Characterizing the dynamics of COVID-19 transmission using multiple mathematical approaches and novel data sources

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

Mauricio Santillana, PhD
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, Network Science Institute, Northeastern University

Northeastern University

netsi Northeastern University *Network Science Institute*

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Water availability

Global air quality modeling

Natural disasters preparedness

The main motivation behind my research has been to provide ***decision-makers*** with ***actionable*** information with the hope of ***saving/improving human lives*** using ***mathematical models***

Intensive Care Unite Predictive

Epidemic outbreaks mitigation

Characterizing the dynamics of COVID-19 transmission using multiple mathematical approaches and novel data sources

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

Mauricio Santillana, PhD
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 Lab, Network Science Institute, Northeastern University
Associate faculty, Harvard Institute of Applied Computational Sciences

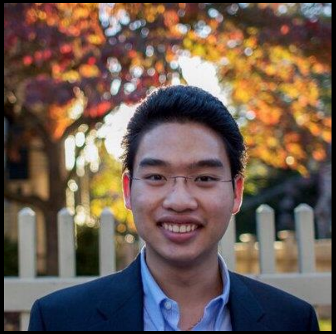
Northeastern University

netsi Northeastern University Network Science Institute

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Mauricio Santillana, Ph.D
Director,
Machine Intelligence Lab, Boston Children's Hospital;

Assistant Professor,
Harvard Medical School
Harvard T.H. Chan School of Public Health



Fred Lu, M.S.
Team member,
MIGHTE, Northeastern University
Graduate Student,
University of Maryland

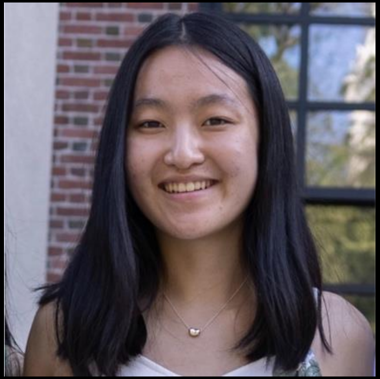
Andre Nguyen, Ph.D.
Team member,
MIGHTE, Northeastern University

Leonardo Clemente, M.S.
Team member,
MIGHTE, Northeastern University
Graduate Student,
Tec de Monterrey, Mexico

Canelle Poirier, PhD
Team member,
MIGHTE, Northeastern University
Post-doctoral Fellow

Anjalika Nande, Ph.D.
Team member,
MIGHTE, Northeastern University
Post-doctoral Fellow
Johns Hopkins University

Nicole Kogan
Team member,
MIGHTE, Northeastern University
Graduate Student,
Harvard T.H Chan School of Public Health



Iris Lang
Team member,
MIGHTE, Northeastern University
Undergraduate Student,
Harvard University

Lucas Stolerman, PhD
Team member,
MIGHTE, Northeastern University
Assistant Professor, Oklahoma State U

Austin Meyer, Ph.D. M.D.
Team member,
MIGHTE, Northeastern University
Medical team, Austin TX

Jinho Park,
Team member,
MIGHTE, Northeastern University
Undergraduate Student
Mathematics, Harvard University

Leonardo Cazares
Team member,
MIGHTE, Northeastern University
Undergraduate Student
Physics, UNAM, Mexico

Xiyu Yang, M.S.
Team member,
MIGHTE, Northeastern University

Yash Bhora
Team member,
MIGHTE, Northeastern University
Undergraduate Student
Physics, Northeastern University

Raul Garrido,
Team member,
MIGHTE, Northeastern University
PhD Student
Physics, Northeastern University

Tamanna Urmi,
Team member,
MIGHTE, Northeastern University
PhD Student
Network Science, Northeastern University

Skyler Wu
Team member,
MIGHTE, Northeastern University
Undergraduate Student,
Applied Math and Statistics, Harvard

Pablo Martinez de Salazar, MD
Team member
MIGHTE, Northeastern University
Swiss Tropical and Public Health Institute
Post-doctoral Fellow

