8k
0
Currently hospitalized
MOBILITY
Public transit (MBTA)
600k
300k
0
Public transit (MBTA)
Cellular mobility
1.2
0.6
0
Cellular mobility


A timeline of viral dynamics
Feb 01
First confirmed case in MA
Feb 26
Biogen conference started
Mar 10
MA state of emergency declared
Mar 17
Closure of Boston public schools
Mar 24
Stay-at-home advisory effective
Apr 10
State-wide advisory on face coverings
Apr 24
Peak of new cases in MA
May 06
State-wide mandate on face coverings
Jan 20
Feb 3
Feb17
Mar 2
Mar 16
Mar 30
Apr 13
Apr 27
May 11
CLINICAL CASES VS VIRAL TITERS IN WASTEWATER
Titers (copies/ml)
1k
100
10
1
Viral titers average
Cumulative confirmed cases in Middlesex, Norfolk, Suffolk
Daily confirmed cases in Middlesex, Norfolk, Suffolk
10k
100
1
Clinically confirmed cases
INFLUENZA LIKE ILLNESS
Normalized value
1
0
Normalized flu in MA
Normalized ILI in MA
Difference
CLINICAL TESTING
Positive rate (%)
40
0
Daily tests in MA
Daily positive rate (%) in MA
16k
0
Daily tests
HOSPITALIZATIONS AND DEATHS
New reported deaths
320
0
New reported deaths in MA
Currently hospitalized in MA
4.8k
0
Currently hospitalized
MOBILITY
Public transit (MBTA)
600k
300k
0
Public transit (MBTA)
Cellular mobility
1.2
0.6
0
Cellular mobility
SUPERMARKET VISITS
Median scaled visits
2
1
0
Supermarket visits
SOCIAL MEDIA
Median scaled value
2
1
0
Facebook: Health
Facebook: Anxiety
Facebook: Death
Jan 20
Feb 3
Feb17
Mar 2
Mar 16
Mar 30
Apr 13
Apr 27
May 11

THE COVIDSTATES PROJECT

THE STATE OF THE NATION: A 50-STATE COVID-19 SURVEY

As covered by:

THE WALL STREET JOURNAL. The New York Times The Atlantic
The Washington Post USA TODAY CBS NEWS
nature npr STAT The Boston Globe

Featured publications

JANUARY 24, 2024
Black Networks Matter The Role of Interracial Contact and Social Media in the 2020 Black Lives Matter Protests

JANUARY 11, 2024
Divisive or Descriptive?: How Americans Understand Critical Race Theory

Recent media coverage

Finding the Israel-Palestine sweet spot
Politico February 22, 2024

netsi | Northeastern University Network Science Institute
HARVARD Kennedy School SHORENSTEIN CENTER on Media, Politics and Public Policy
HARVARD MEDICAL SCHOOL
RUTGERS UNIVERSITY | NEW BRUNSWICK
Northwestern University

Machine Intelligence Group for the betterment of Health and the Environment

EPISTORM

Northeastern University

THE COVIDSTATES PROJECT

David Lazer
Northeastern University
PRINCIPAL INVESTIGATOR
Website

Matthew Baum
Harvard University
PRINCIPAL INVESTIGATOR
Website

Katherine Ognyanova
Rutgers University
PRINCIPAL INVESTIGATOR
Website

Roy H. Perlis
Harvard Medical School
PRINCIPAL INVESTIGATOR
Website

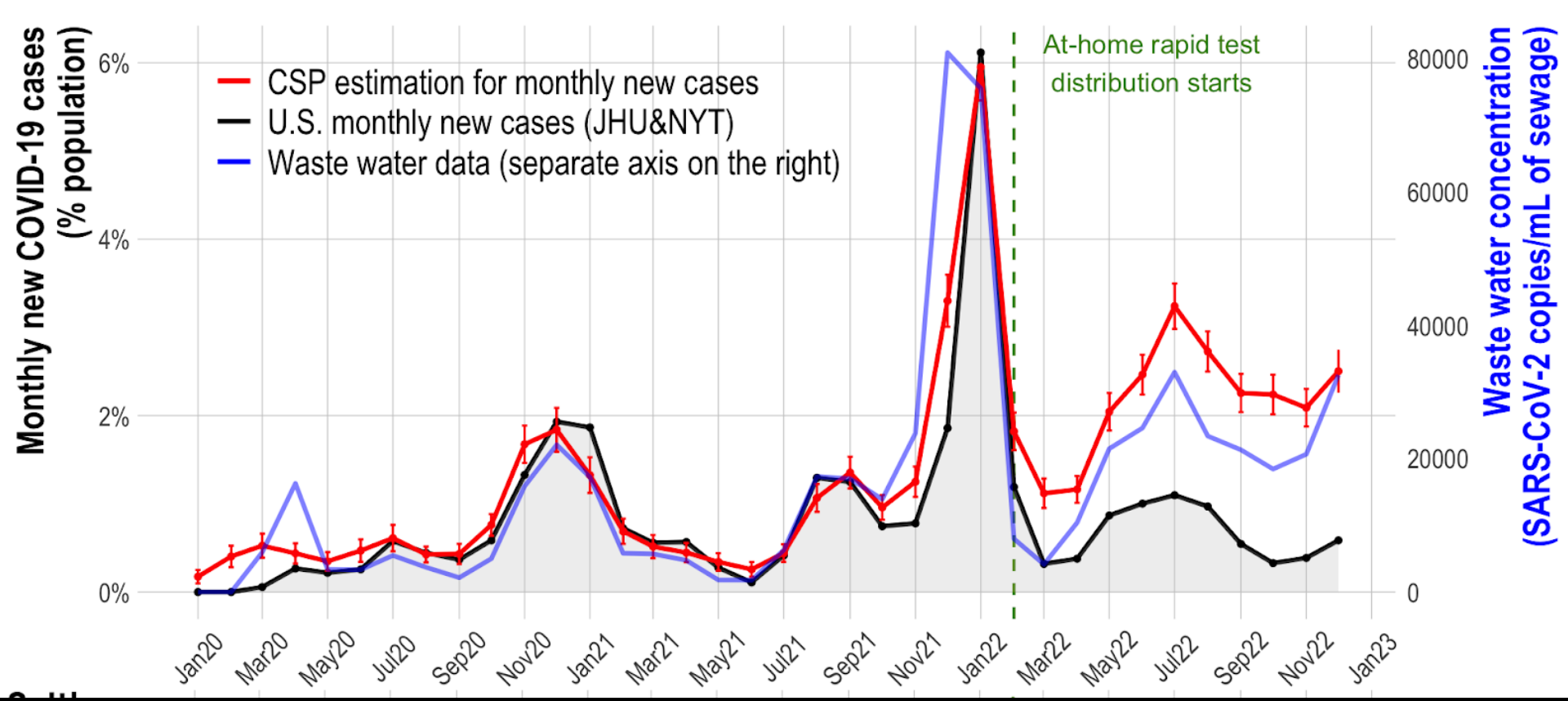
Mauricio Santillana
Northeastern University
PRINCIPAL INVESTIGATOR
Website

James Druckman
University of Rochester
PRINCIPAL INVESTIGATOR
Website

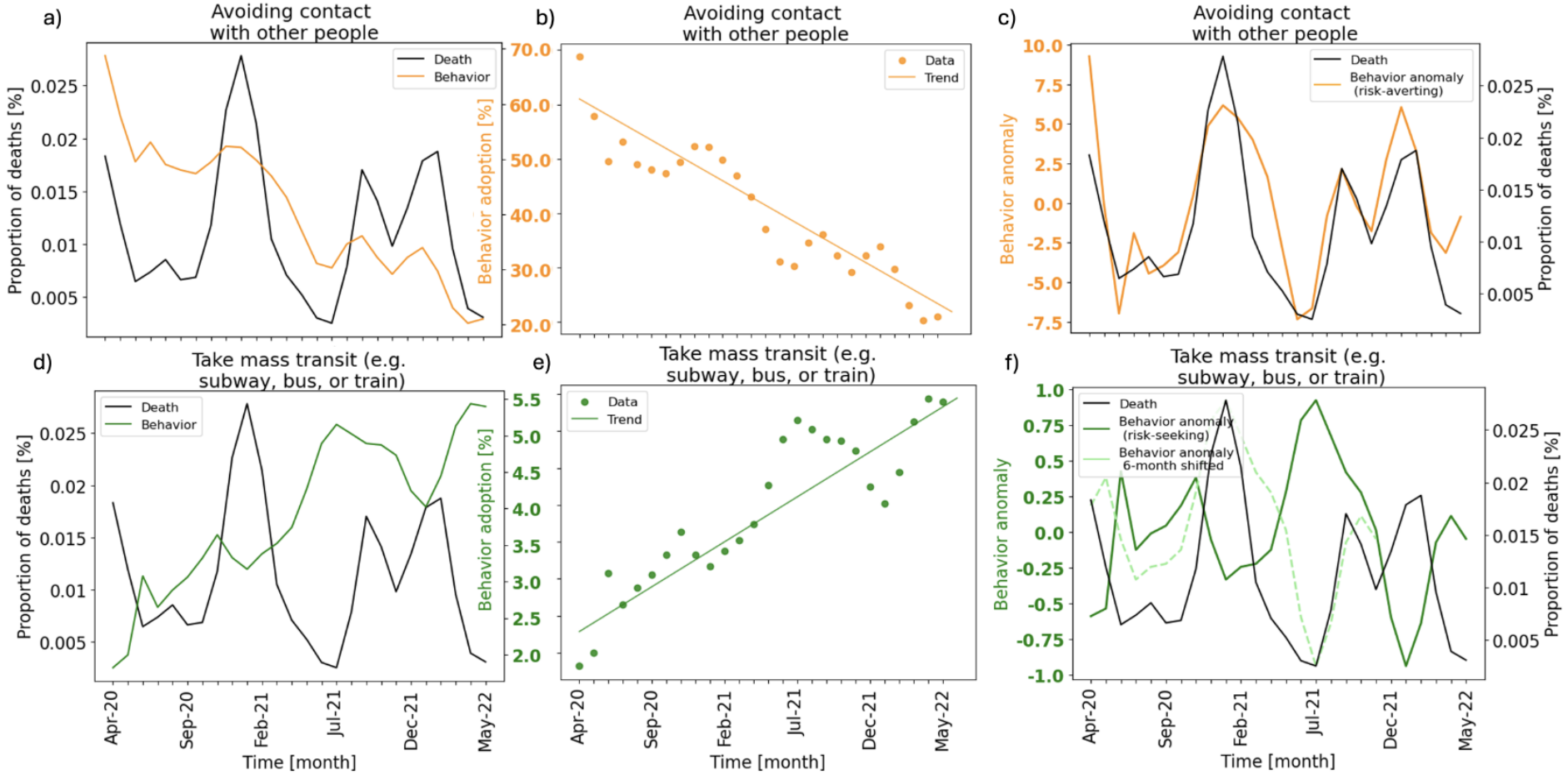
Machine Intelligence Group for the betterment of Health and the Environment

EPISTORM

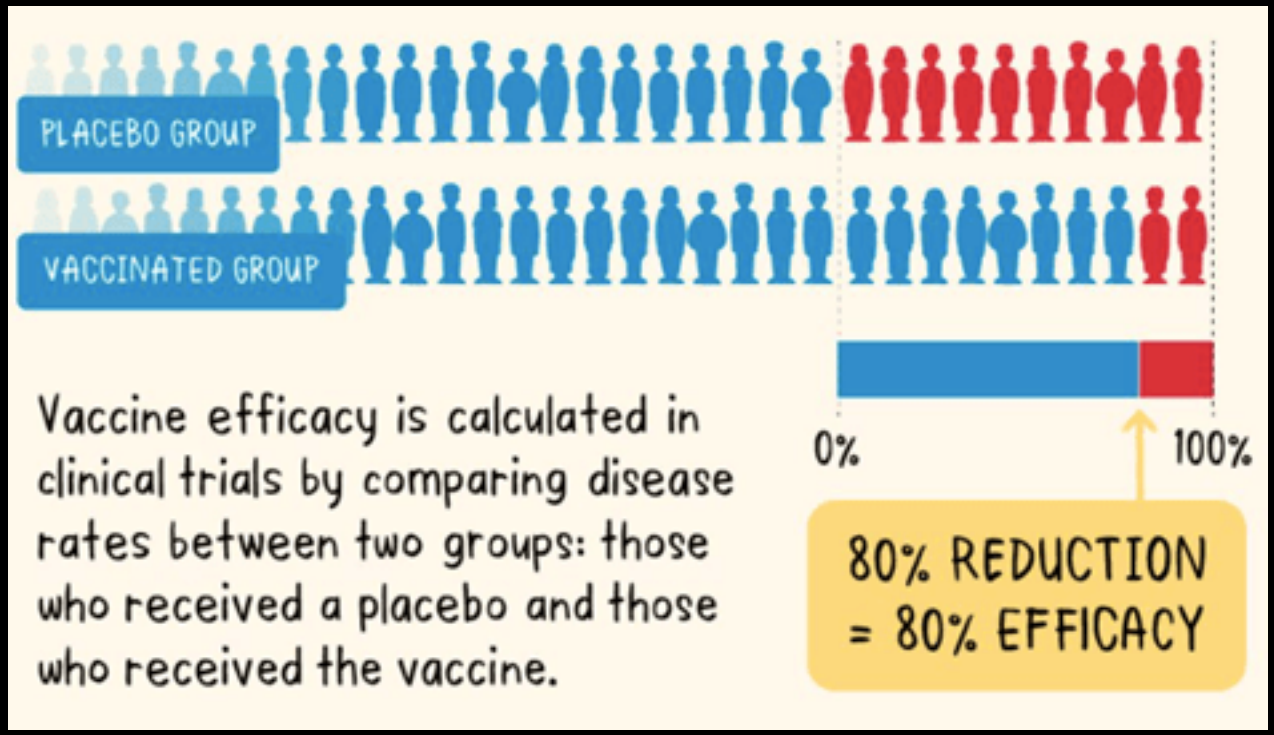
Estimated Infections using Wastewater and Survey Data



Characterizing Human Behavior Dynamics during COVID-19 pandemic in the United States and Their Association with Pandemic Severity Metrics



Vaccine efficacy



communications medicine

Article | Open Access | Published: 16 July 2021

High coverage COVID-19 mRNA vaccination rapidly controls SARS-CoV-2 transmission in long-term care facilities

Pablo M. De Salazar, Nicholas B. Link, Karuna Lamarca & Mauricio Santillana

Communications Medicine **1**, Article number: 16 (2021) | Cite this article

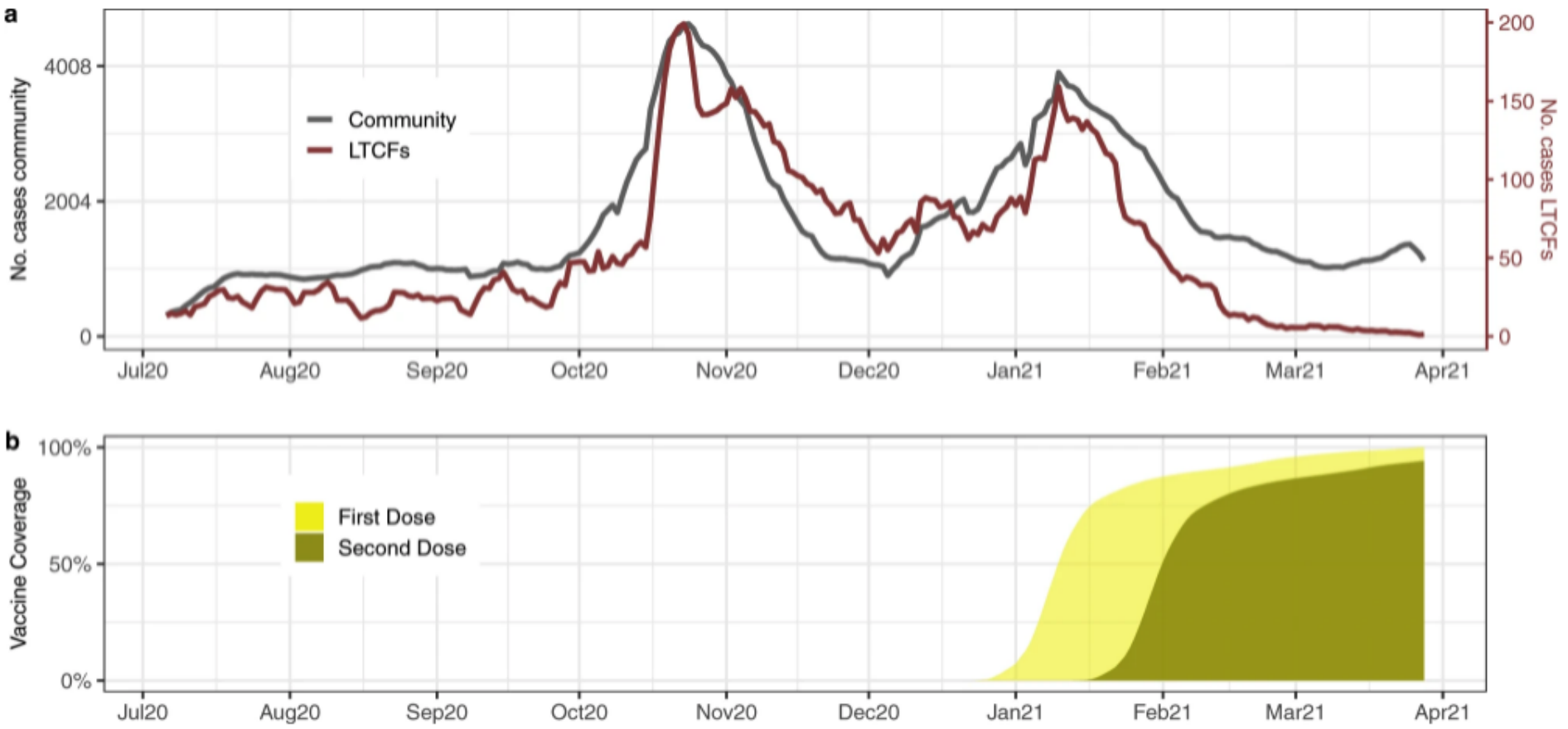
Abstract

Background

Residents of Long-Term Care Facilities (LTCFs) represent a major share of COVID-19 deaths worldwide. Measuring the vaccine effectiveness among the most vulnerable in these settings is essential to monitor and improve mitigation strategies.

Fig. 1: Documented infections and vaccinations in Catalonia, July 6, 2020–March 28, 2021.