Team and collaborators in Epidemiological Forecasting: COVID-19


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

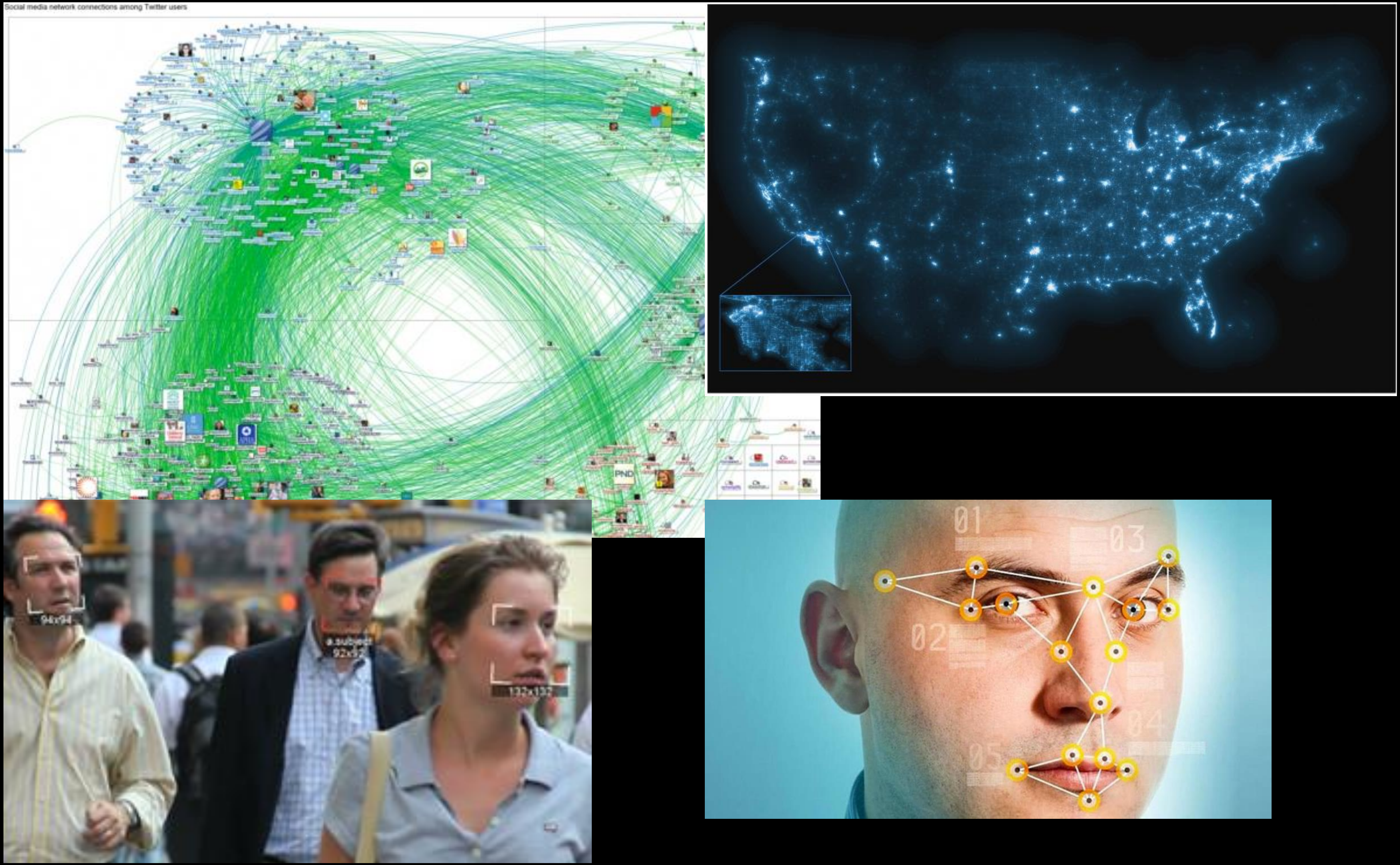

Pablo Martinez de Salazar


Mauricio Santillana

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

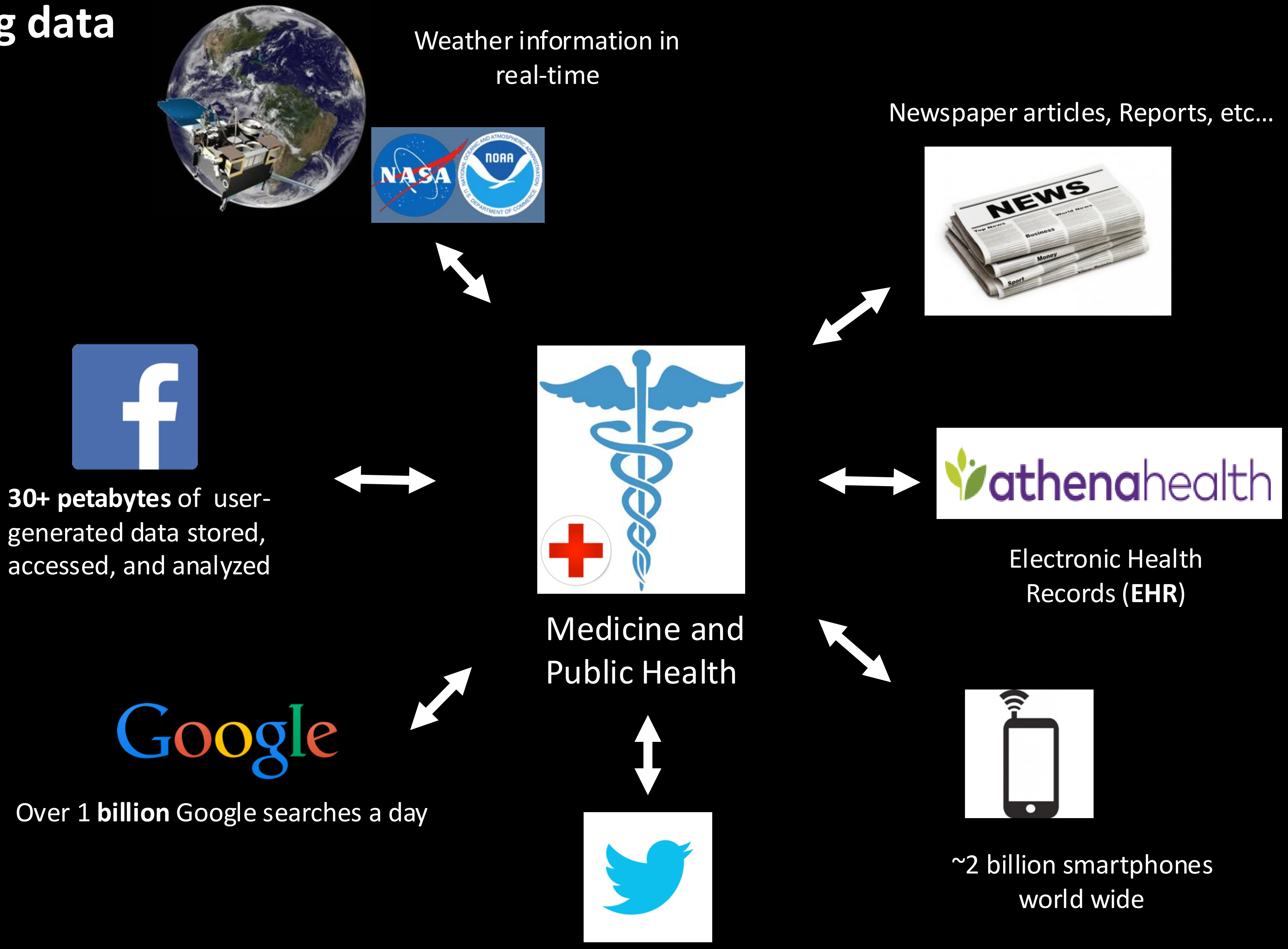
Over 1 **billion** Google searches a day

230 million tweets every day

~2 billion smartphones world wide

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**)

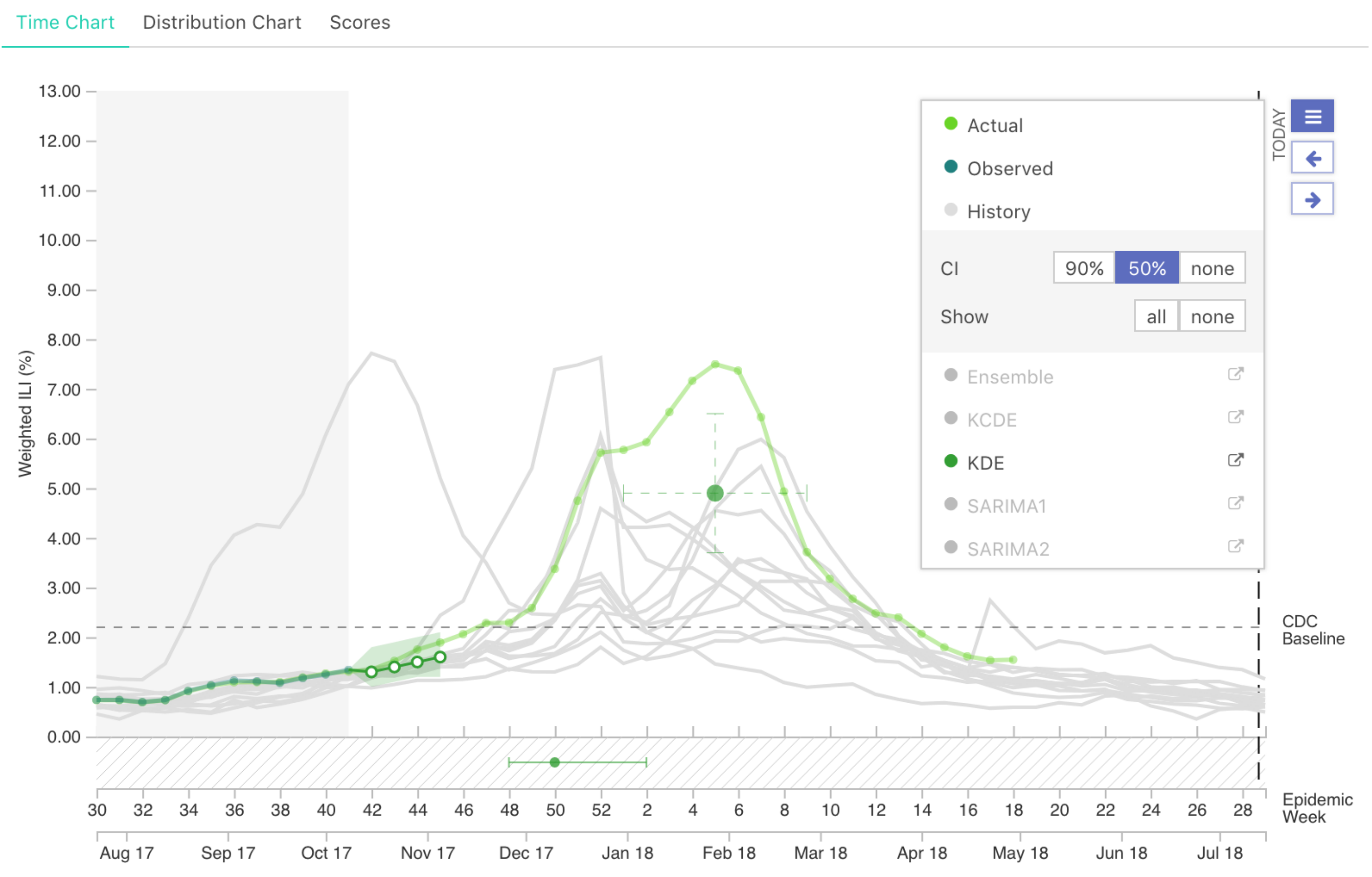
Over 1 **billion** Google searches a day

~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)



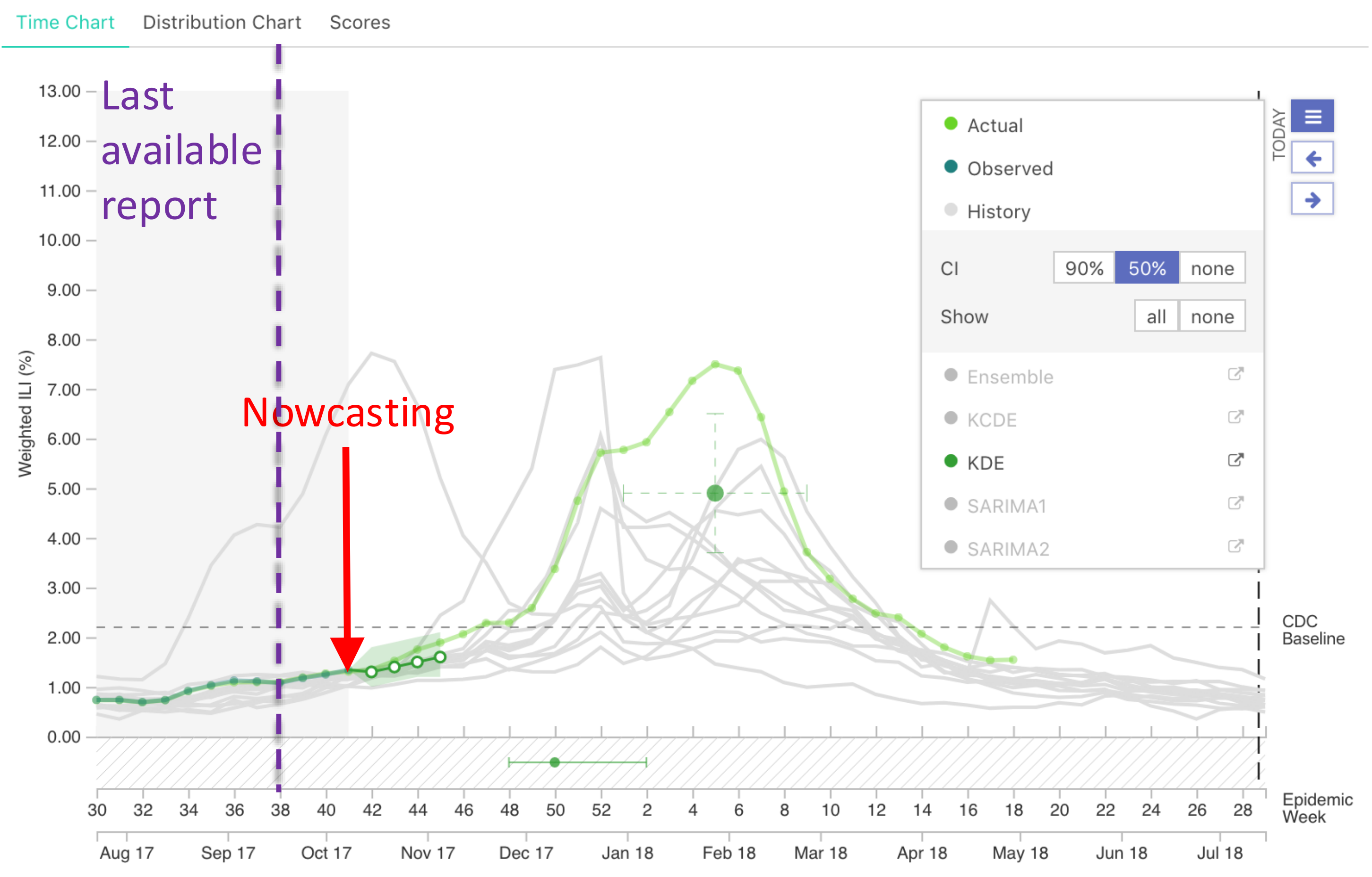
CI 90% [✓ 50%] none

Show all none

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

The challenge…

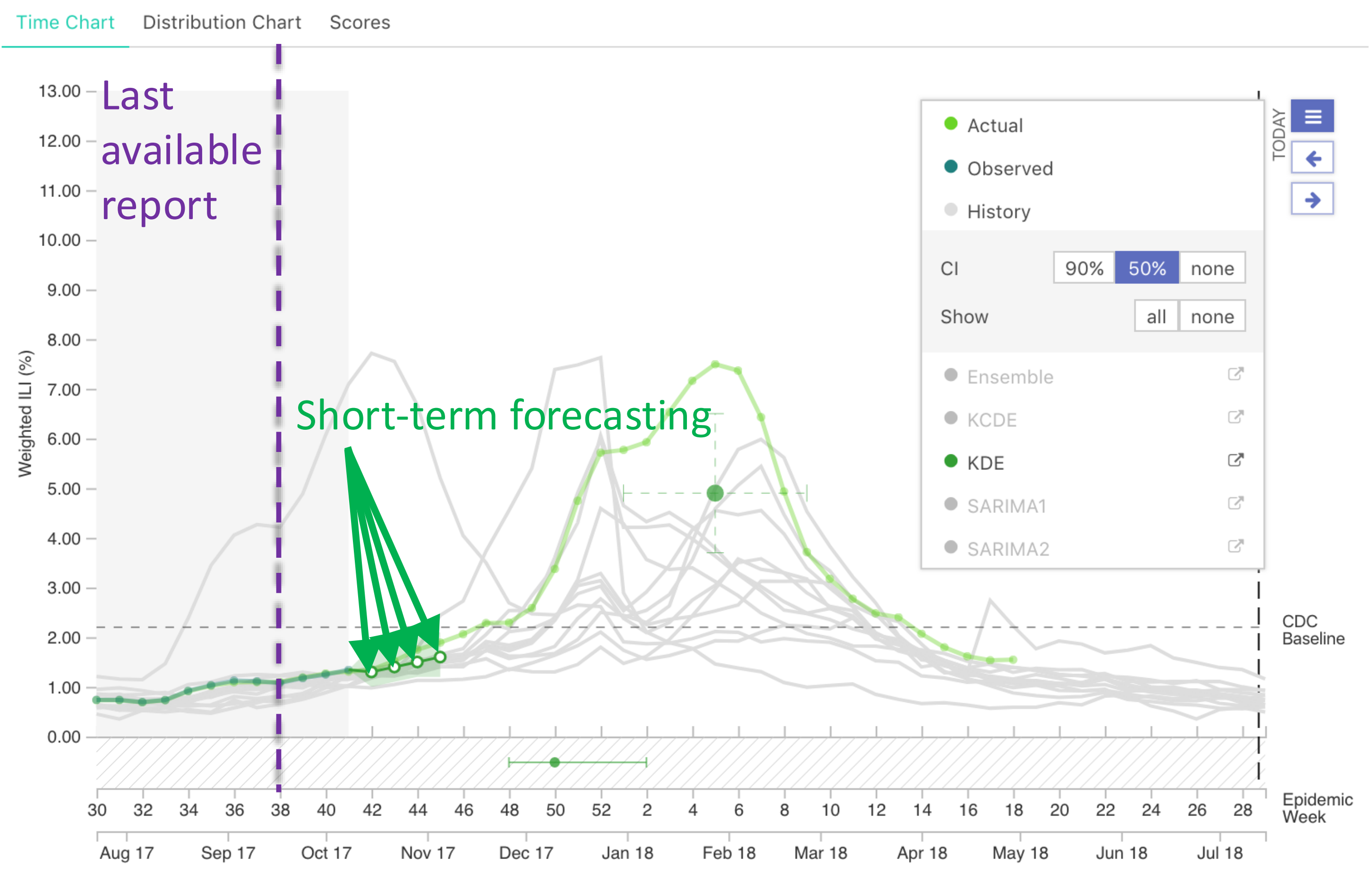
Real-time monitoring of disease activity, **short-term** forecasting (weeks), **long-term** forecasting (months)