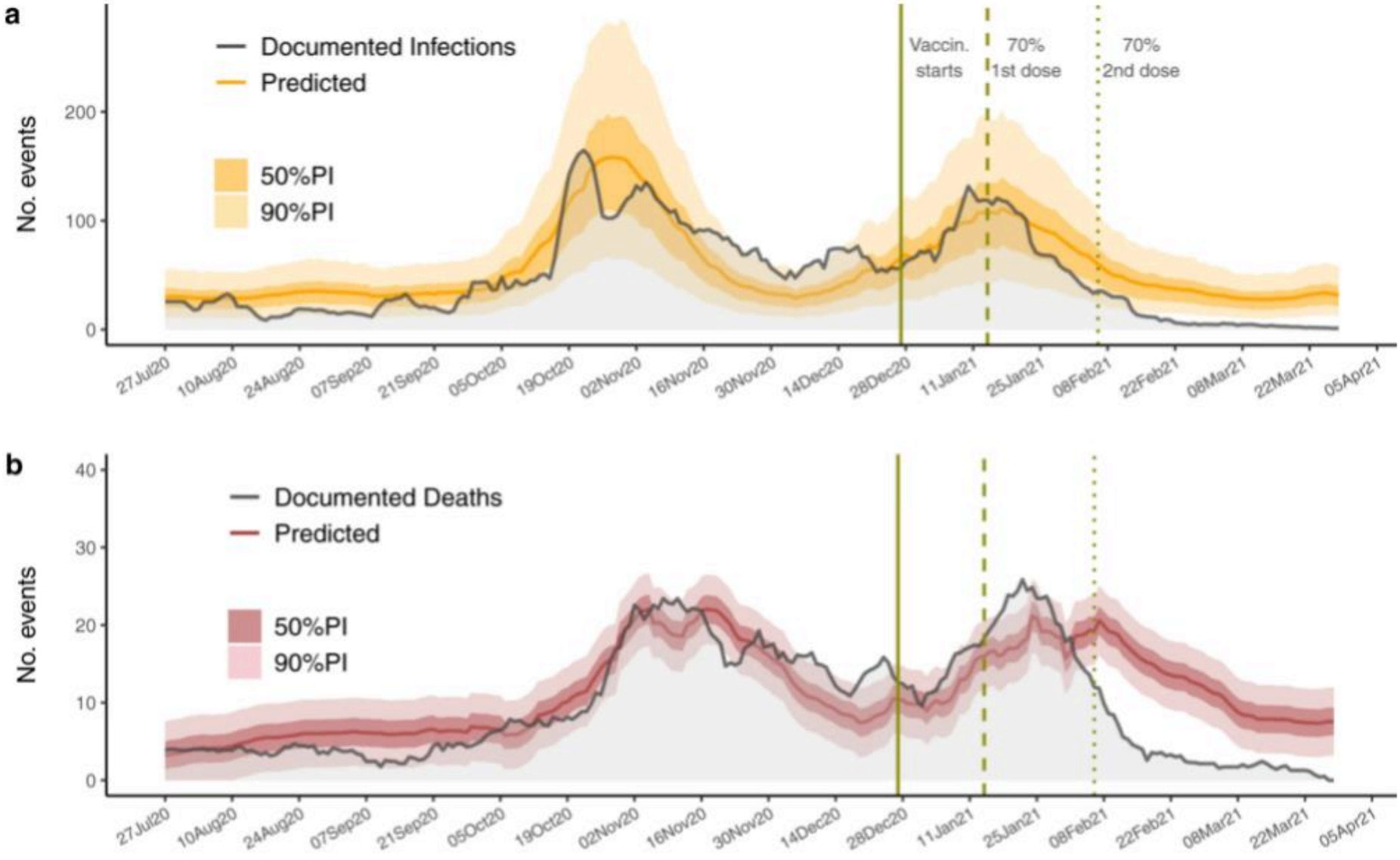
From: High coverage COVID-19 mRNA vaccination rapidly controls SARS-CoV-2 transmission in long-term care facilities

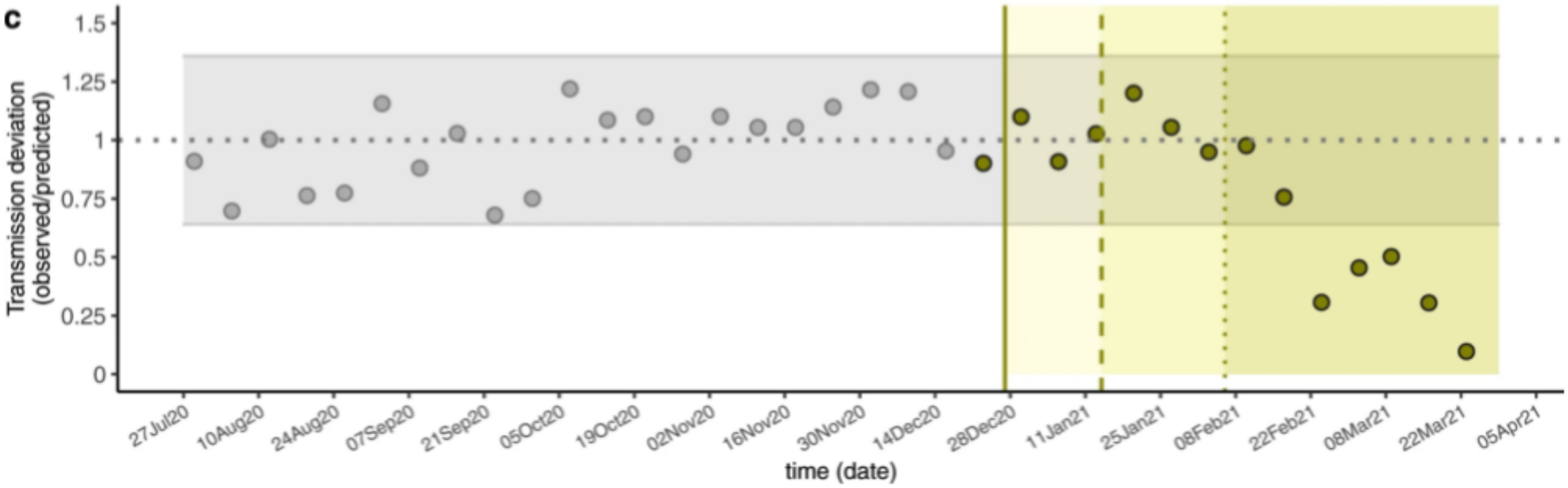


a Comparison of the total community (gray) and LTCFs' documented infections (red) trajectories in Catalonia, Spain. **b** First and second dose vaccine coverage among LTCFs' residents.

Fig. 2: Predicted vs. observed SARS-CoV-2 infections, deaths, and transmission events in Catalonia.

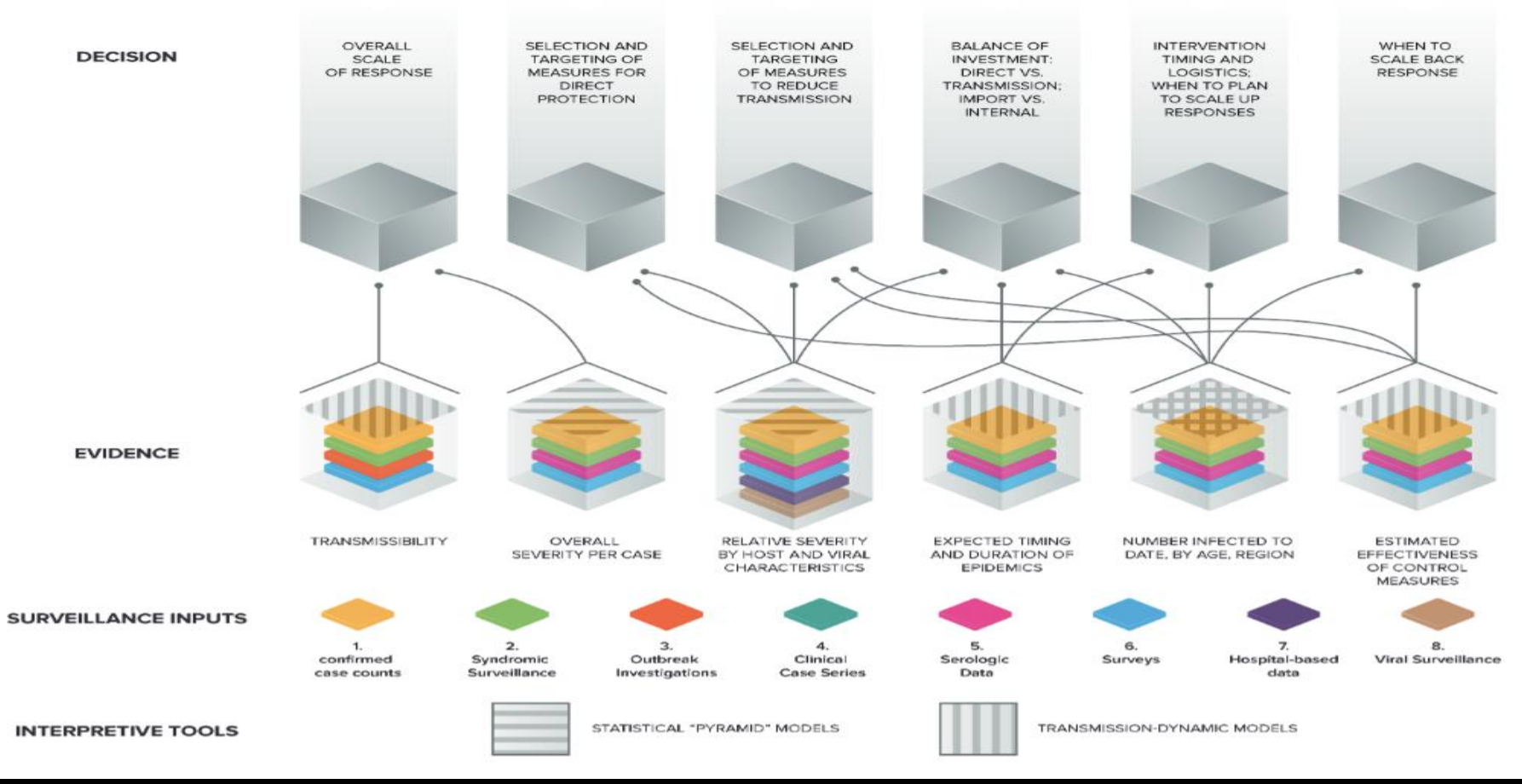
From: High coverage COVID-19 mRNA vaccination rapidly controls SARS-CoV-2 transmission in long-term care facilities

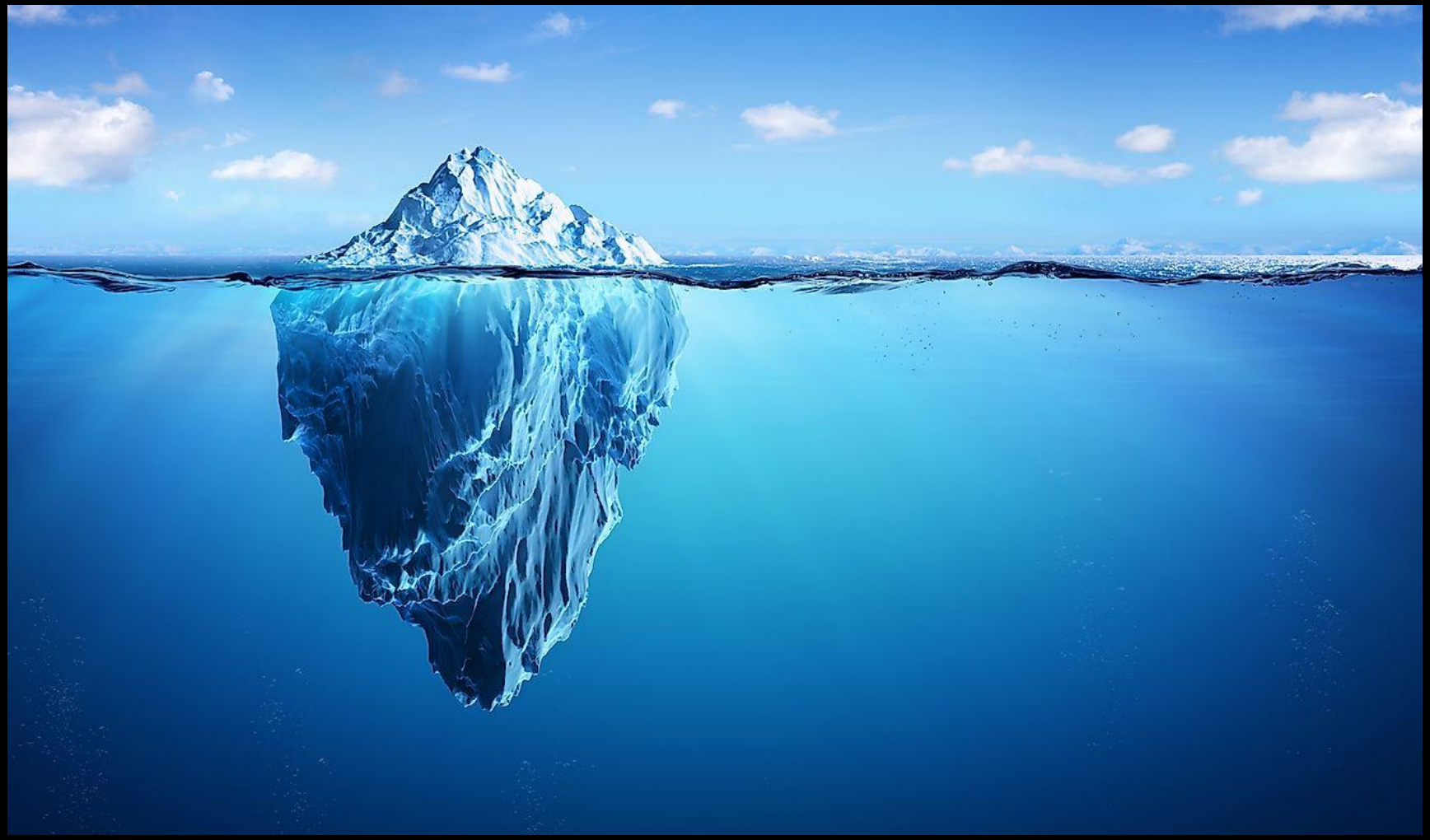




The predictions for infections (**a**) and deaths (**b**) across all of Catalonia. The solid lines show the model predictions from training July 6, 2020 through December 27, 2020, the darker shaded background shows the 50% prediction intervals (PI), and the lighter background shows the 90% PI. Vertical lines show key analysis time points: when vaccination started (solid), when 70% of residents received the first dose and when 70% of residents received the second dose. **c** The ratio between observed and predicted transmission at county level in Catalonia, represented by point estimates, gray for the training period and green for the prediction period; gray horizontal ribbon represents the 90% confidence interval. Solid green areas represent the prediction periods after vaccination starts.

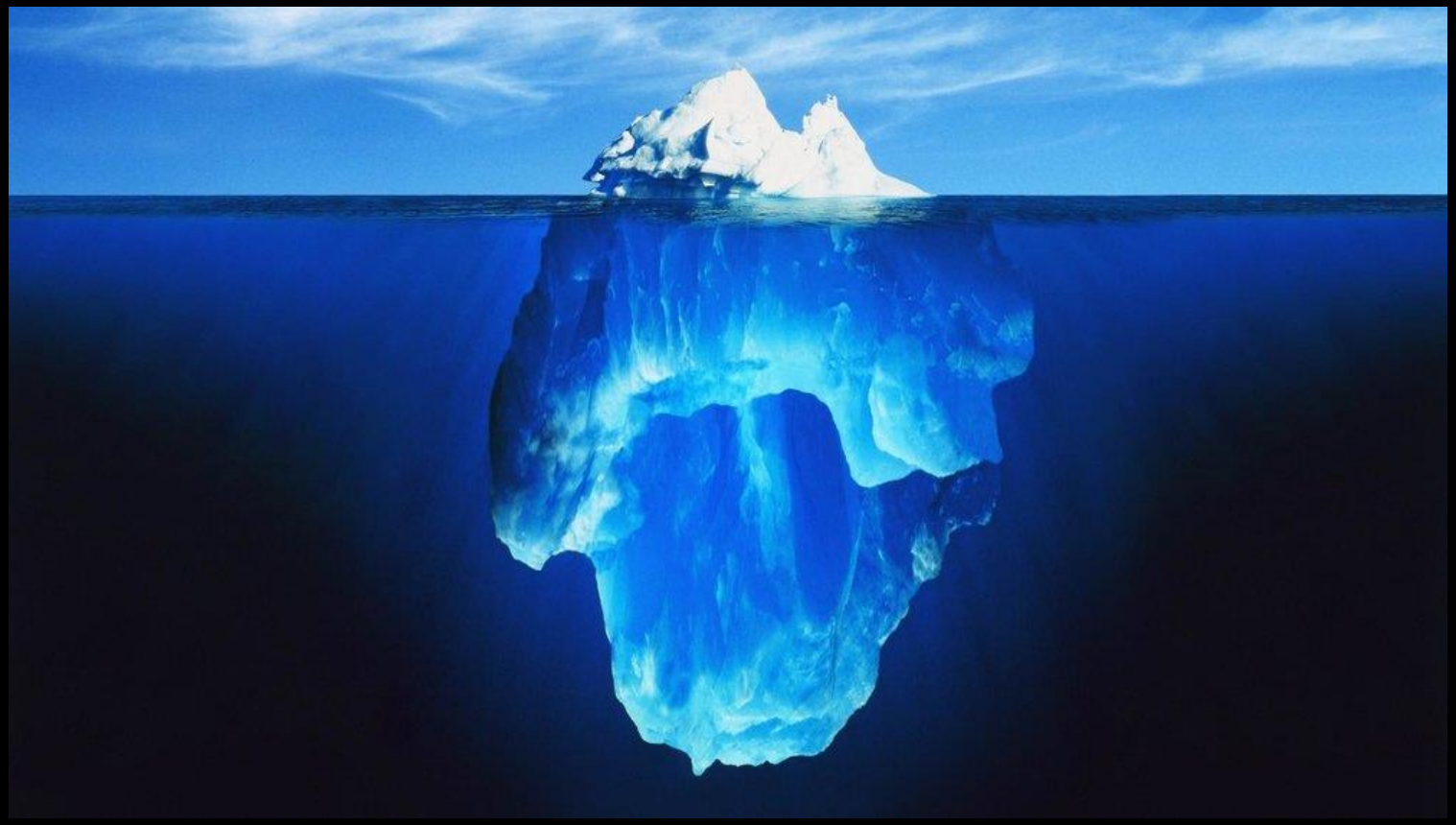
Pandemic Preparedness





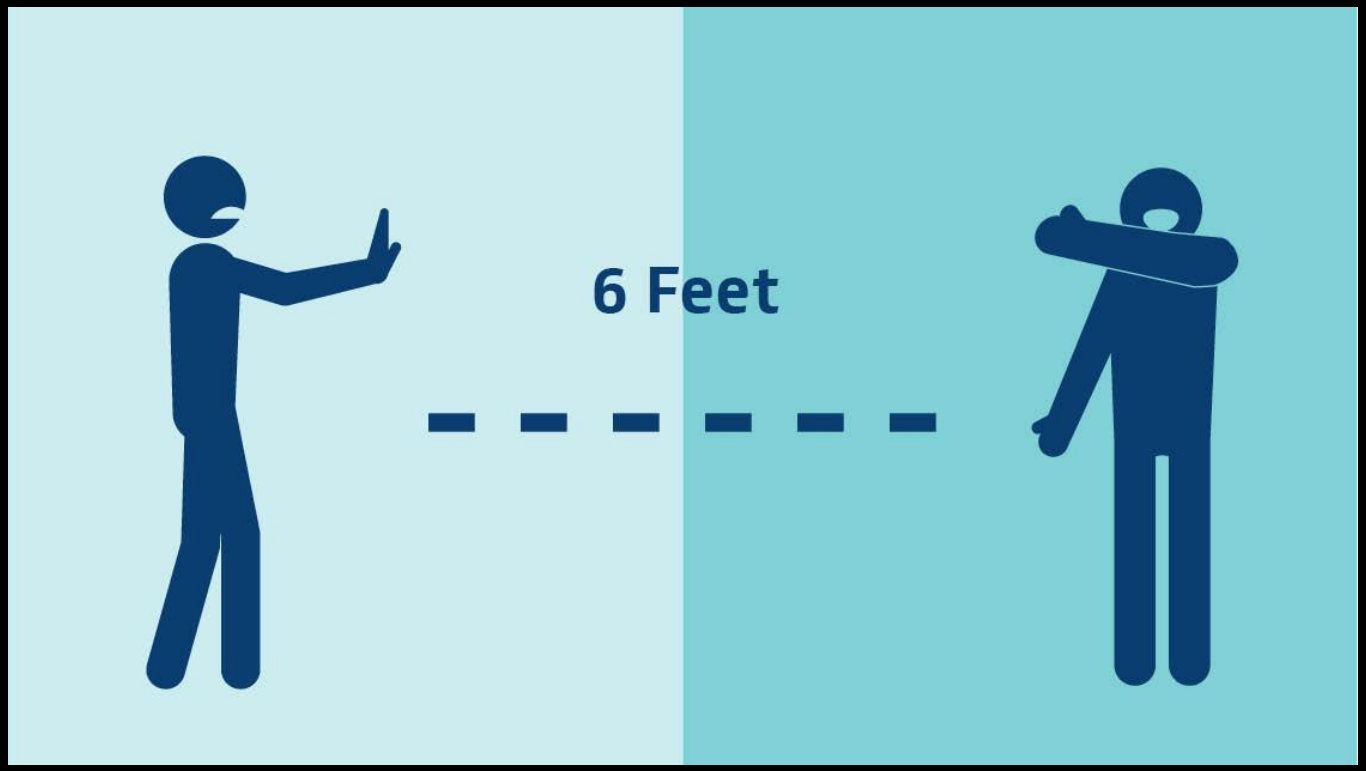
Uncertainty in Epidemiology

Why should we estimate **prevalence**?

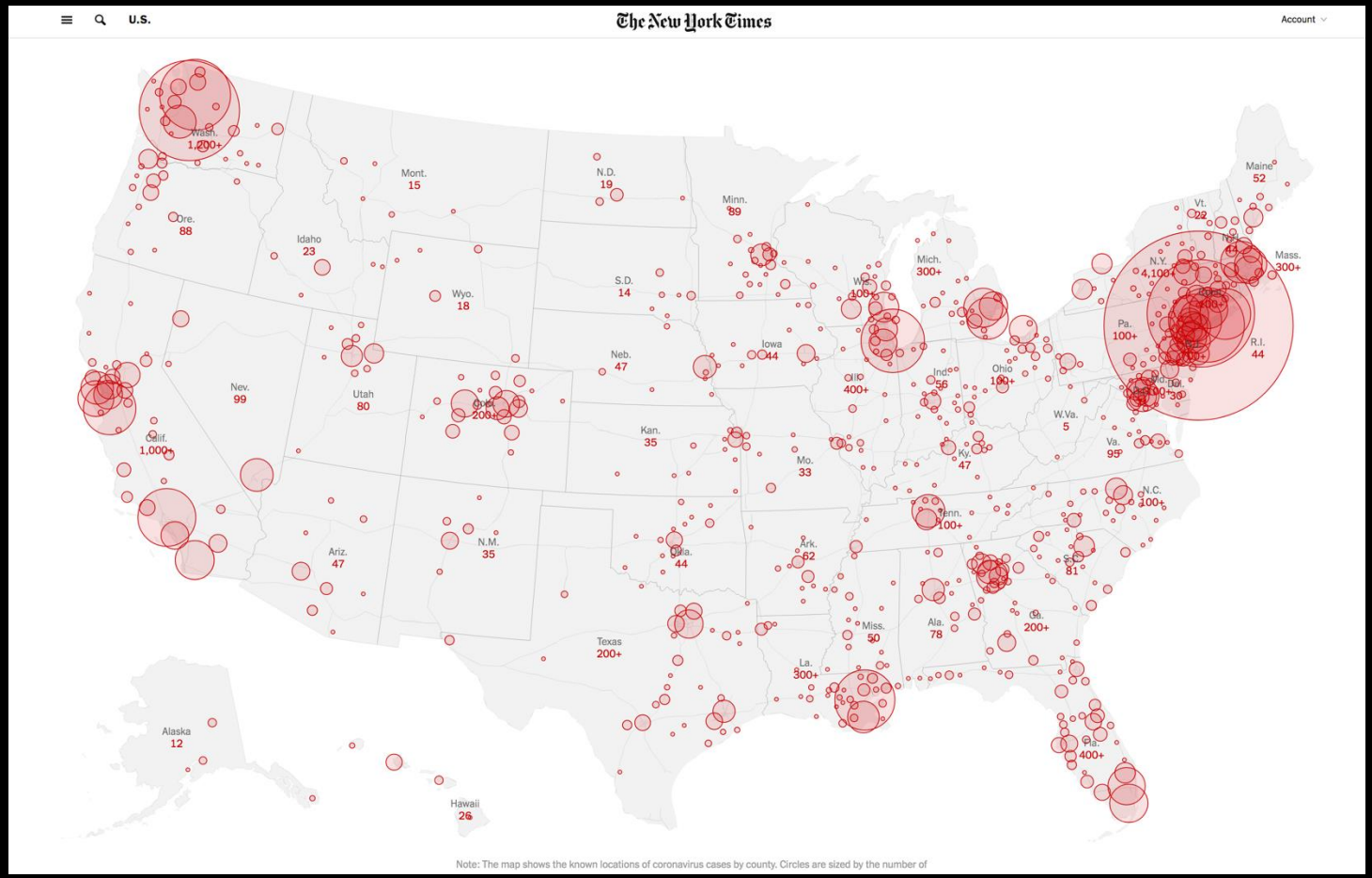


tok resources.org

Interventions (non-pharmaceutical) need to be implemented according to prevalence



How many people have been affected in the US?
Why should we estimate **prevalence**?



Widespread spatial transmission of COVID-19 can only be explained by high number of unreported cases (> 85% ?)

1. Spatial patterns of COVID-19 transmission could not be observed under currently reported numbers

2. **Unreported** cases are driving transmission in the US

3. Evidence that this happened in **China** suggests that **86% of infections were undetected**

4. Given the level of testing in the USA, this number could be higher

AAAS Become a Member

Science Contents News Careers Journals

SHARE

RESEARCH ARTICLE

Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2)

Ruiyun Li[1,*], Sen Pei[2,*,†], Bin Chen[3,*], Yimeng Song[4], Tao Zhang[5], Wan Yang[6], Jeffrey Shaman[2,†]
+ See all authors and affiliations

Science 16 Mar 2020:
eabb3221
DOI: 10.1126/science.abb3221

Article Figures & Data Info & Metrics eLetters PDF

Abstract

Estimation of the prevalence and contagiousness of undocumented novel coronavirus (SARS-CoV2) infections is critical for understanding the overall prevalence and pandemic potential of this disease. Here we use observations of reported infection within China, in conjunction with mobility data, a networked dynamic metapopulation model and Bayesian inference, to infer critical epidemiological characteristics associated with SARS-CoV2, including the fraction of undocumented infections and their contagiousness. We estimate 86% of all infections were undocumented (95% CI: [82%–90%]) prior to 23 January 2020 travel restrictions. Per person, the transmission rate of undocumented infections was 55% of documented infections ([46%–62%]), yet, due to their greater numbers, undocumented infections were the infection source for 79% of documented cases. These findings explain the rapid geographic spread of SARS-CoV2 and indicate containment of this virus will be particularly challenging.

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Estimating the cumulative incidence of COVID-19 in the United States using influenza surveillance, virologic testing, and mortality data: Four complementary approaches

Fred S. Lu, Andre T. Nguyen, Nicholas B. Link, Mathieu Molina, Jessica T. Davis, Matteo Chinazzi, Xinyue Xiong, Alessandro Vespignani, Marc Lipsitch, Mauricio Santillana

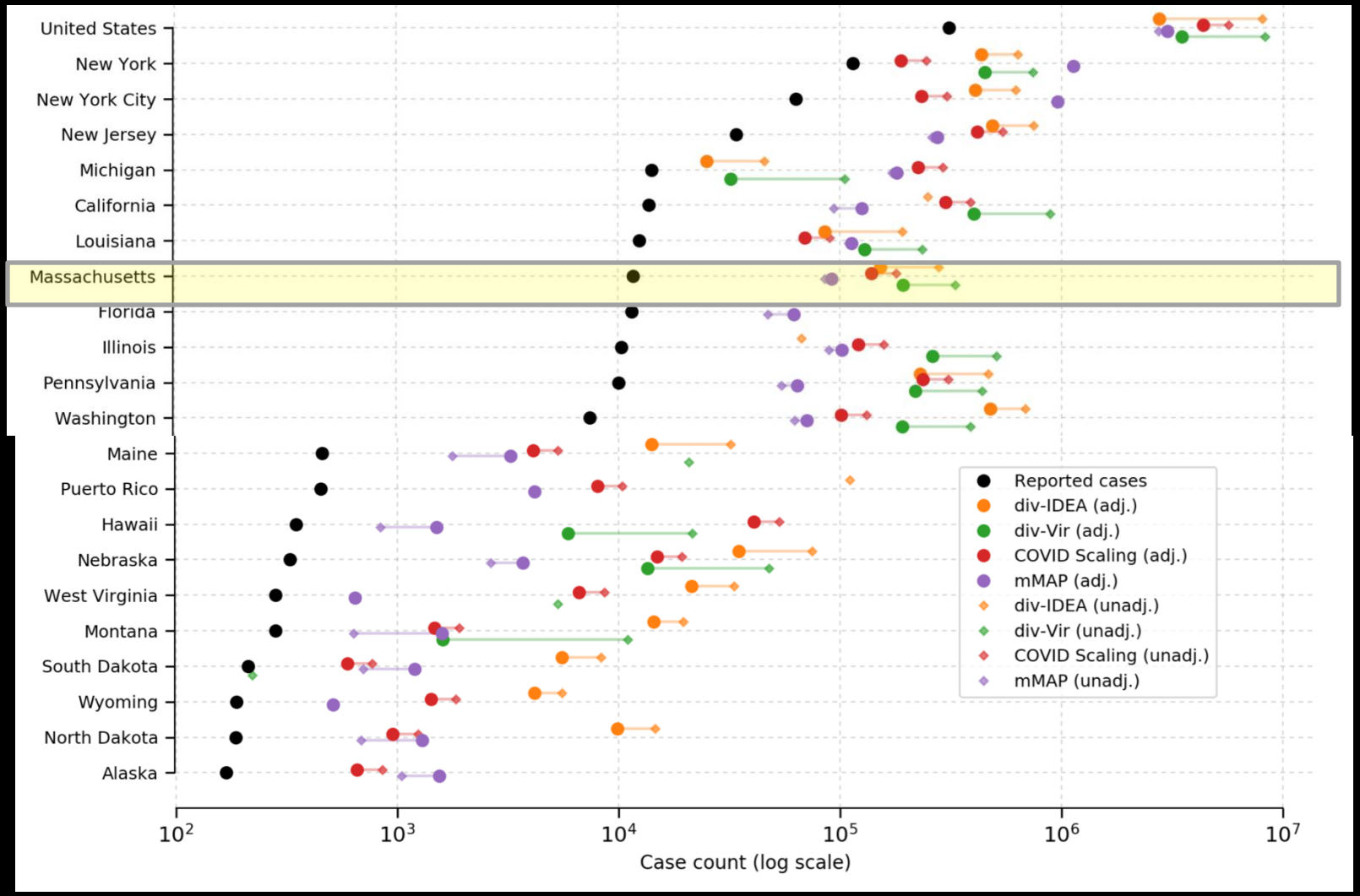
Published: June 17, 2021 • https://doi.org/10.1371/journal.pcbi.1008994

Abstract

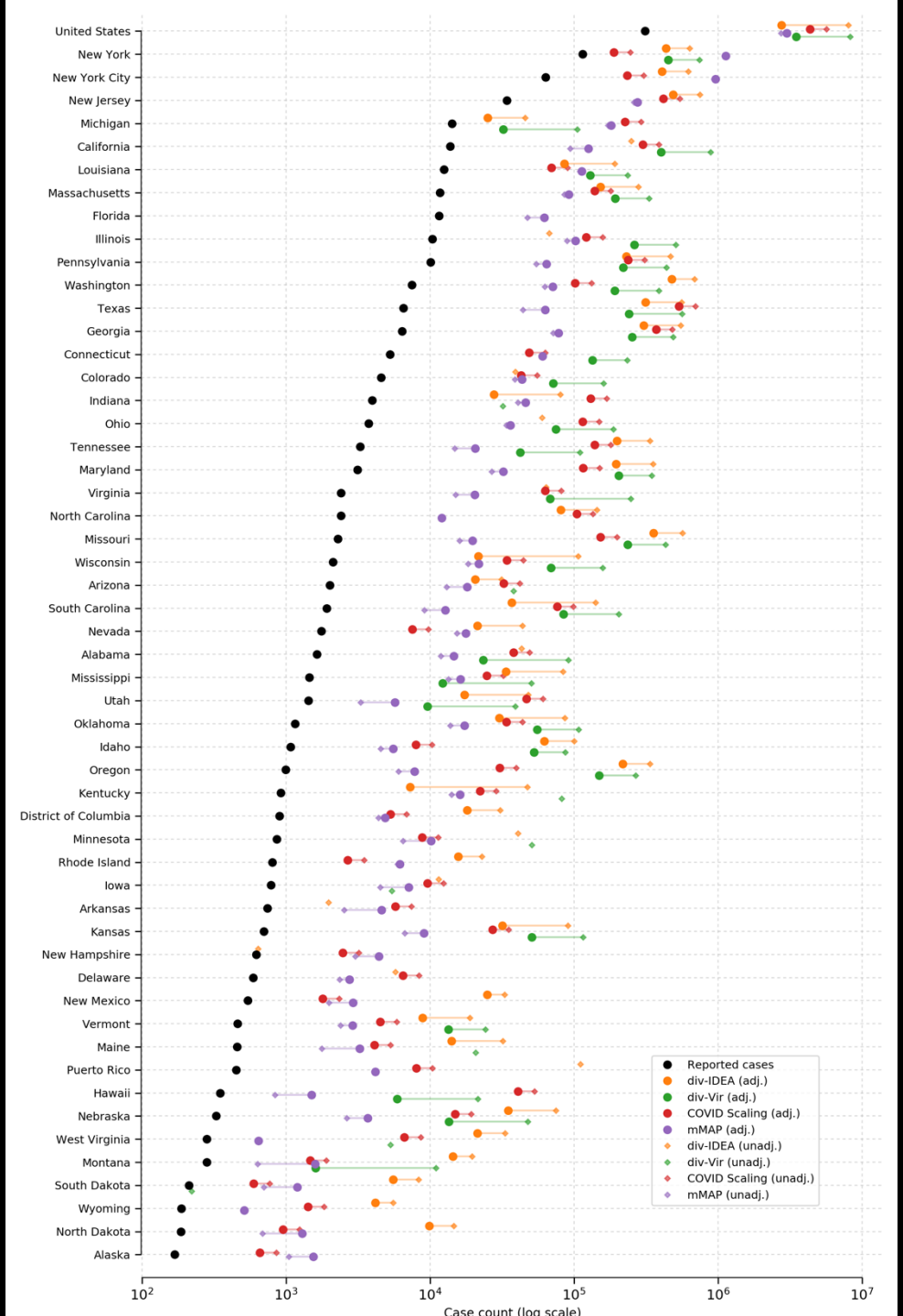
Effectively designing and evaluating public health responses to the ongoing COVID-19 pandemic requires accurate estimation of the prevalence of COVID-19 across the United States (US). Equipment shortages and varying testing capabilities have however hindered the usefulness of the official reported positive COVID-19 case counts. We introduce four complementary approaches to estimate the cumulative incidence of symptomatic COVID-19 in each state in the US as well as Puerto Rico and the District of Columbia, using a combination of excess influenza-like illness reports, COVID-19 test statistics, COVID-19 mortality reports, and a spatially structured epidemic model. Instead of relying on the estimate from a single data source or method that may be biased, we provide multiple estimates, each relying on different assumptions and data sources. Across our four approaches emerges the consistent conclusion that on April 4, 2020, the estimated case count was 5 to 50 times higher than the official positive test counts across the different states. Nationally, our estimates of COVID-19 symptomatic cases as of April 4 have a likely range of 2.3 to 4.8 million, with possibly as many as 7.6 million cases, up to 25 times greater than the cumulative confirmed cases of about 311,000. Extending our methods to May 16, 2020, we estimate that cumulative symptomatic incidence ranges from 4.9 to 10.1 million, as opposed to 1.5 million positive test counts. The proposed combination of approaches may prove useful in assessing the burden of COVID-19 during resurgences in the US and other countries with comparable surveillance systems.