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

The challenge…

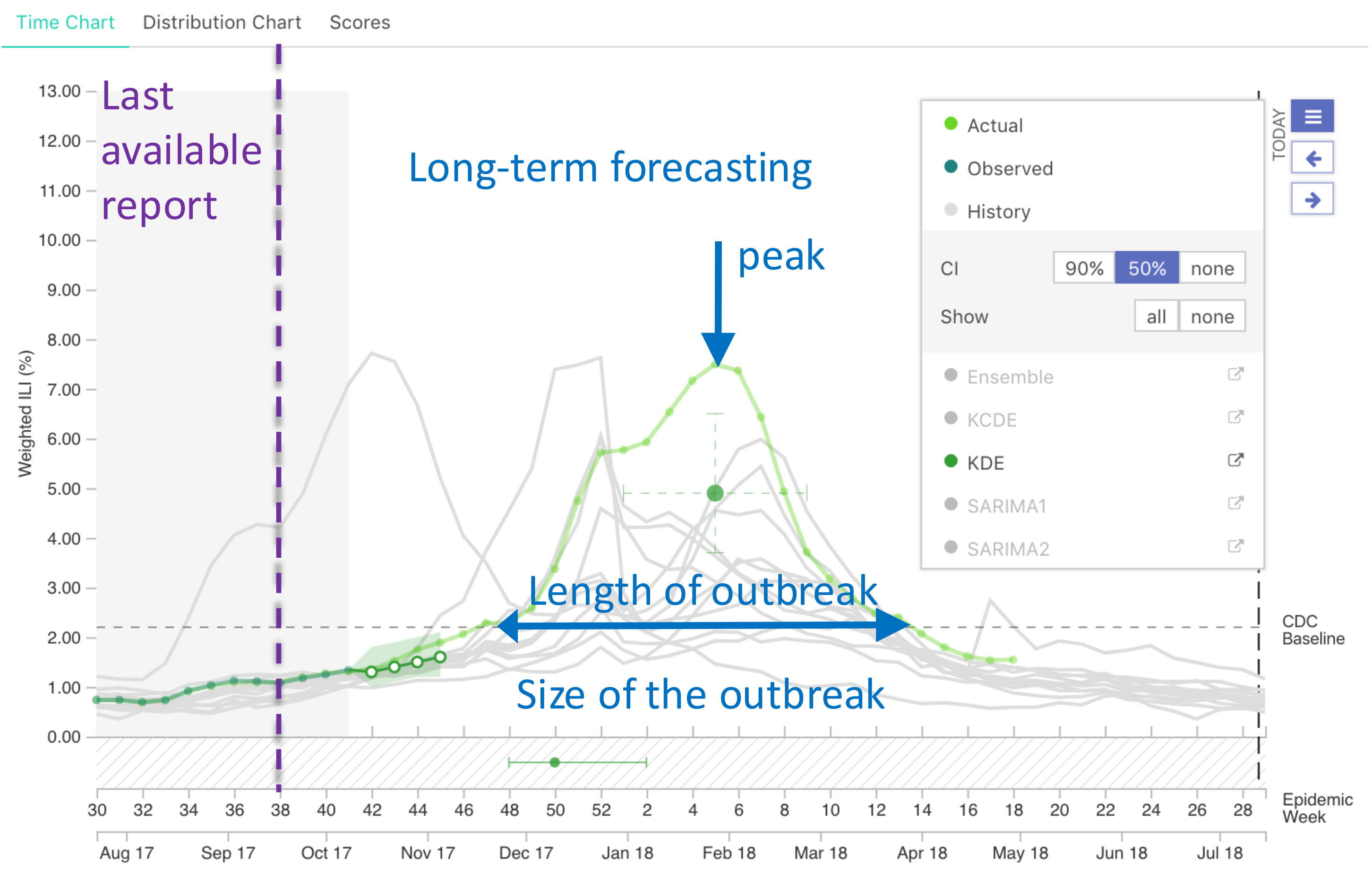
Real-time monitoring of disease activity, **short-term** forecasting (weeks), **long-term** forecasting (months)



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

The challenge…

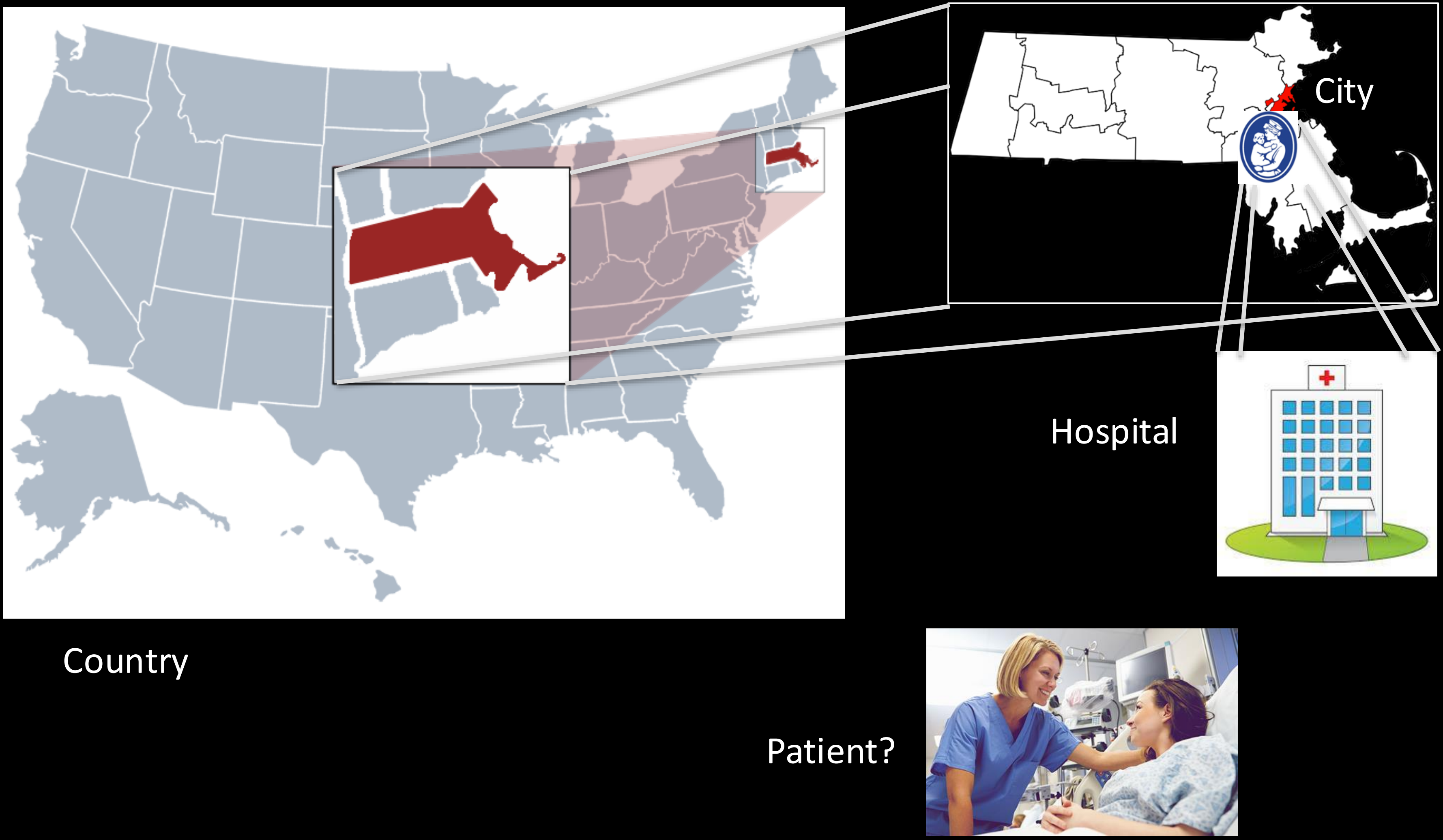
Real-time monitoring of disease activity, **short-term** forecasting (weeks), **long-term** forecasting (months)