Subset of US states

As of April 4th, 2020

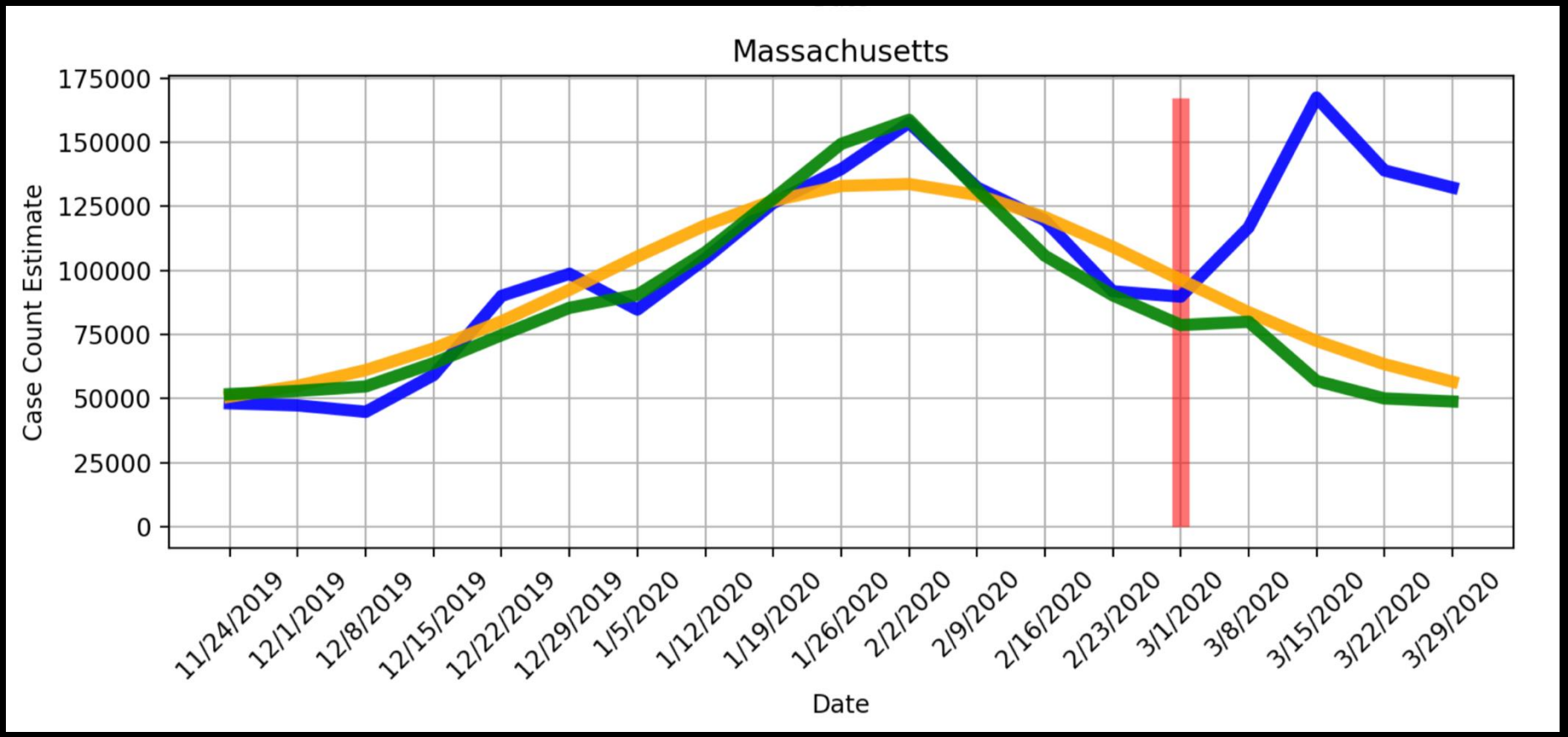


As of April 4th, 2020



As of April 4th, 2020

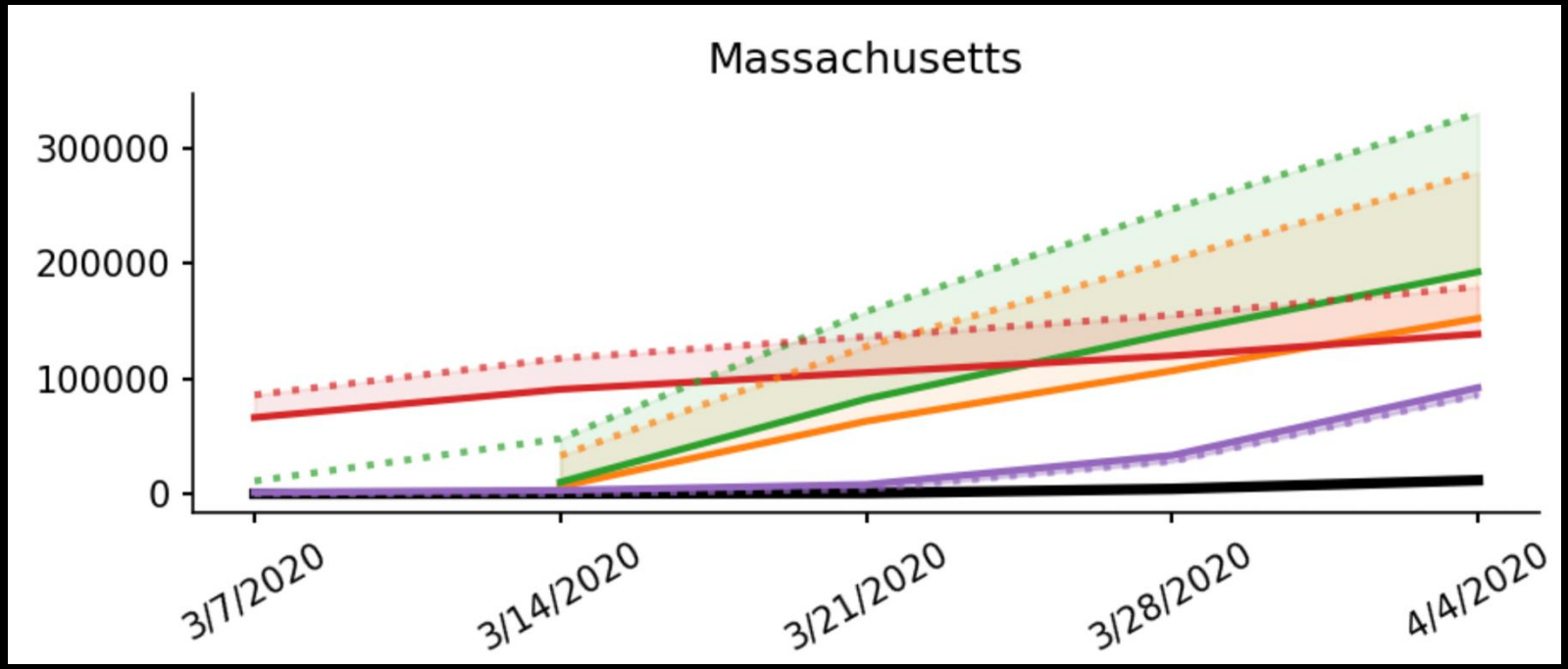
In MA we estimate about **120,000** COVID-19 infected
Compare to



As of April 4th, 2020

As of April 16th, 2020

In MA we estimate about **120,000** COVID-19 infected
Compare to



PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Near real-time surveillance of the SARS-CoV-2 epidemic with incomplete data

Pablo M. De Salazar, Fred Lu, James A Hay, Diana Gómez-Barroso, Pablo Fernández-Navarro, Elena V Martínez, Jenaro Astray-Mochales, Rocío Amillategui, Ana García-Fulgueiras, Maria D Chirlaque, Alonso Sánchez-Migallón, Amparo Larrauri, María J Sierra, [...], Miguel A Hernán [view all]

Version 2 Published: March 31, 2022 • https://doi.org/10.1371/journal.pcbi.1009964

See the preprint

Abstract

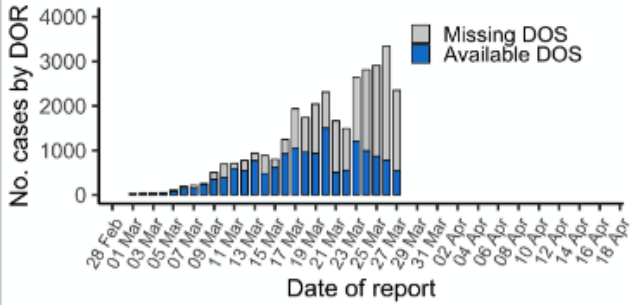
When responding to infectious disease outbreaks, rapid and accurate estimation of the epidemic trajectory is critical. However, two common data collection problems affect the reliability of the epidemiological data in real time: missing information on the time of first symptoms, and retrospective revision of historical information, including right censoring. Here, we propose an approach to construct epidemic curves in near real time that addresses these two challenges by 1) imputation of dates of symptom onset for reported cases using a dynamically-estimated "backward" reporting delay conditional distribution, and 2) adjustment for right censoring using the *NobBS* software package to nowcast cases by date of symptom onset. This process allows us to obtain an approximation of the time-varying reproduction number (R_t) in real time. We apply this approach to characterize the early SARS-CoV-2 outbreak in two Spanish regions between March and April 2020. We evaluate how these real-time estimates compare with more complete epidemiological data that became available later. We explore the impact of the different assumptions on the estimates, and compare our estimates with those obtained from commonly used surveillance approaches. Our framework can help improve accuracy, quantify uncertainty, and evaluate frequently unstated assumptions when recovering the epidemic curves from limited data obtained from public health systems in other locations.

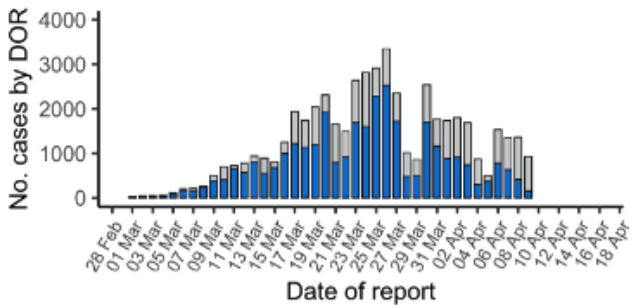
Early analysis
March 1-27, 2020

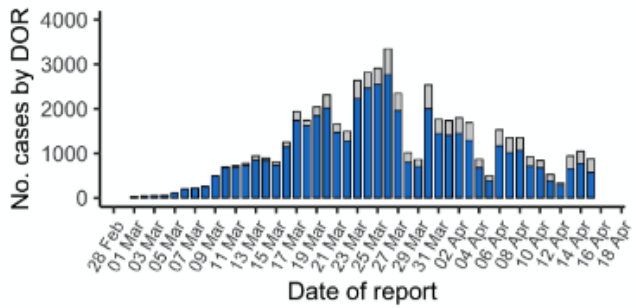
Intermediate analysis
March 1-April 9, 2020

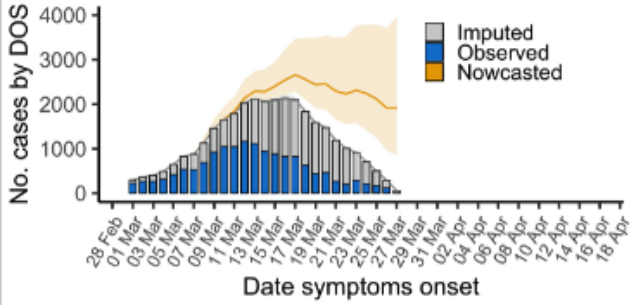
Late analysis
March 1-April 16, 2020

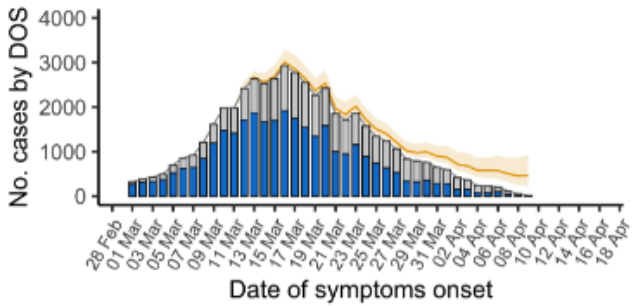
Madrid

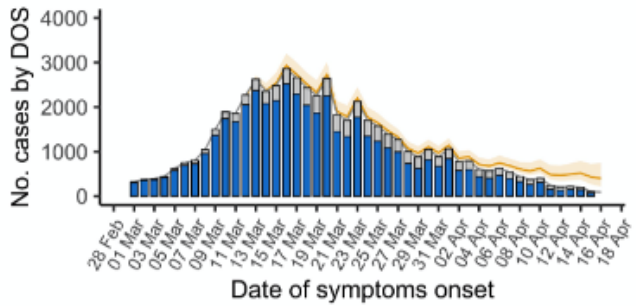


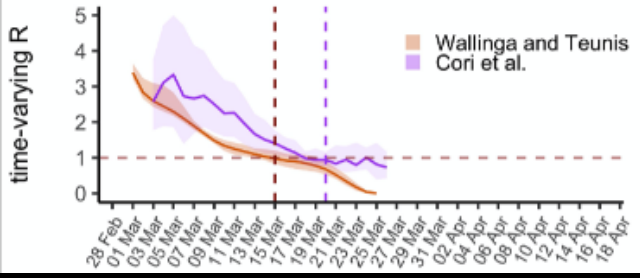


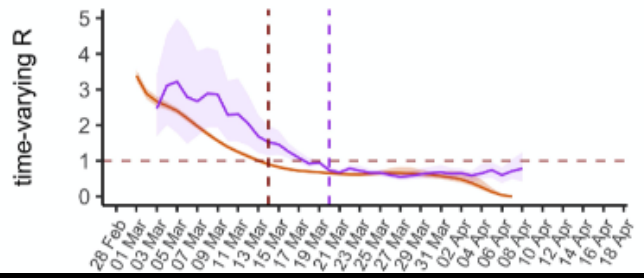


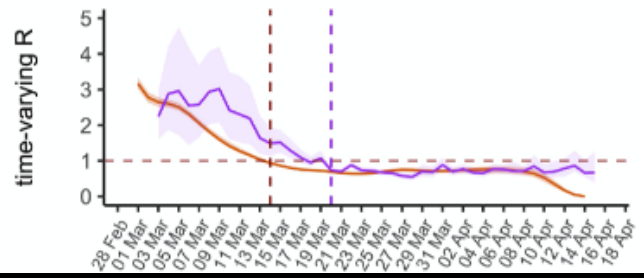


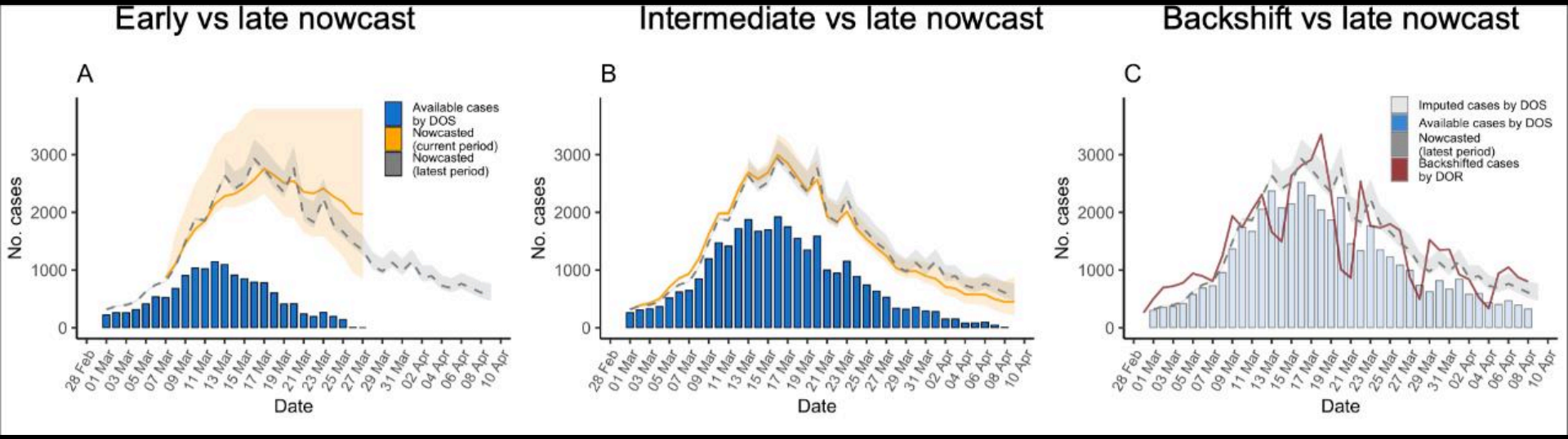










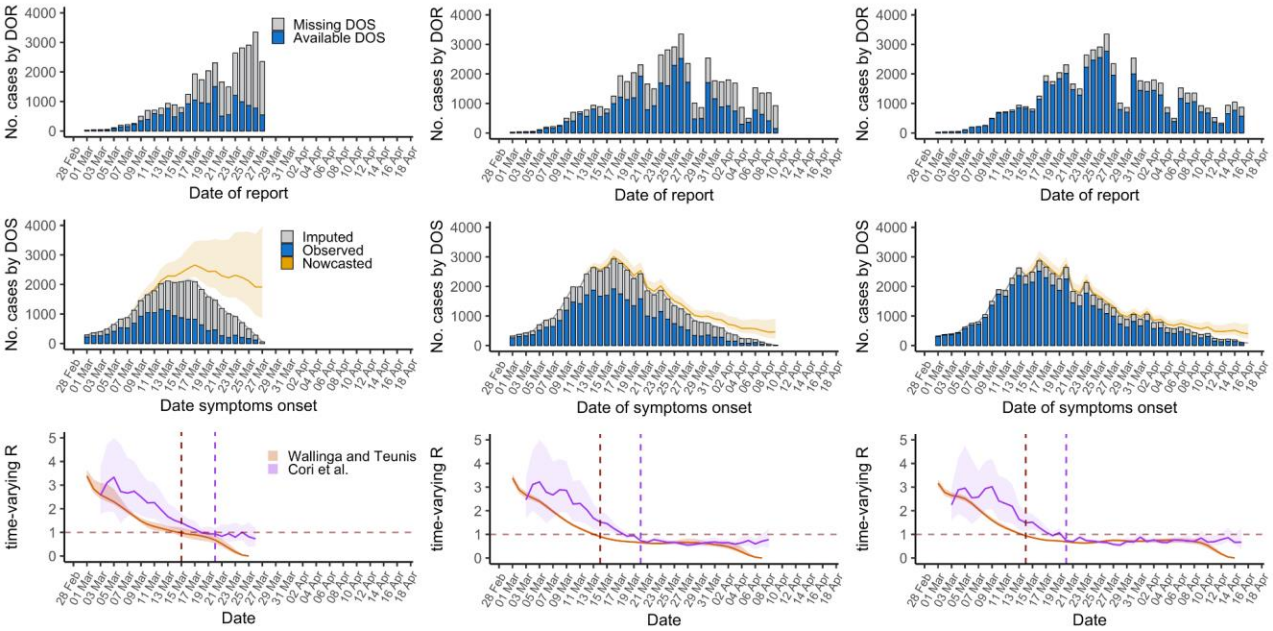

Early vs late nowcast
A
Available cases by DOS
Nowcasted (current period)
Nowcasted (latest period)
No. cases
Date
Intermediate vs late nowcast
B
No. cases
Date
Backshift vs late nowcast
C
Imputed cases by DOS
Available cases by DOS
Nowcasted (latest period)
Backshifted cases by DOR
No. cases
Date

Early analysis
March 1-27, 2020

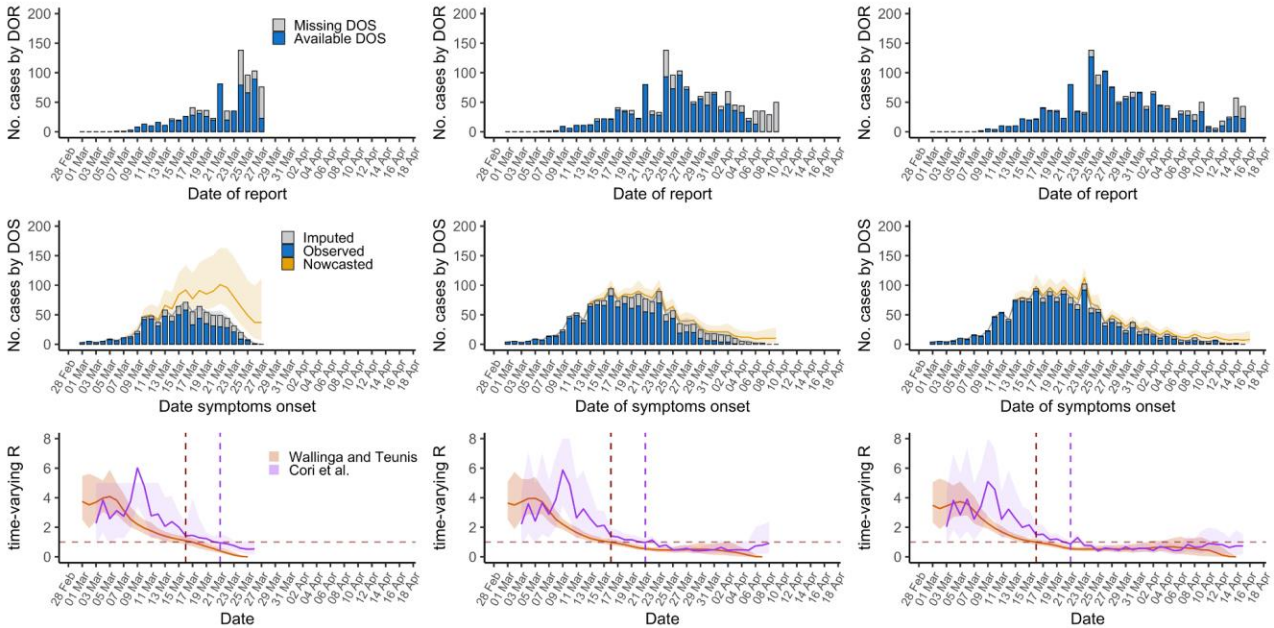
Intermediate analysis
March 1-April 9, 2020

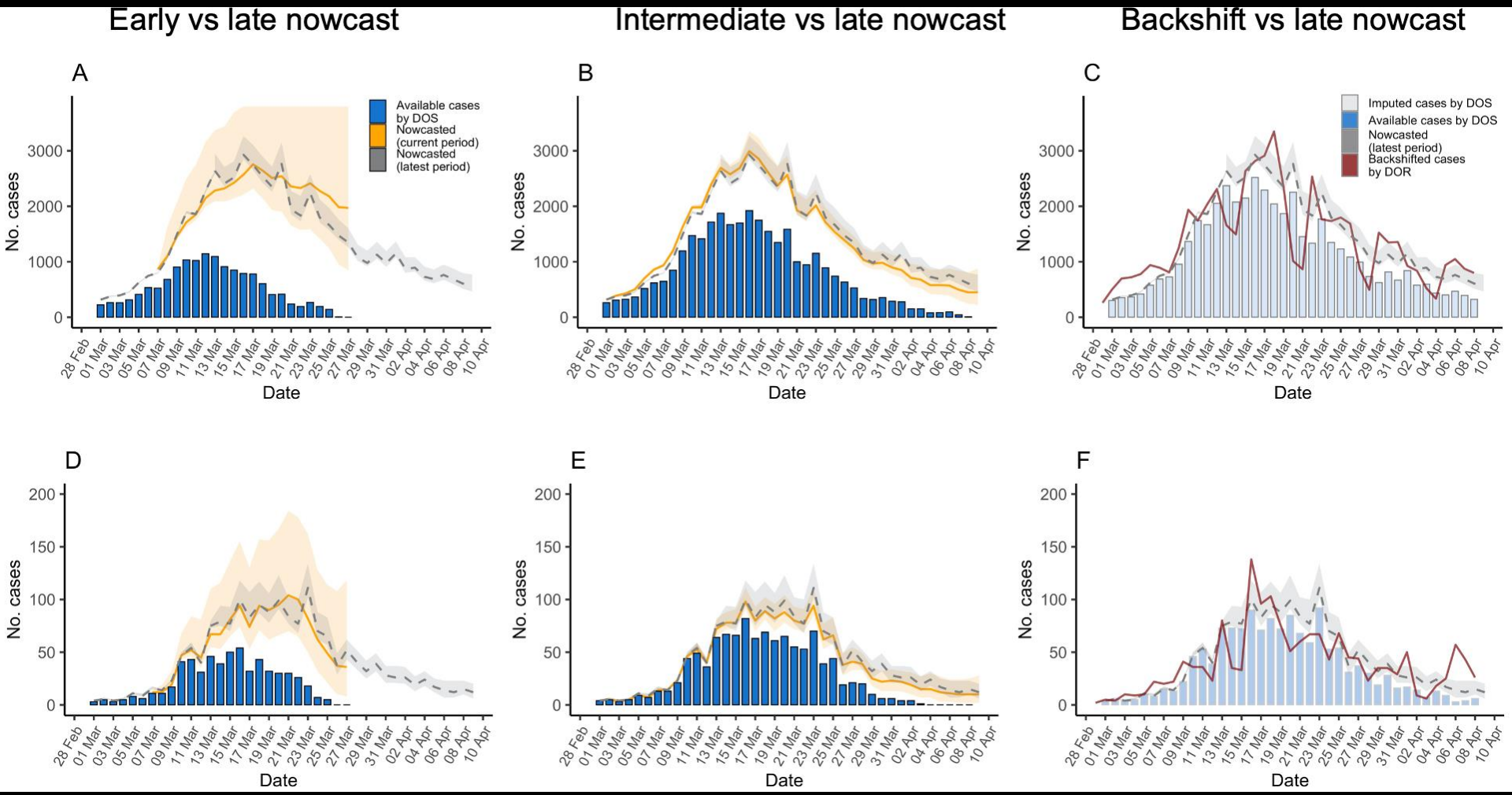
Late analysis
March 1-April 16, 2020

Madrid

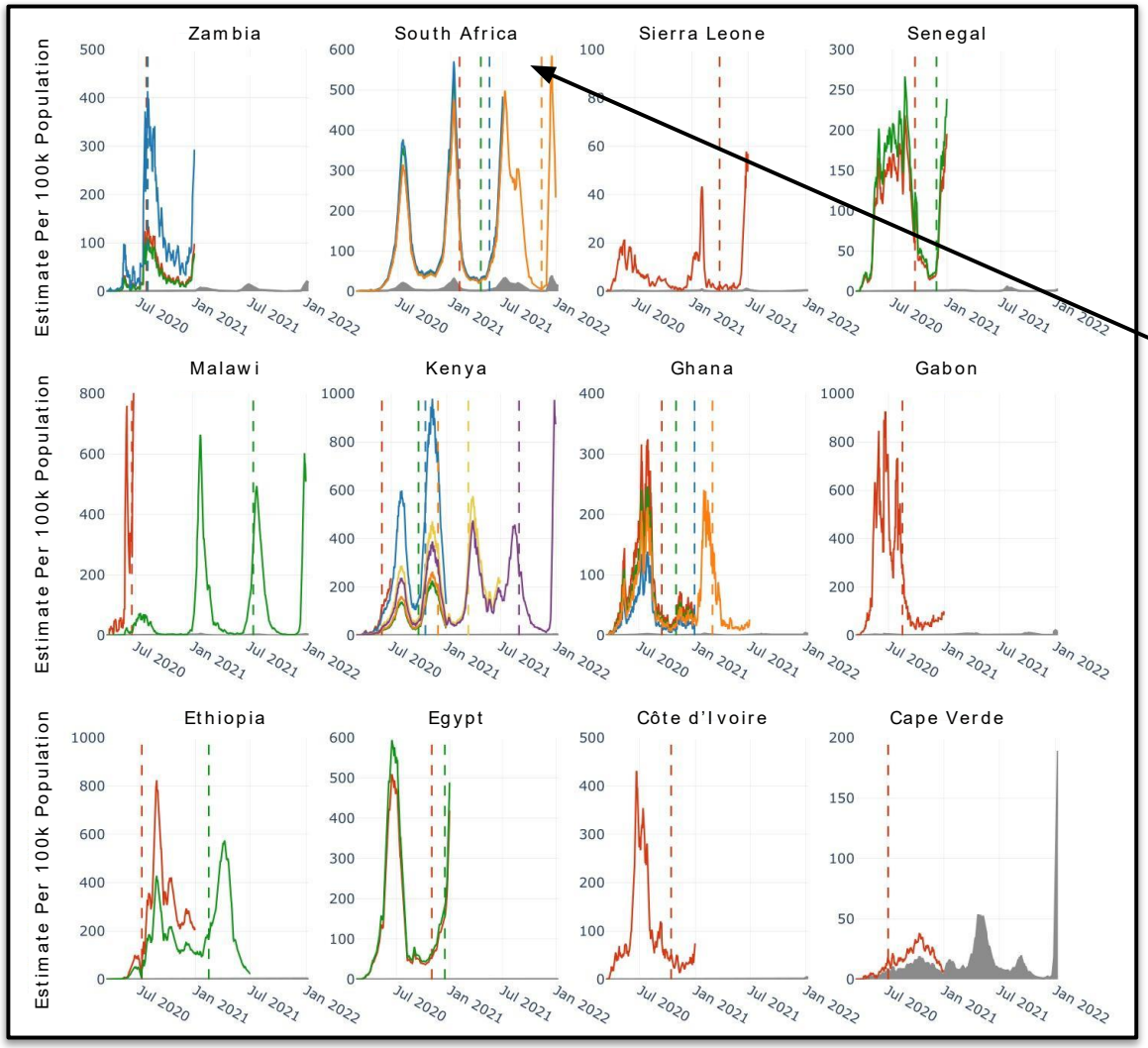


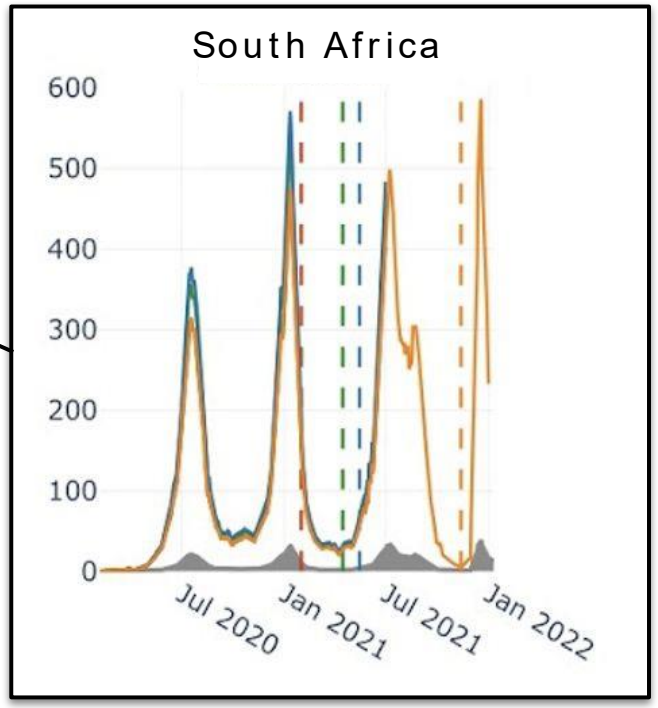
Murcia

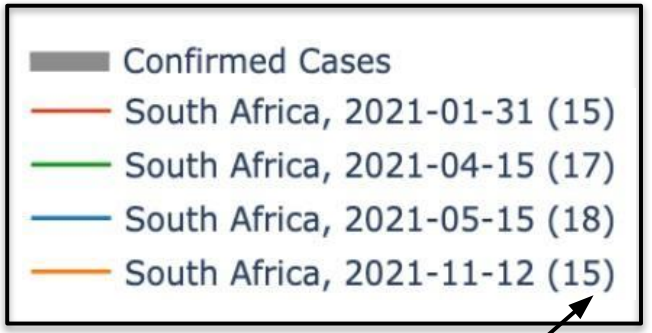



Early vs late nowcast
Intermediate vs late nowcast
Backshift vs late nowcast
A
B
C
D
E
F
Available cases by DOS
Nowcasted (current period)
Nowcasted (latest period)
Imputed cases by DOS
Backshifted cases by DOR
No. cases
Date

Reported COVID-19 Cases vs. Serology-Derived COVID-19 Infection Estimates in 12 African Nations

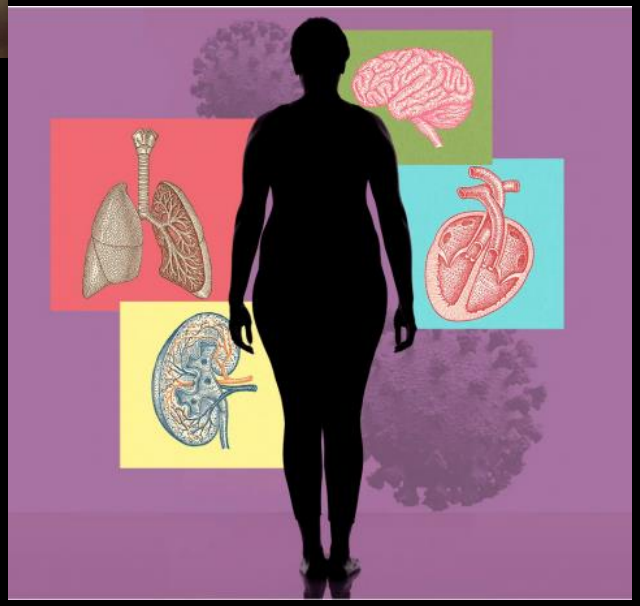






Multiplier

What about the role of social determinants of health?



RESEARCH ARTICLE

Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile

GONZALO E. MENA, PAMELA P. MARTINEZ, AYESHA S. MAHMUD, PABLO A. MARQUET, CAROLINE O. BUCKEE, AND MAURICIO SANTILLANA Authors Info & Affiliations

SCIENCE • 28 May 2021 • Vol 372, Issue 6545 • DOI: 10.1126/science.abg5298

Abstract

The current COVID-19 pandemic has impacted cities particularly hard. Here, we provide an in-depth characterization of disease incidence and mortality, and their dependence on demographic and socioeconomic strata in Santiago, a highly segregated city and the capital of Chile. Our analyses show a strong association between socioeconomic status and both COVID-19 outcomes and public health capacity. People living in municipalities with low socioeconomic status did not reduce their mobility during lockdowns as much as those in more affluent municipalities. Testing volumes may have been insufficient early in the pandemic in those places, and both test positivity rates and testing delays were much higher. We find a strong association between socioeconomic status and mortality, measured either by COVID-19 attributed deaths or excess deaths. Finally, we show that infection fatality rates in young people are higher in low-income municipalities. Together, these results highlight the critical consequences of socioeconomic inequalities on health outcomes.

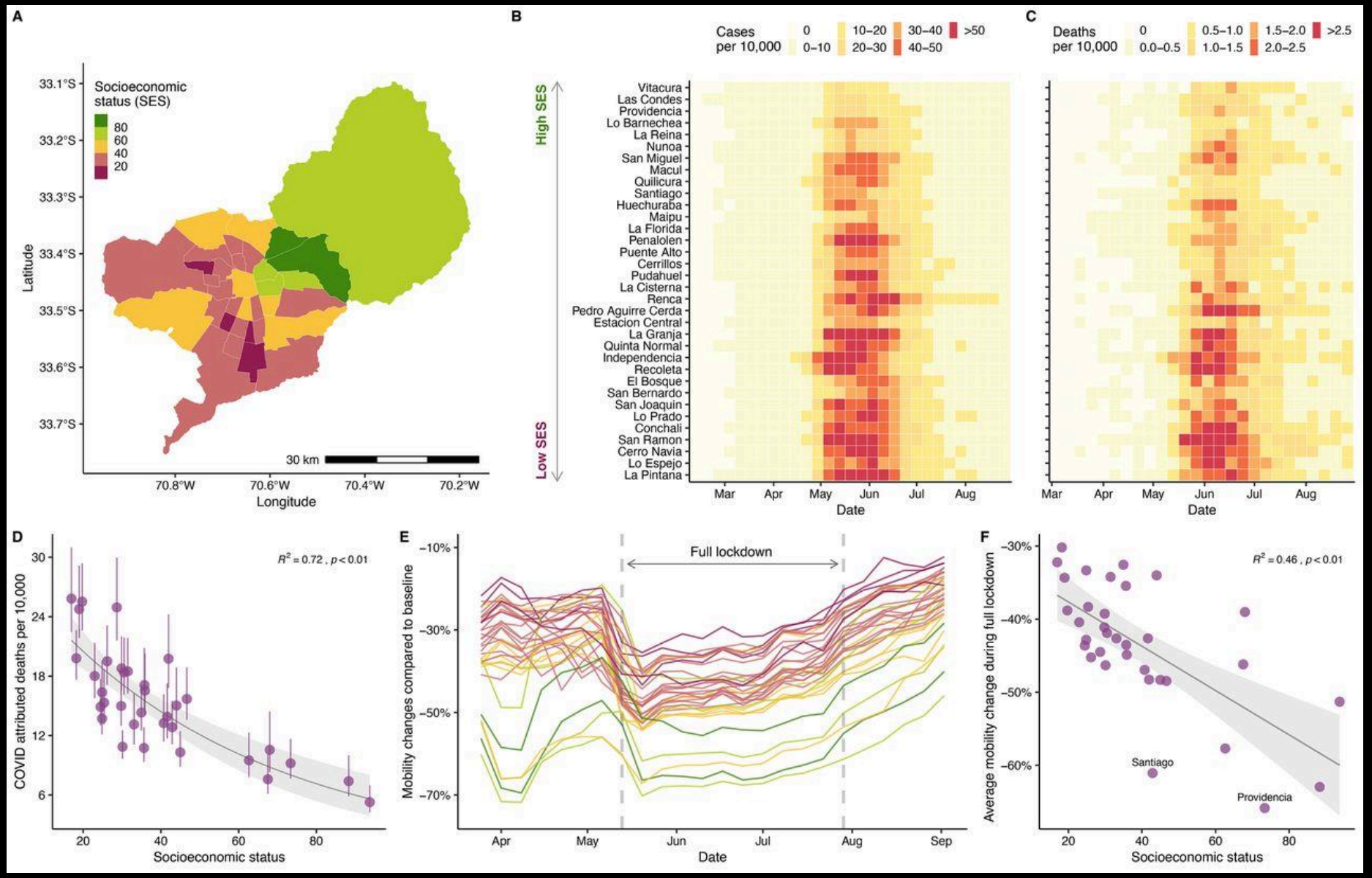
Pamela Martinez
University of Illinois
Urbana-Champaign

Gonzalo Mena
University of Oxford

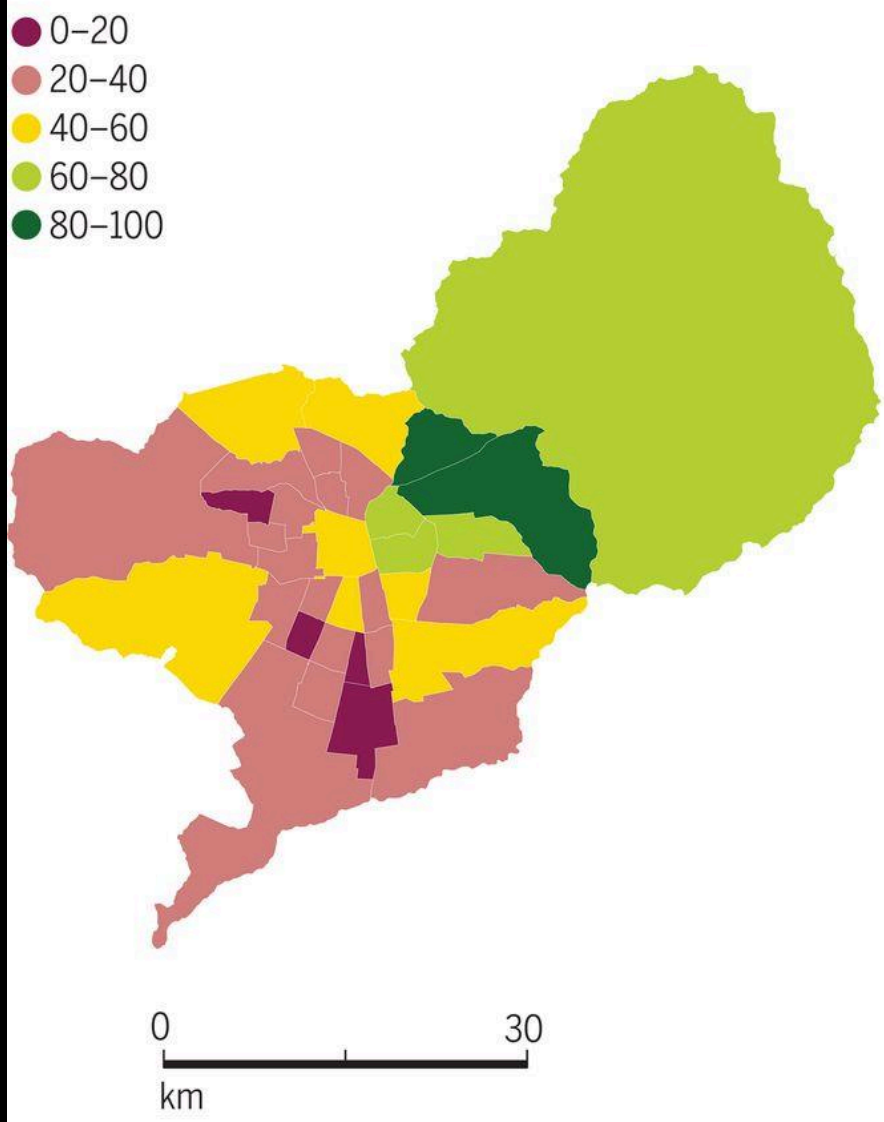
Ayesha Mahmud
University of
California, Berkeley