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

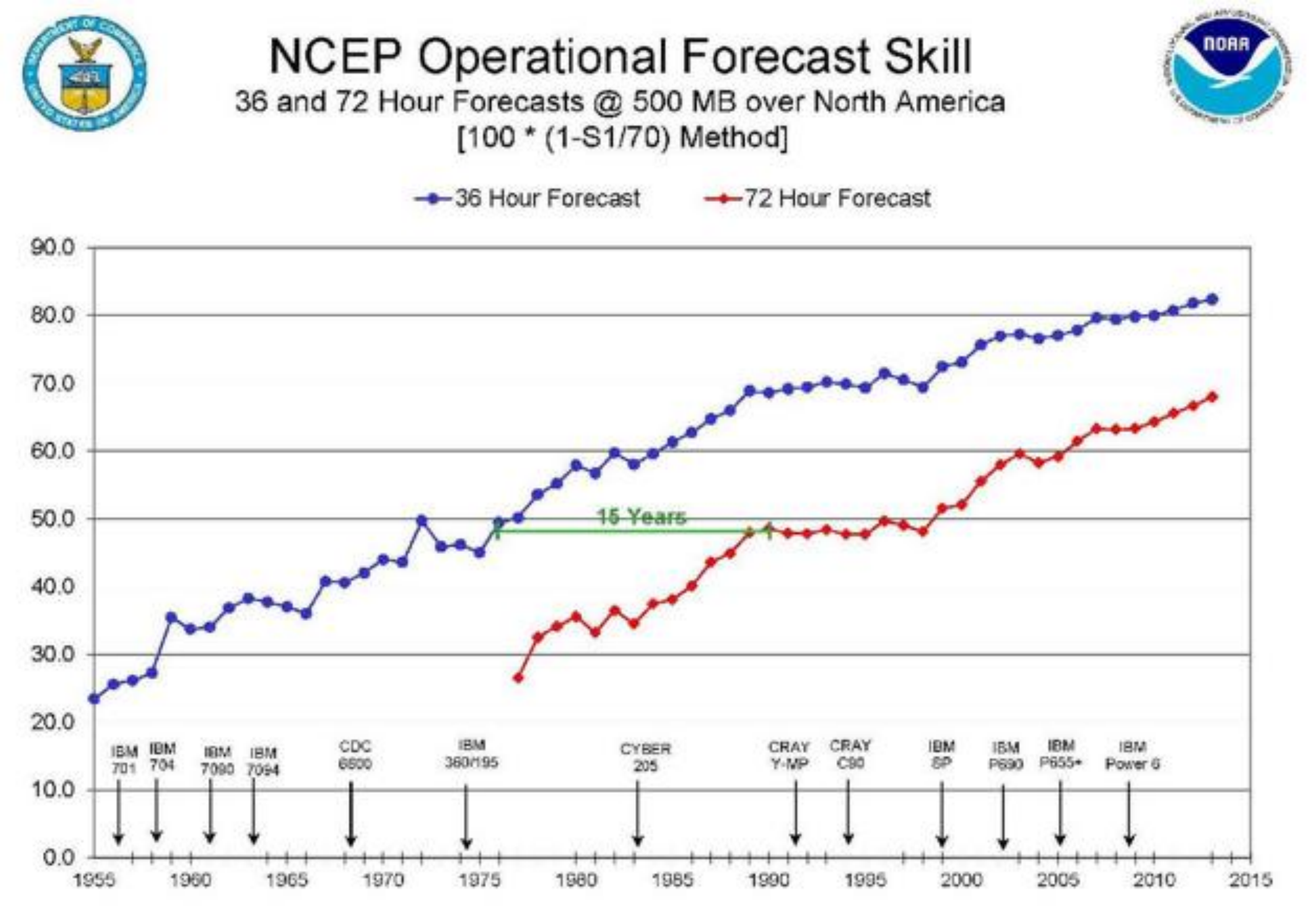
City

Hospital

Country

Patient?

What if we get it right?



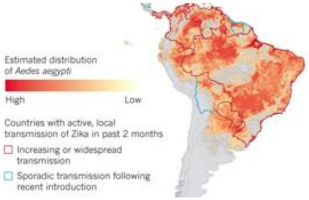
Real-time tracking vs predictions of disease incidence/risk
Similarities and differences with weather prediction