Pablo Marquet
Pontificia Universidad
Catolica de Chile

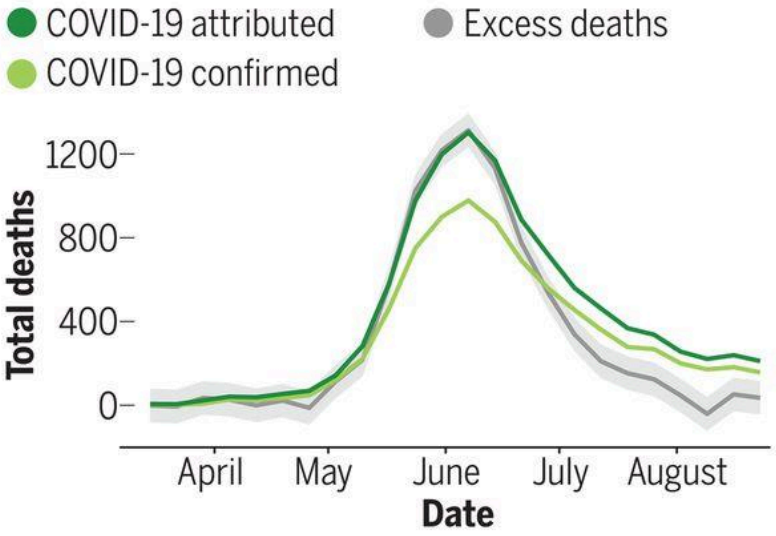
Caroline Buckee
Harvard University


A
Socioeconomic status (SES)
80
60
40
20
33.1°S
33.2°S
33.3°S
33.4°S
33.5°S
33.6°S
33.7°S
Latitude
30 km
70.8°W
70.6°W
70.4°W
70.2°W
Longitude
B
Cases per 10,000
0
0–10
10–20
20–30
30–40
40–50
>50
High SES
Low SES
Vitacura
Las Condes
Providencia
Lo Barnechea
La Reina
Nunoa
San Miguel
Macul
Quilicura
Santiago
Huechuraba
Maipu
La Florida
Penalolen
Puente Alto
Cerrillos
Pudahuel
La Cisterna
Renca
Pedro Aguirre Cerda
Estacion Central
La Granja
Quinta Normal
Independencia
Recoleta
El Bosque
San Bernardo
San Joaquin
Lo Prado
Conchali
San Ramon
Cerro Navia
Lo Espejo
La Pintana
Mar
Apr
May
Jun
Jul
Aug
Date
C
Deaths per 10,000
0
0.0–0.5
0.5–1.0
1.0–1.5
1.5–2.0
2.0–2.5
>2.5
Mar
Apr
May
Jun
Jul
Aug
Date
D
$R^2 = 0.72$, $p < 0.01$
COVID attributed deaths per 10,000
30
24
18
12
6
20
40
60
80
Socioeconomic status
E
Full lockdown
−10%
−30%
−50%
−70%
Mobility changes compared to baseline
Apr
May
Jun
Jul
Aug
Sep
Date
F
$R^2 = 0.46$, $p < 0.01$
−30%
−40%
−50%
−60%
Average mobility change during full lockdown
Santiago
Providencia
20
40
60
80
Socioeconomic status

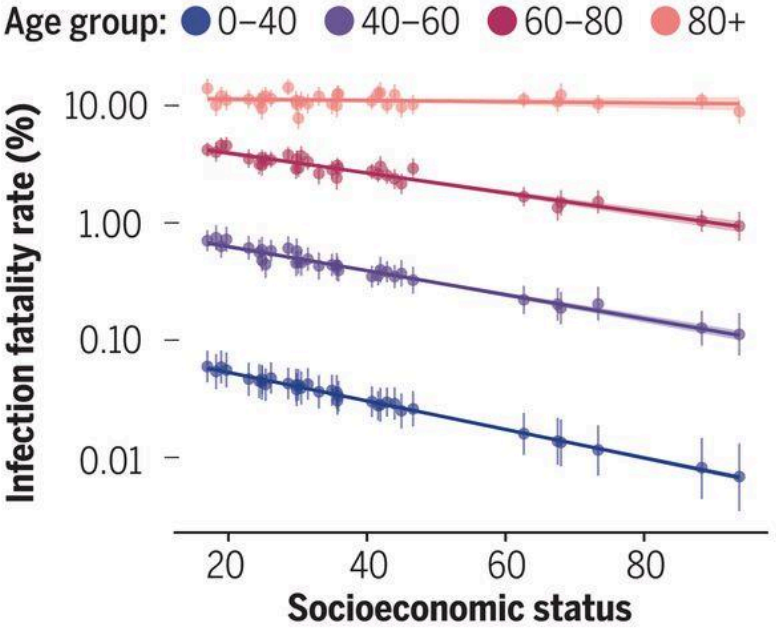
Municipalities of the Greater Santiago area of Chile

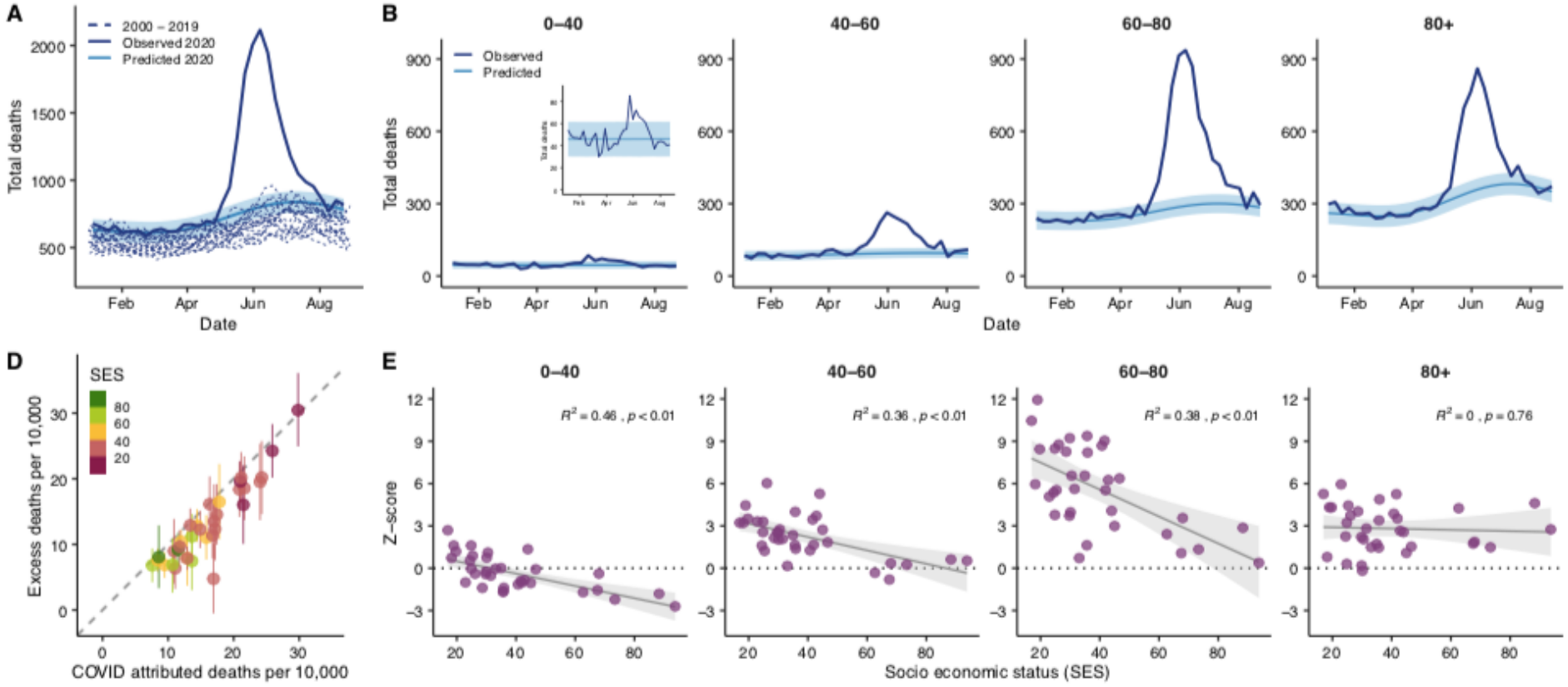


Comparison of COVID-19 deaths with excess deaths for the Greater Santiago area



Inferred infection fatality rate by age and socioeconomic status




A
2000 – 2019
Observed 2020
Predicted 2020
Total deaths
2000
1500
1000
500
Feb
Apr
Jun
Aug
Date
B
0–40
40–60
60–80
80+
Observed
Predicted
Total deaths
900
600
300
0
Feb
Apr
Jun
Aug
Date
D
SES
80
60
40
20
Excess deaths per 10,000
30
20
10
0
0
10
20
30
COVID attributed deaths per 10,000
E
0–40
40–60
60–80
80+
$R^2 = 0.46$, $p < 0.01$
$R^2 = 0.36$, $p < 0.01$
$R^2 = 0.38$, $p < 0.01$
$R^2 = 0$, $p = 0.76$
Z–score
12
9
6
3
0
−3
20
40
60
80
Socio economic status (SES)

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

The evolving roles of US political partisanship and social vulnerability in the COVID-19 pandemic from February 2020–February 2021

Justin Kaashoek, Christian Testa, Jarvis T. Chen, Lucas M. Stolerman, Nancy Krieger, William P. Hanage, Mauricio Santillana

Published: December 5, 2022 • https://doi.org/10.1371/journal.pgph.0000557

Abstract

The COVID-19 pandemic has had intense, heterogeneous impacts on different communities and geographies in the United States. We explore county-level associations between COVID-19 attributed deaths and social, demographic, vulnerability, and political variables to develop a better understanding of the evolving roles these variables have played in relation to mortality. We focus on the role of political variables, as captured by support for either the Republican or Democratic presidential candidates in the 2020 elections and the stringency of state-wide governor mandates, during three non-overlapping time periods between February 2020 and February 2021. We find that during the first three months of the pandemic, Democratic-leaning and internationally-connected urban counties were affected. During subsequent months (between May and September 2020), Republican counties with high percentages of Hispanic and Black populations were most hardly hit. In the third time period –between October 2020 and February 2021– we find that Republican-leaning counties with loose mask mandates experienced up to 3 times higher death rates than Democratic-leaning counties, even after controlling for multiple social vulnerability factors. Some of these deaths could perhaps have been avoided given that the effectiveness of non-pharmaceutical interventions in preventing uncontrolled disease transmission, such as social distancing and wearing masks indoors, had been well-established at this point in time.

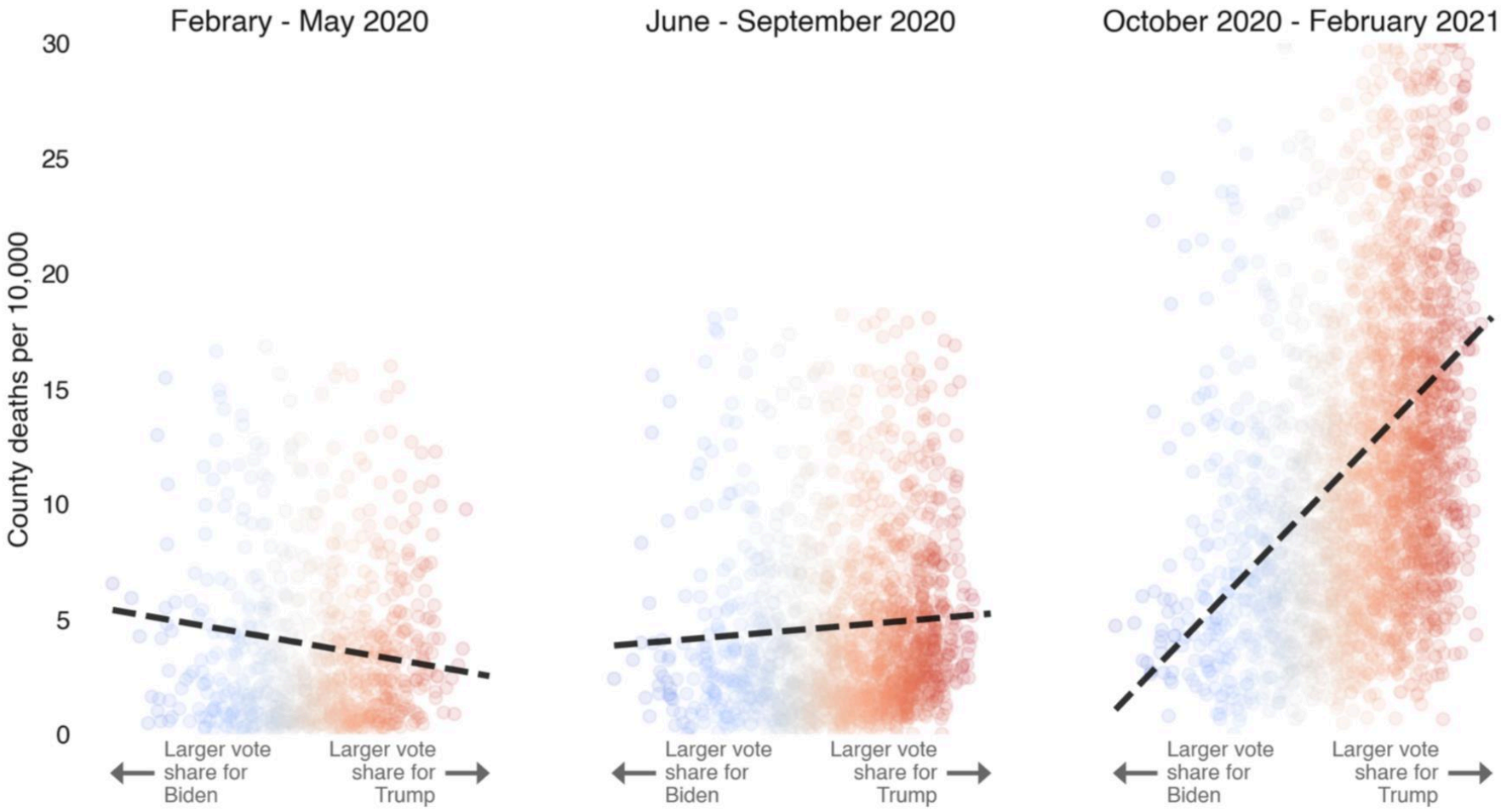
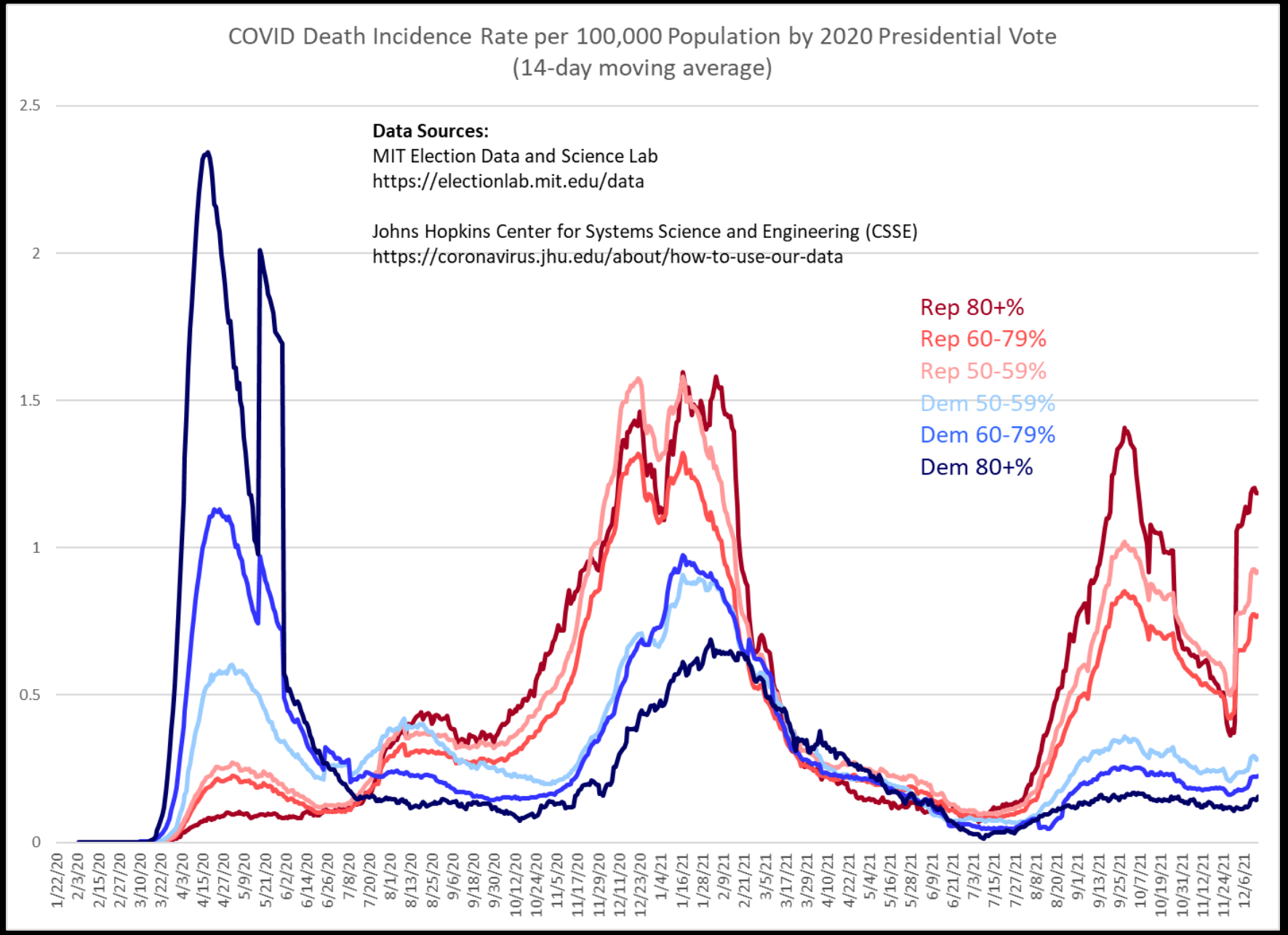


Figure 2: Evolution of COVID-19 deaths vs Political leaning. COVID-19 attributed deaths (per 10,000) at the county level as a function of vote share in favor of J. Biden (Democratic) vs D.J. Trump (Republican), the 2020 presidential candidates, during the three time periods of interest. Inspiration for this figure comes from a David Leonhardt's New York Times article, "Red COVID" [19].



Not from our team with updates. Credit @maolesen

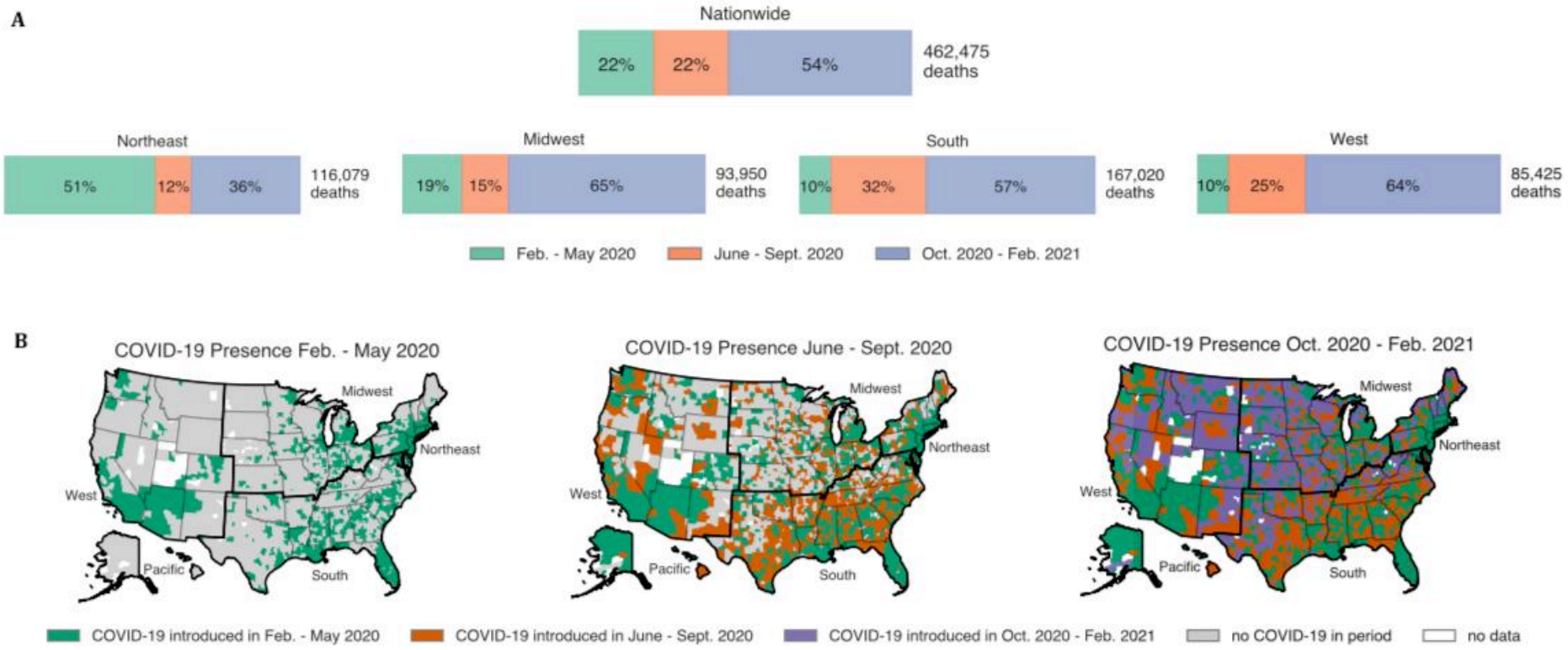
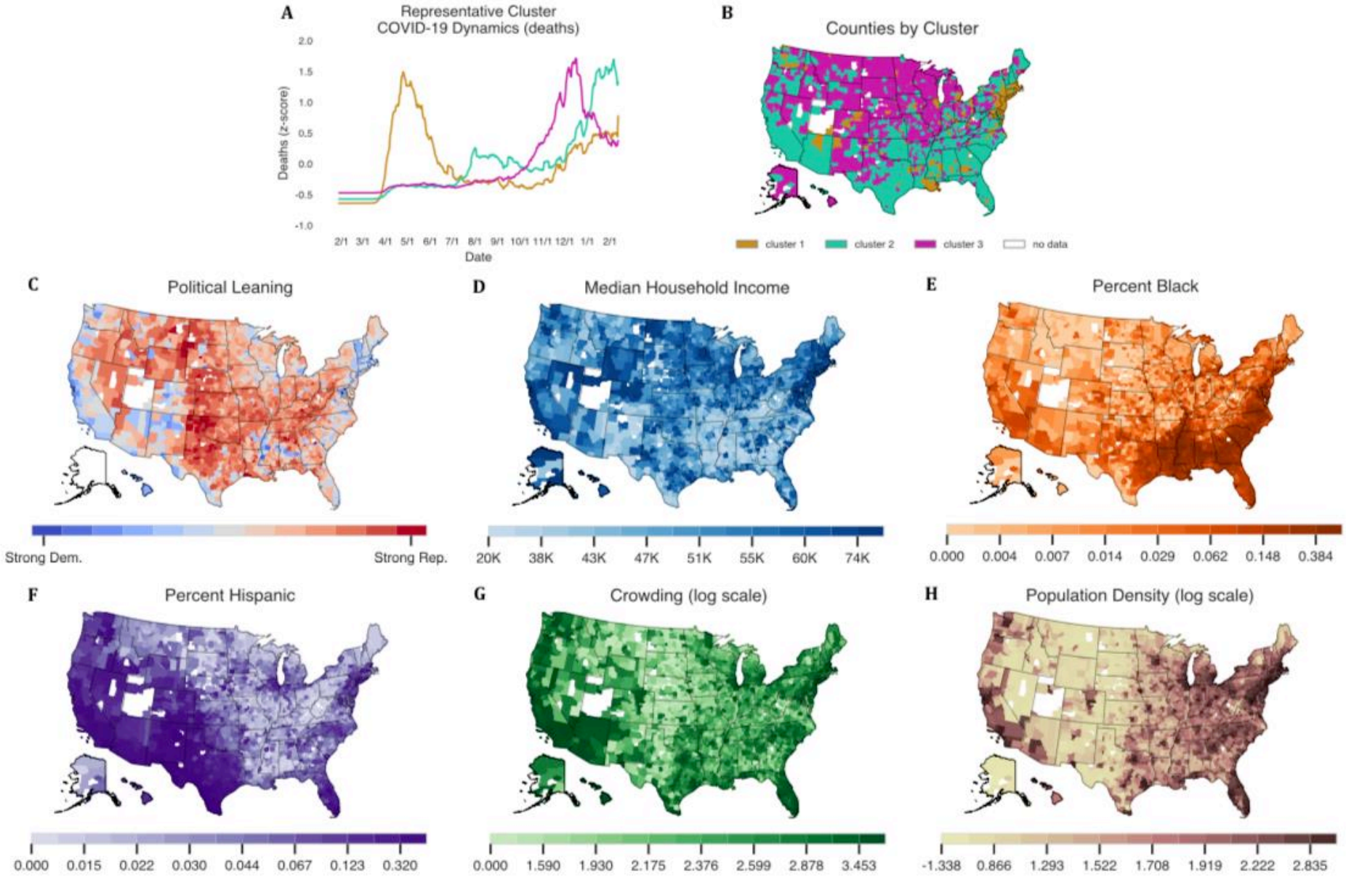
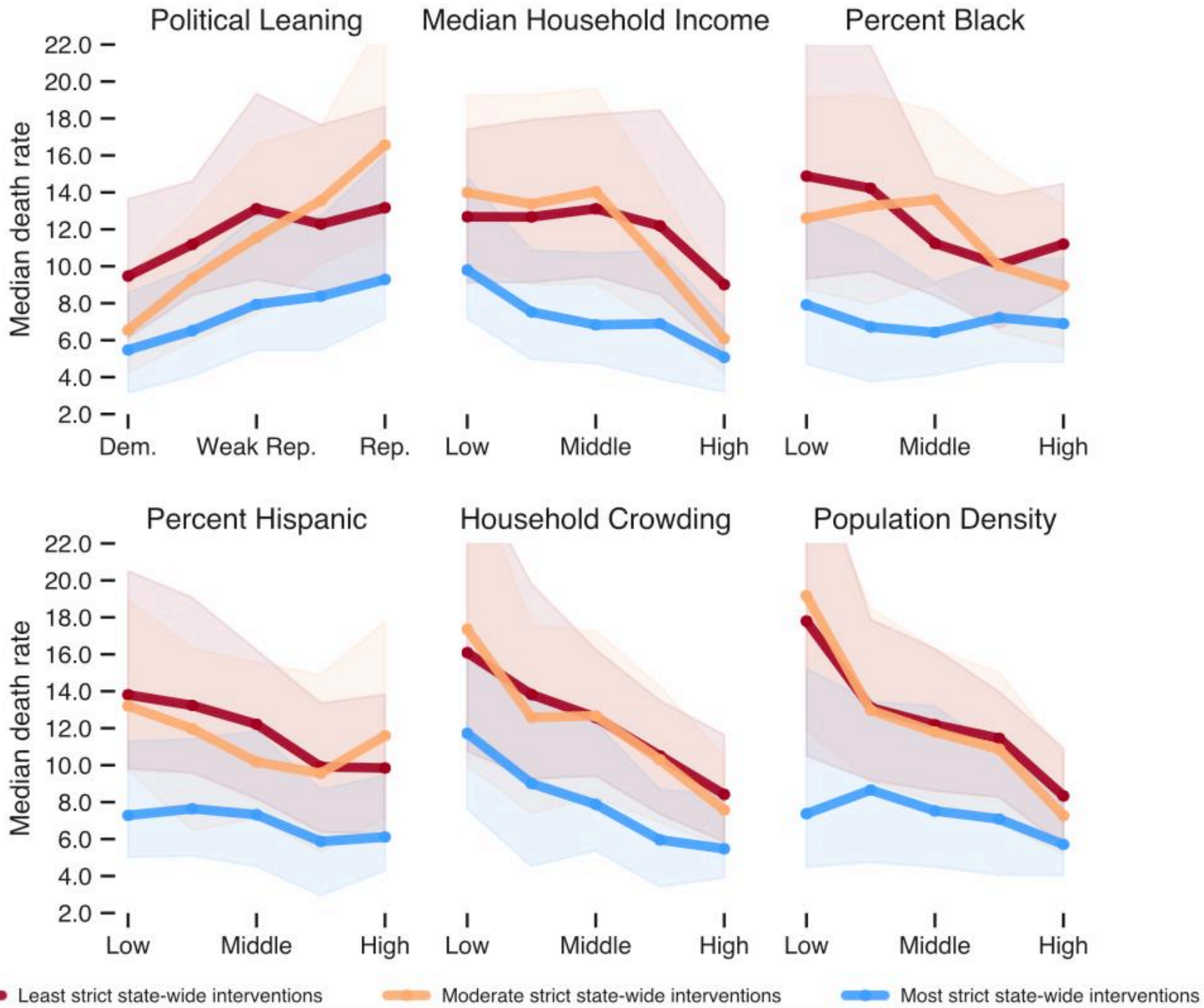
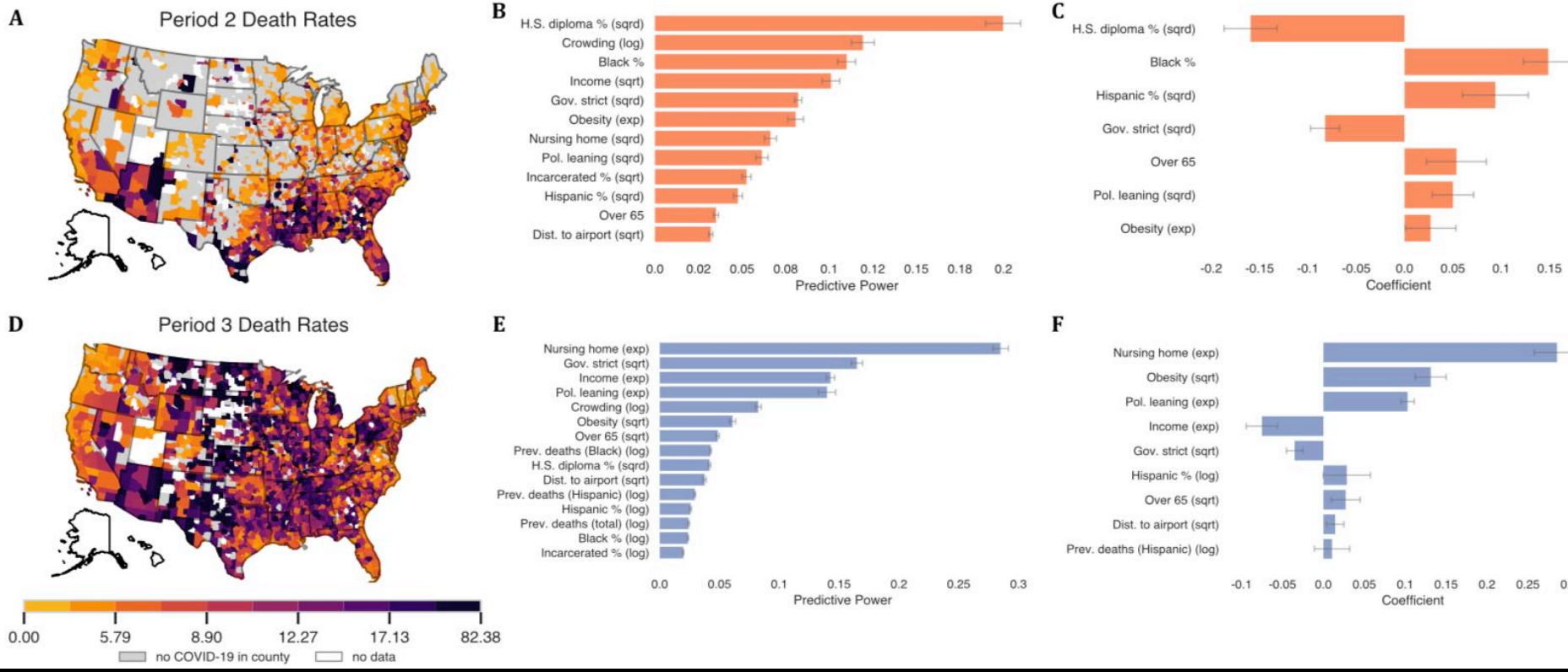


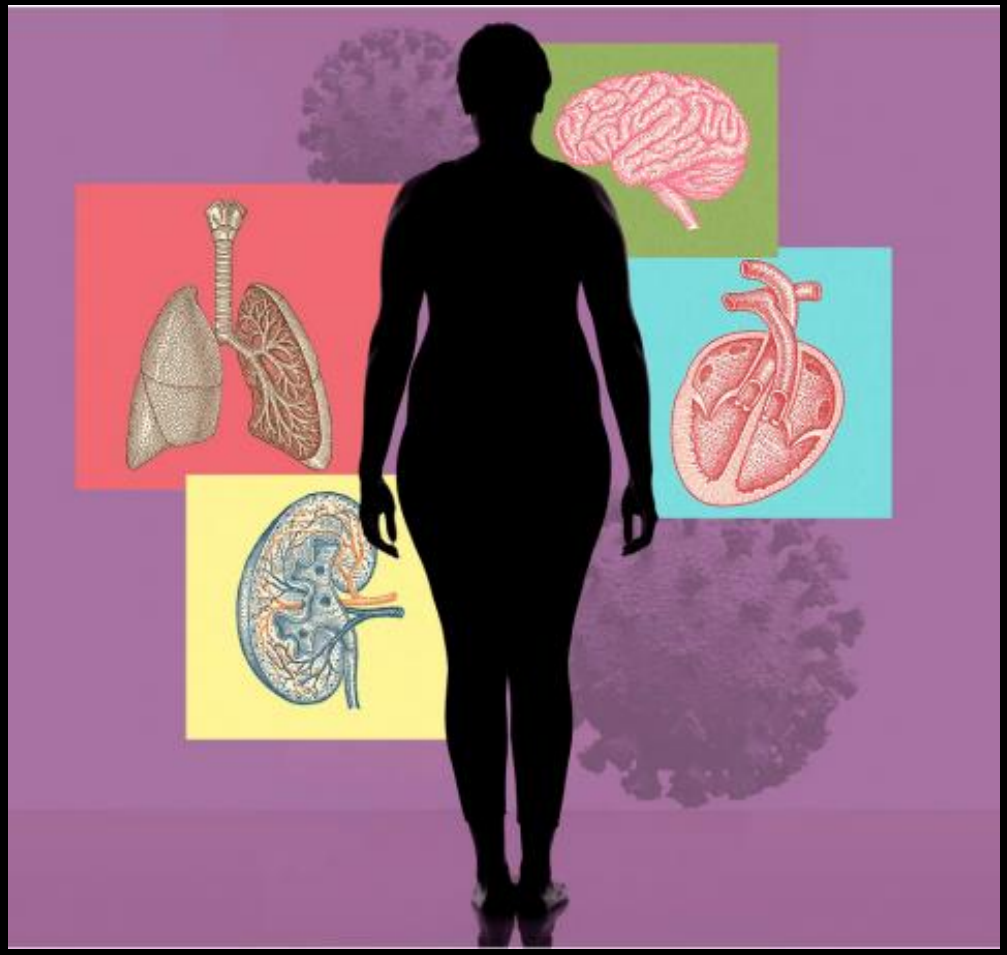
Figure 3: A breakdown of COVID-19 presence across the time periods of interest. (**A**) The percentage of deaths in the nation by time period, both nationwide and by Census region. Other than the Northeast, which was hit hard in the first period, the nation was hit hardest in period 3, as pointed out in [20]. (**B**) COVID-19 onset at the county-level. A county is treated as infected once it has experienced at least 5 COVID-related deaths. We see the movement of COVID from the cities and coastal areas to the center of the county over the course of the year.