Approach

Data streams

Twitter and Wikipedia activity

Weather variables

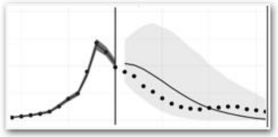


Mosquito prevalence

Human mobility

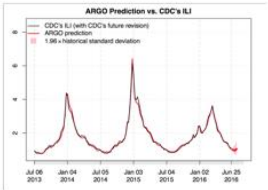
Territory

Modeling approaches



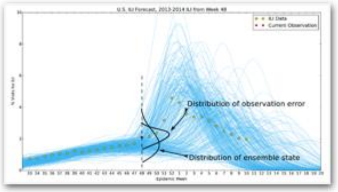
Mechanistic approaches

+



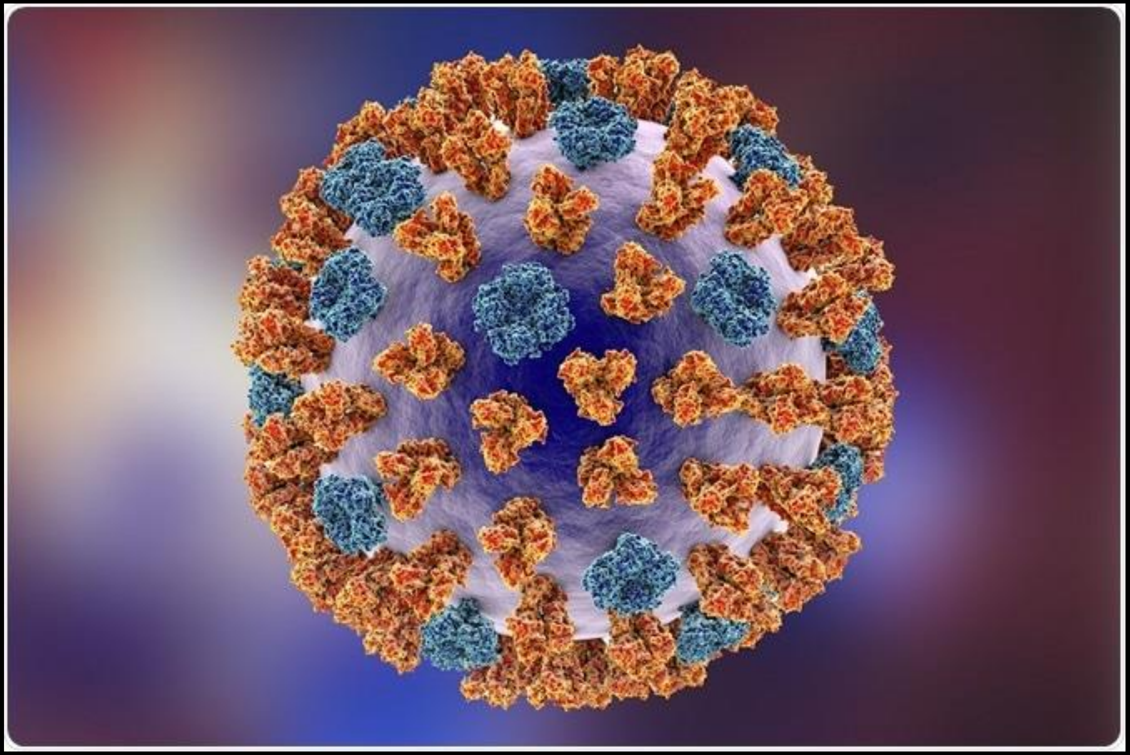
Machine-learning approaches

=



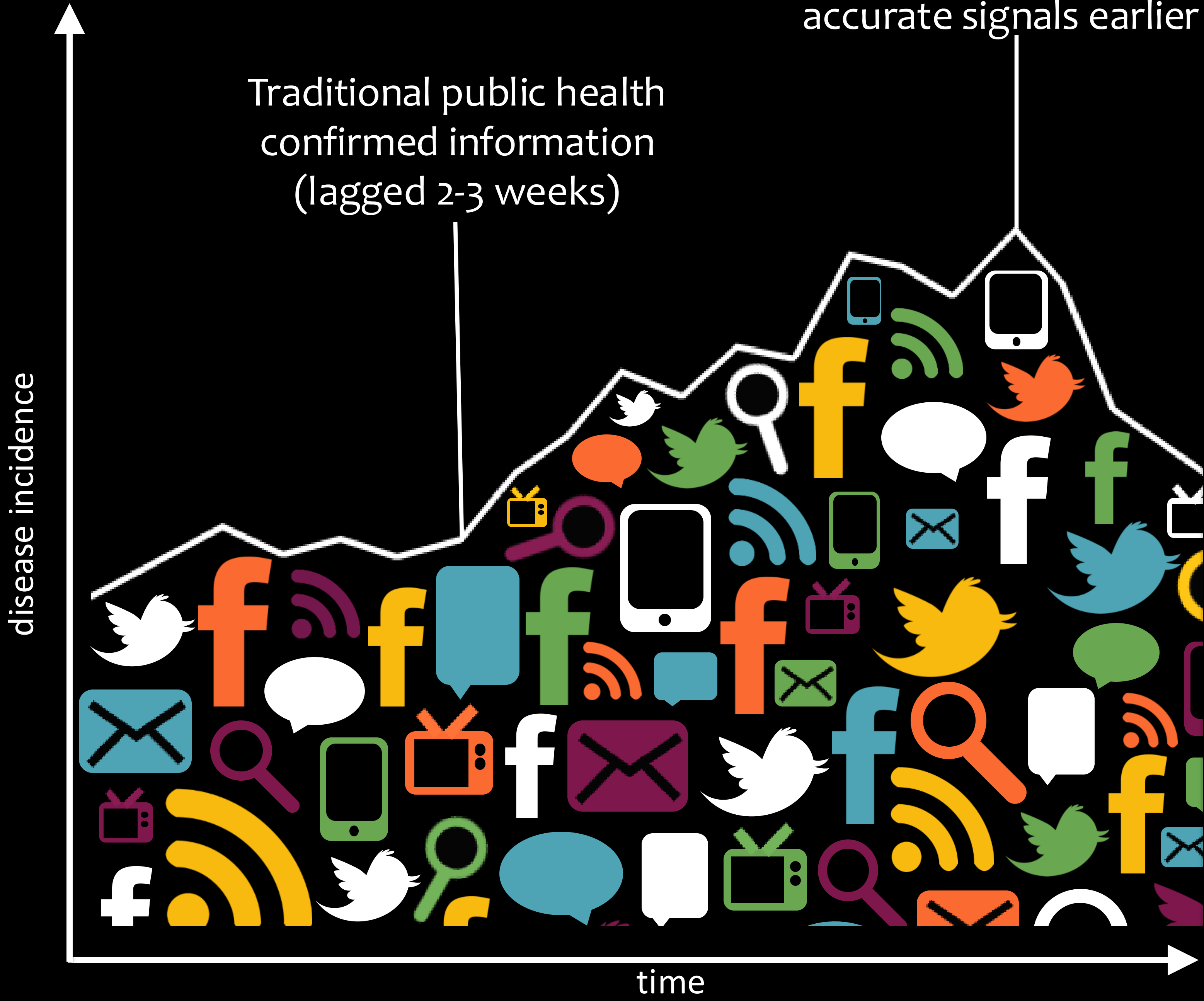
Ensemble forecasting approaches

Background: monitoring influenza in rich nations

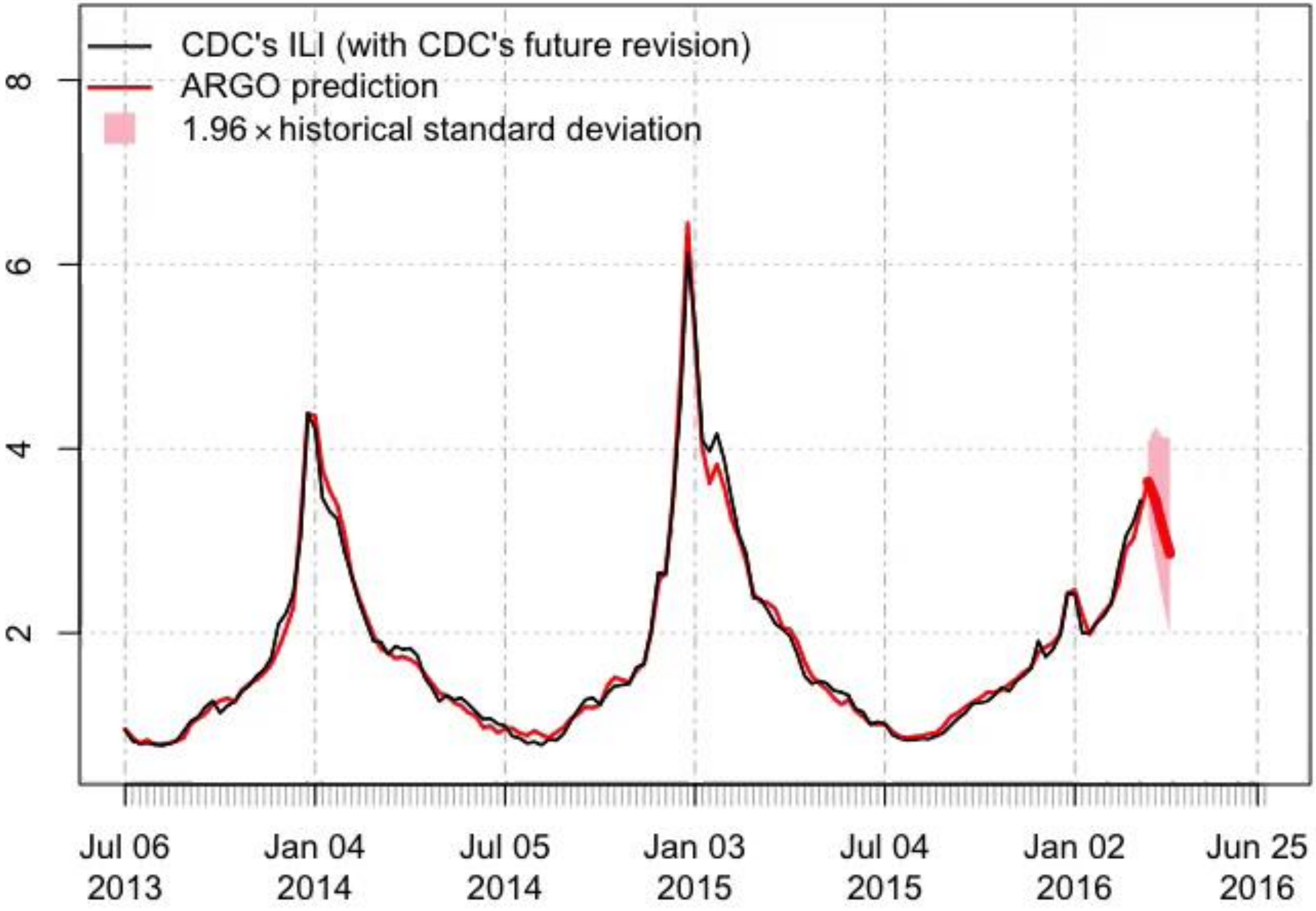


Can **Digital disease tracking** pick up accurate signals earlier ?

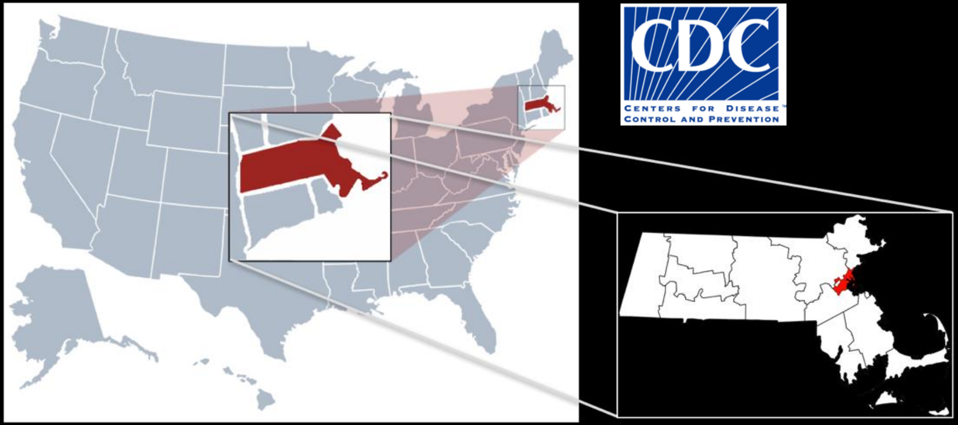
Traditional public health confirmed information (lagged 2-3 weeks)



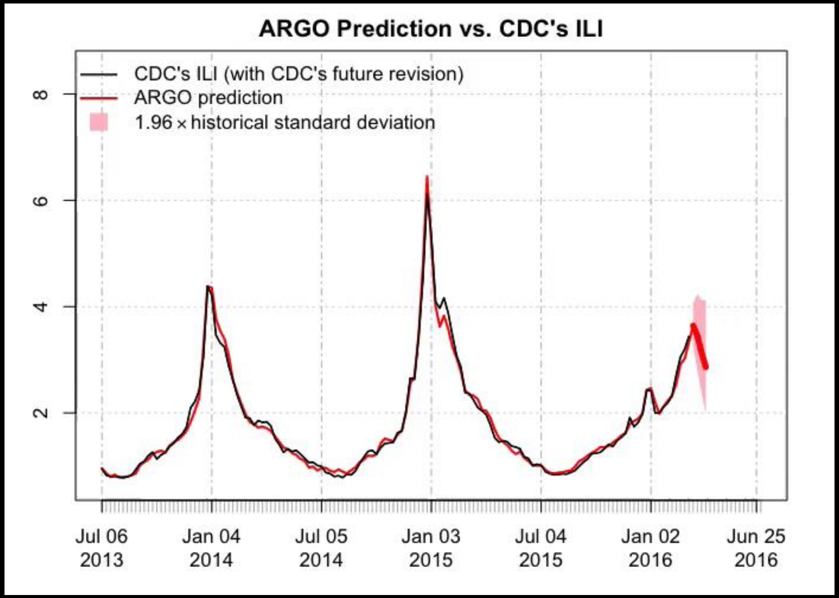
ARGO Prediction vs. CDC's ILI



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



CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION



Google