A
Representative Cluster
COVID-19 Dynamics (deaths)
Deaths (z-score)
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
2/1 3/1 4/1 5/1 6/1 7/1 8/1 9/1 10/1 11/1 12/1 1/1 2/1
Date
B
Counties by Cluster
cluster 1
cluster 2
cluster 3
no data
C
Political Leaning
Strong Dem.
Strong Rep.
D
Median Household Income
20K 38K 43K 47K 51K 55K 60K 74K
E
Percent Black
0.000 0.004 0.007 0.014 0.029 0.062 0.148 0.384
F
Percent Hispanic
0.000 0.015 0.022 0.030 0.044 0.067 0.123 0.320
G
Crowding (log scale)
0.000 1.590 1.930 2.175 2.376 2.599 2.878 3.453
H
Population Density (log scale)
-1.338 0.866 1.293 1.522 1.708 1.919 2.222 2.835


Political Leaning
Median Household Income
Percent Black
Percent Hispanic
Household Crowding
Population Density
Median death rate
22.0
20.0
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
Dem.
Weak Rep.
Rep.
Low
Middle
High
Least strict state-wide interventions
Moderate strict state-wide interventions
Most strict state-wide interventions


A
Period 2 Death Rates
B
H.S. diploma % (sqrd)
Crowding (log)
Black %
Income (sqrt)
Gov. strict (sqrd)
Obesity (exp)
Nursing home (sqrd)
Pol. leaning (sqrd)
Incarcerated % (sqrt)
Hispanic % (sqrd)
Over 65
Dist. to airport (sqrt)
0.0 0.02 0.05 0.08 0.1 0.12 0.15 0.18 0.2
Predictive Power
C
H.S. diploma % (sqrd)
Black %
Hispanic % (sqrd)
Gov. strict (sqrd)
Over 65
Pol. leaning (sqrd)
Obesity (exp)
-0.2 -0.15 -0.1 -0.05 0.0 0.05 0.1 0.15
Coefficient
D
Period 3 Death Rates
0.00 5.79 8.90 12.27 17.13 82.38
no COVID-19 in county
no data
E
Nursing home (exp)
Gov. strict (sqrt)
Income (exp)
Pol. leaning (exp)
Crowding (log)
Obesity (sqrt)
Over 65 (sqrt)
Prev. deaths (Black) (log)
H.S. diploma % (sqrd)
Dist. to airport (sqrt)
Prev. deaths (Hispanic) (log)
Hispanic % (log)
Prev. deaths (total) (log)
Black % (log)
Incarcerated % (log)
0.0 0.05 0.1 0.15 0.2 0.25 0.3
Predictive Power
F
Nursing home (exp)
Obesity (sqrt)
Pol. leaning (exp)
Income (exp)
Gov. strict (sqrt)
Hispanic % (log)
Over 65 (sqrt)
Dist. to airport (sqrt)
Prev. deaths (Hispanic) (log)
-0.1 -0.05 0.0 0.05 0.1 0.15 0.2 0.25 0.3
Coefficient

Comorbidities



CANCER DISCOVERY

Volume 10, Issue 6
1 June 2020

Article Contents

RESEARCH BRIEF | AUTHOR CHOICE | JUNE 01 2020

Patients with Cancer Appear More Vulnerable to SARS-CoV-2: A Multicenter Study during the COVID-19 Outbreak FREE

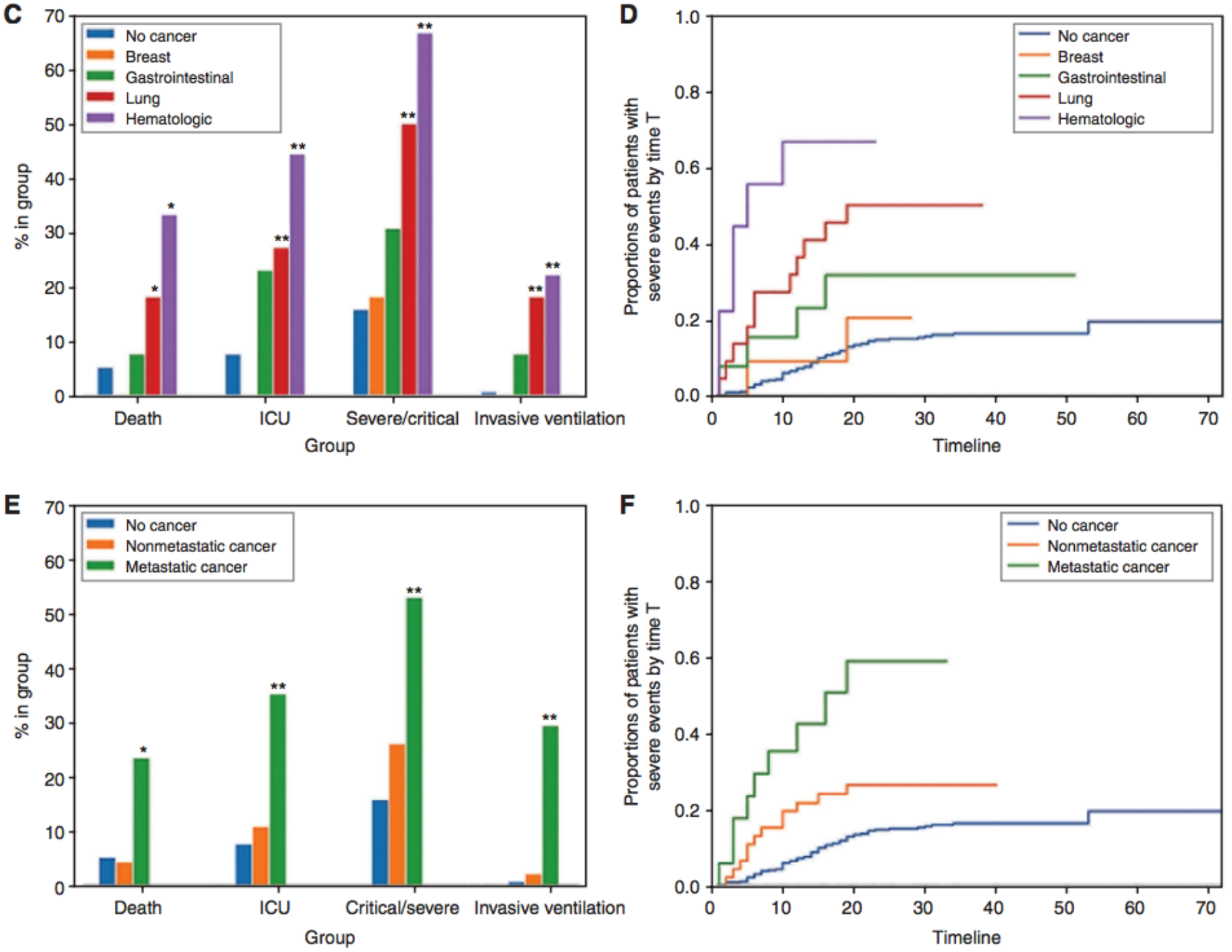
Mengyuan Dai; Dianbo Liu; Miao Liu ✉; Fuxiang Zhou; Guiling Li; Zhen Chen; Zhian Zhang; Hua You; Meng Wu; Qichao Zheng; Yong Xiong ; Huihua Xiong ; Chun Wang; Changchun Chen; Fei Xiong; Yan Zhang; Yaqin Peng; Siping Ge; Bo Zhen; Tingting Yu; Ling Wang; Hua Wang; Yu Liu; Yeshan Chen; Junhua Mei; Xiaojia Gao; Zhuyan Li; Lijuan Gan; Can He; Zhen Li; Yuying Shi; Yuwen Qi; Jing Yang; Daniel G. Tenen ; Li Chai; Lorelei A. Mucci; Mauricio Santillana ✉ ; Hongbing Cai ✉

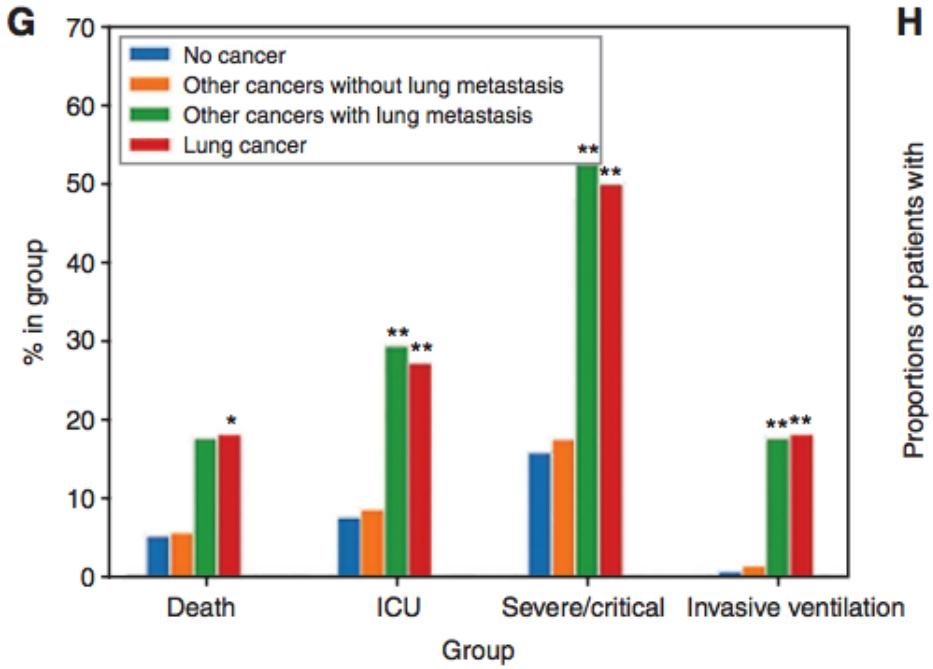
Check for updates

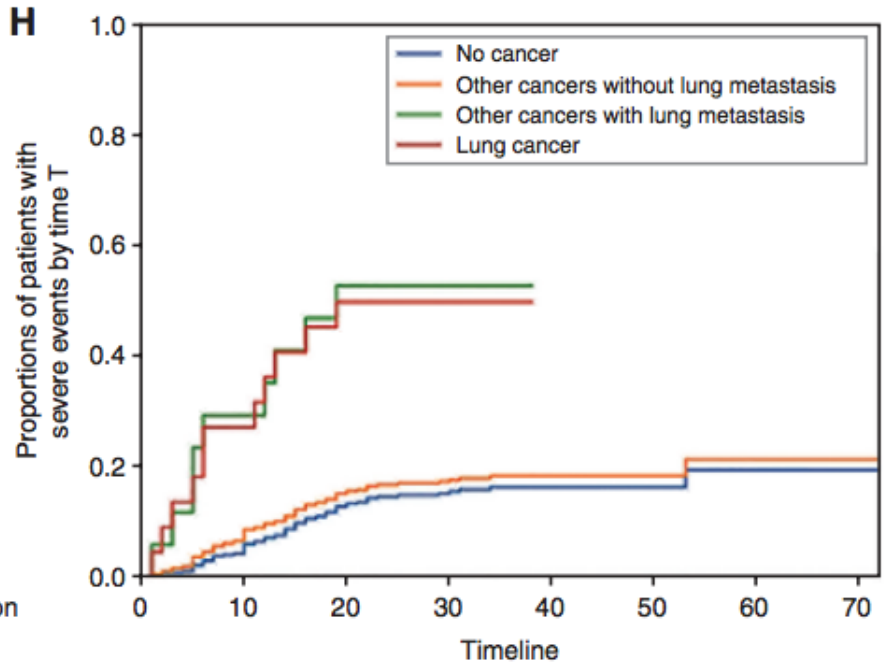
+ Author & Article Information

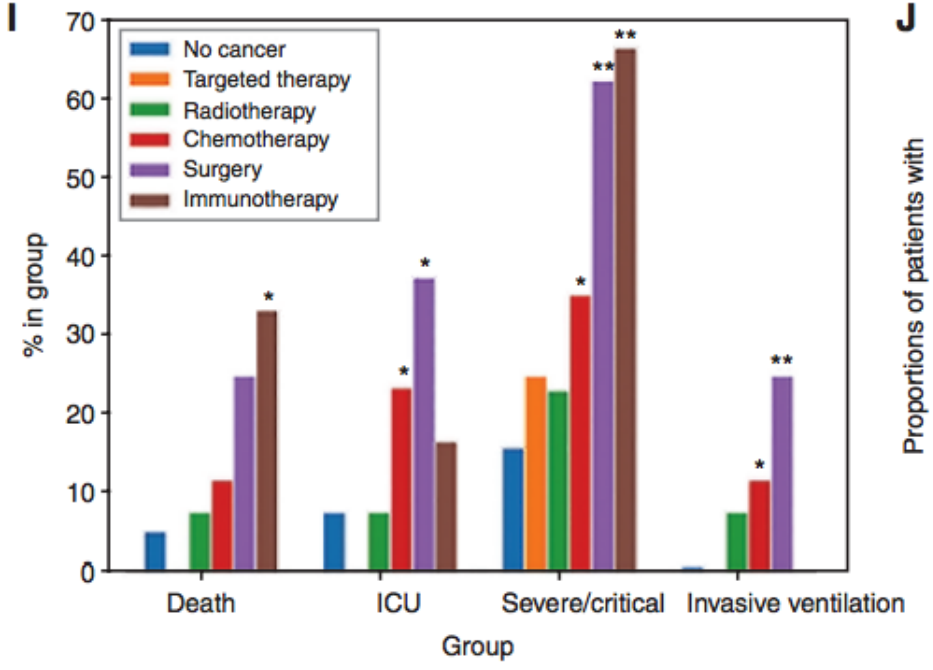
Cancer Discov (2020) 10 (6): 783–791.

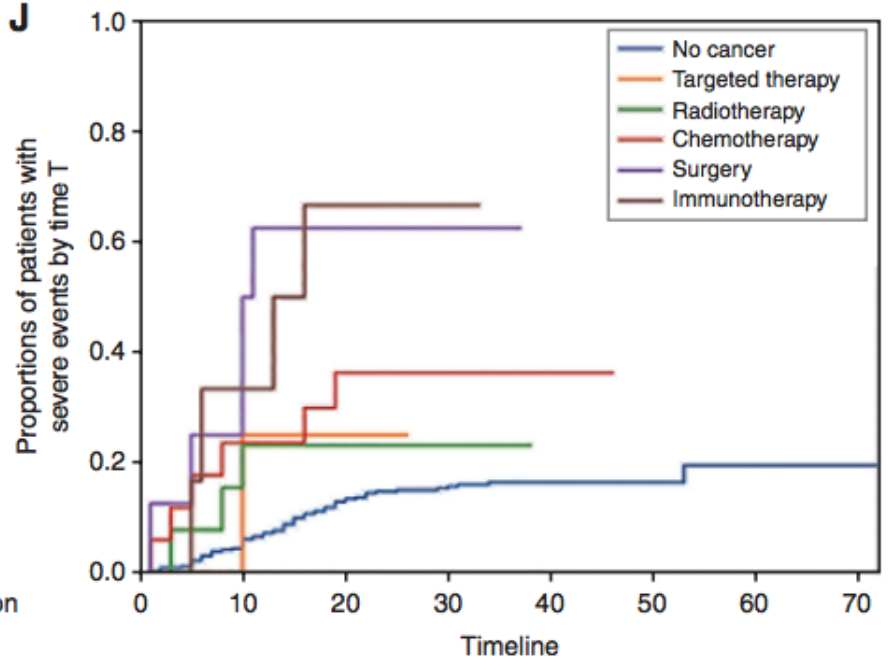
https://doi.org/10.1158/2159-8290.CD-20-0422 **Article history**

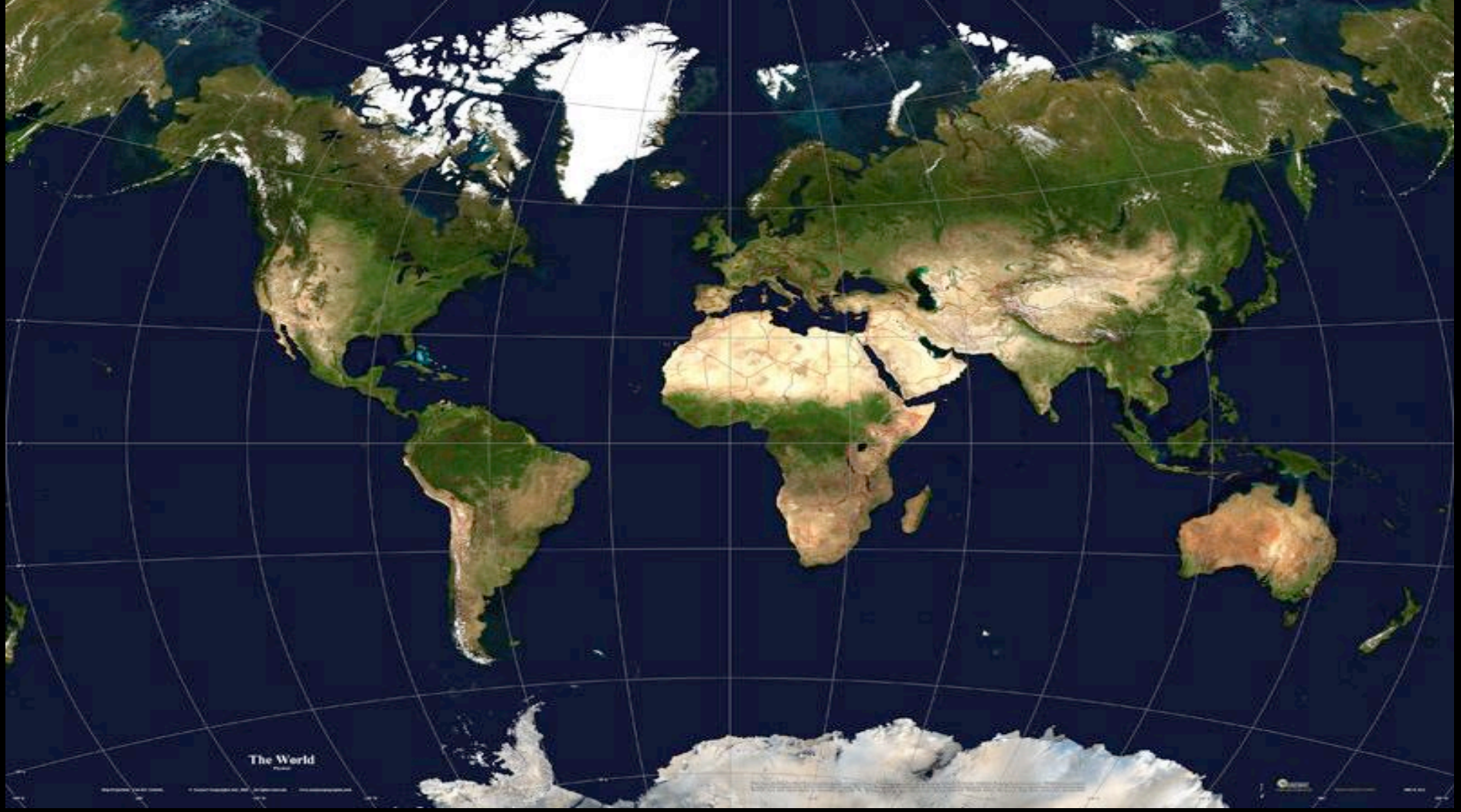

C
No cancer
Breast
Gastrointestinal
Lung
Hematologic
% in group
Death
ICU
Severe/critical
Invasive ventilation
Group
D
Proportions of patients with severe events by time T
Timeline
E
No cancer
Nonmetastatic cancer
Metastatic cancer
% in group
Death
ICU
Critical/severe
Invasive ventilation
Group
F
Proportions of patients with severe events by time T
Timeline


G
No cancer
Other cancers without lung metastasis
Other cancers with lung metastasis
Lung cancer
% in group
Death
ICU
Severe/critical
Invasive ventilation
Group


H
No cancer
Other cancers without lung metastasis
Other cancers with lung metastasis
Lung cancer
Proportions of patients with severe events by time T
Timeline


I
No cancer
Targeted therapy
Radiotherapy
Chemotherapy
Surgery
Immunotherapy
% in group
Death
ICU
Severe/critical
Invasive ventilation
Group


J
No cancer
Targeted therapy
Radiotherapy
Chemotherapy
Surgery
Immunotherapy
Proportions of patients with severe events by time T
Timeline



Thank you!

Contact: m.santillana@northeastern.edu
msantill@g.harvard.edu