flu near you
do you have it in you?

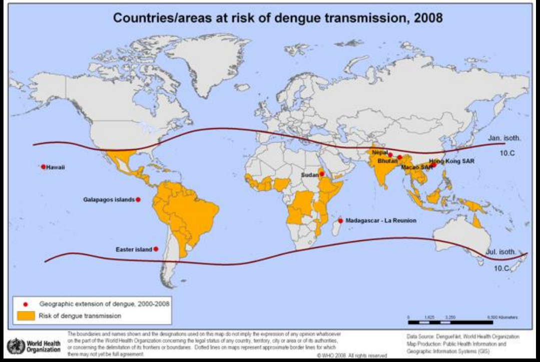
UpToDate
Providing Answers to Clinical Questions

athenahealth

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.

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

Dengue, Zika, and Flu



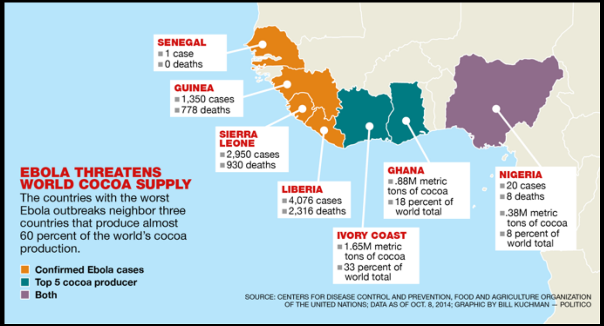
The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted lines on maps represent approximate border lines for which there may not yet be full agreement.

© WHO 2008. All rights reserved

Data Source: DengueNet, World Health Organization
Map Production: Public Health Information and Geographic Information Systems (GIS)

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

Ebola



EBOLA THREATENS WORLD COCOA SUPPLY

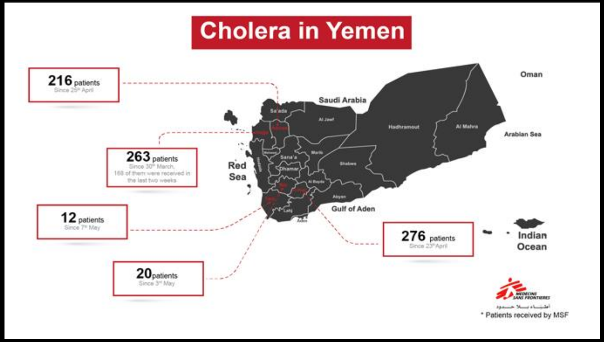
The countries with the worst Ebola outbreaks neighbor three countries that produce almost 60 percent of the world's cocoa production.

- 🟧 Confirmed Ebola cases
- 🟦 Top 5 cocoa producer
- 🟪 Both

SOURCE: CENTERS FOR DISEASE CONTROL AND PREVENTION, FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS; DATA AS OF OCT. 8, 2014; GRAPHIC BY BILL KUCHMAN — POLITICO

- West Africa

Cholera



- Middle East

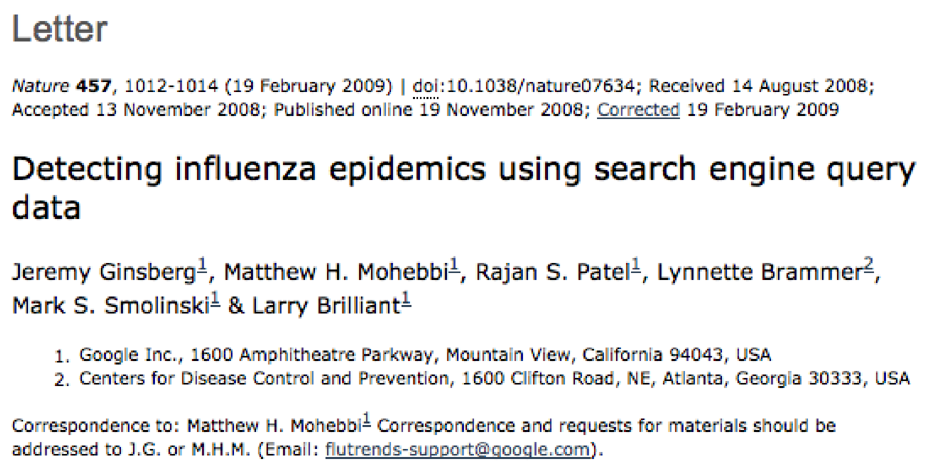
Seminal work by Google

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

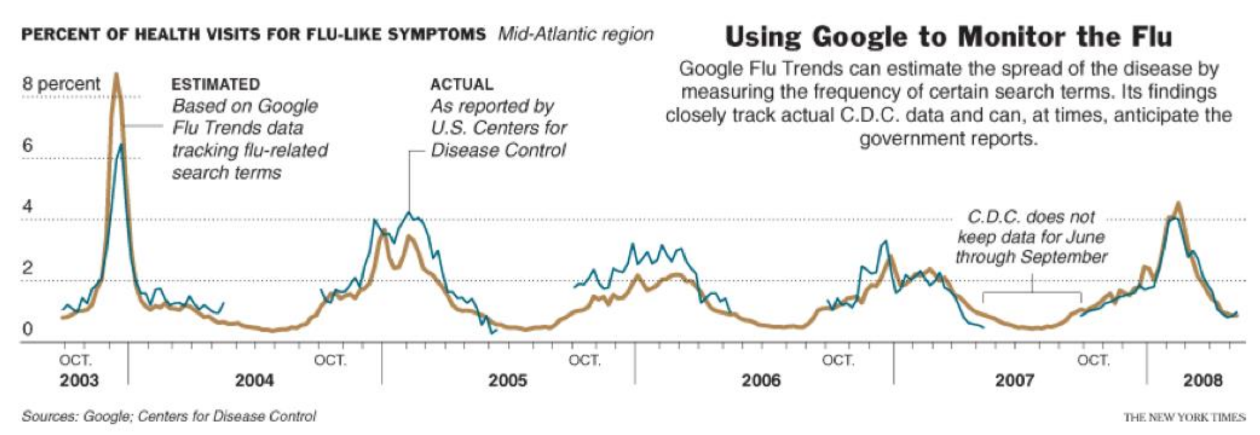
Detecting influenza epidemics using search engine query data

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



Using Google to Monitor the Flu

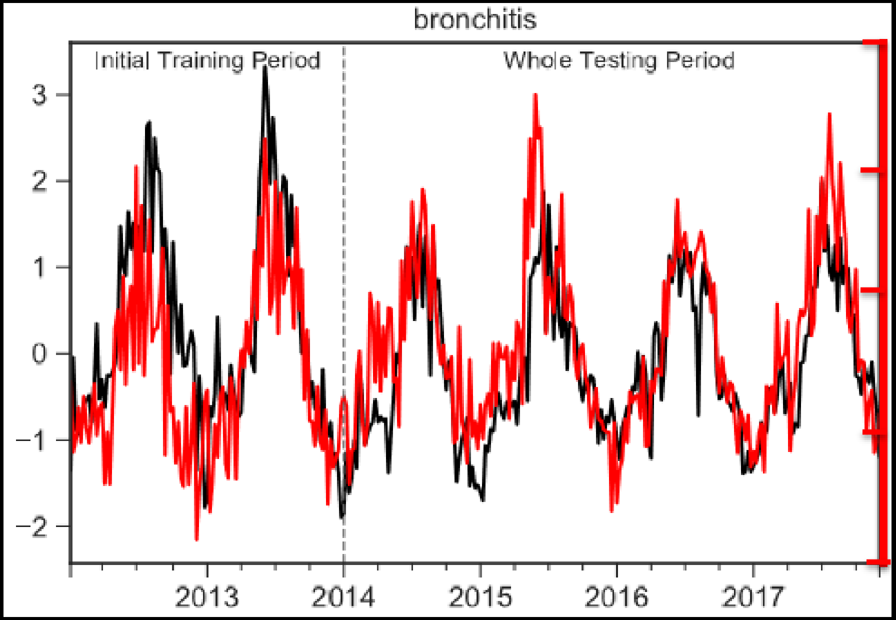
Google Flu Trends can estimate the spread of the disease by measuring the frequency of certain search terms. Its findings closely track actual C.D.C. data and can, at times, anticipate the government reports.

Sources: Google; Centers for Disease Control

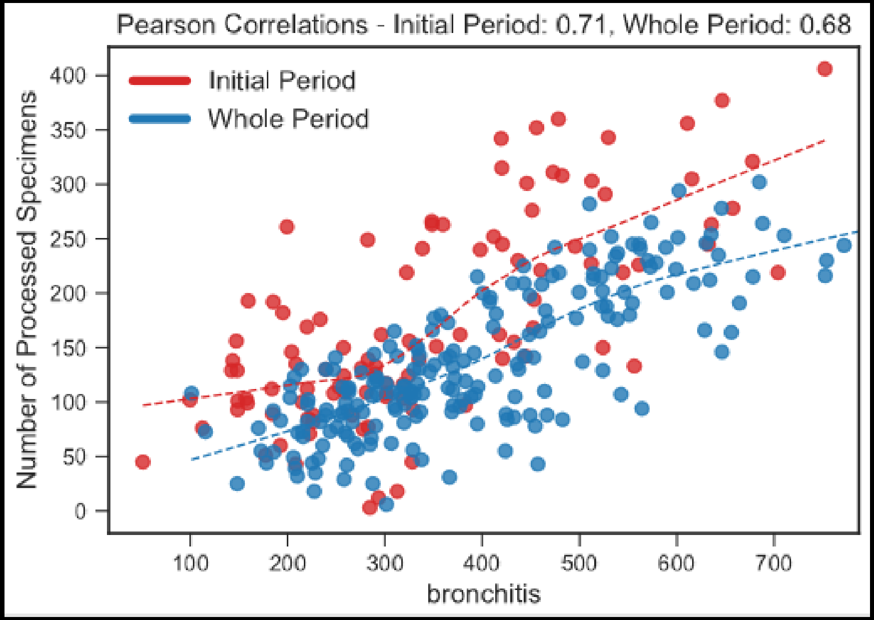
THE NEW YORK TIMES

Google Flu Trends

What is the logic behind this approach?



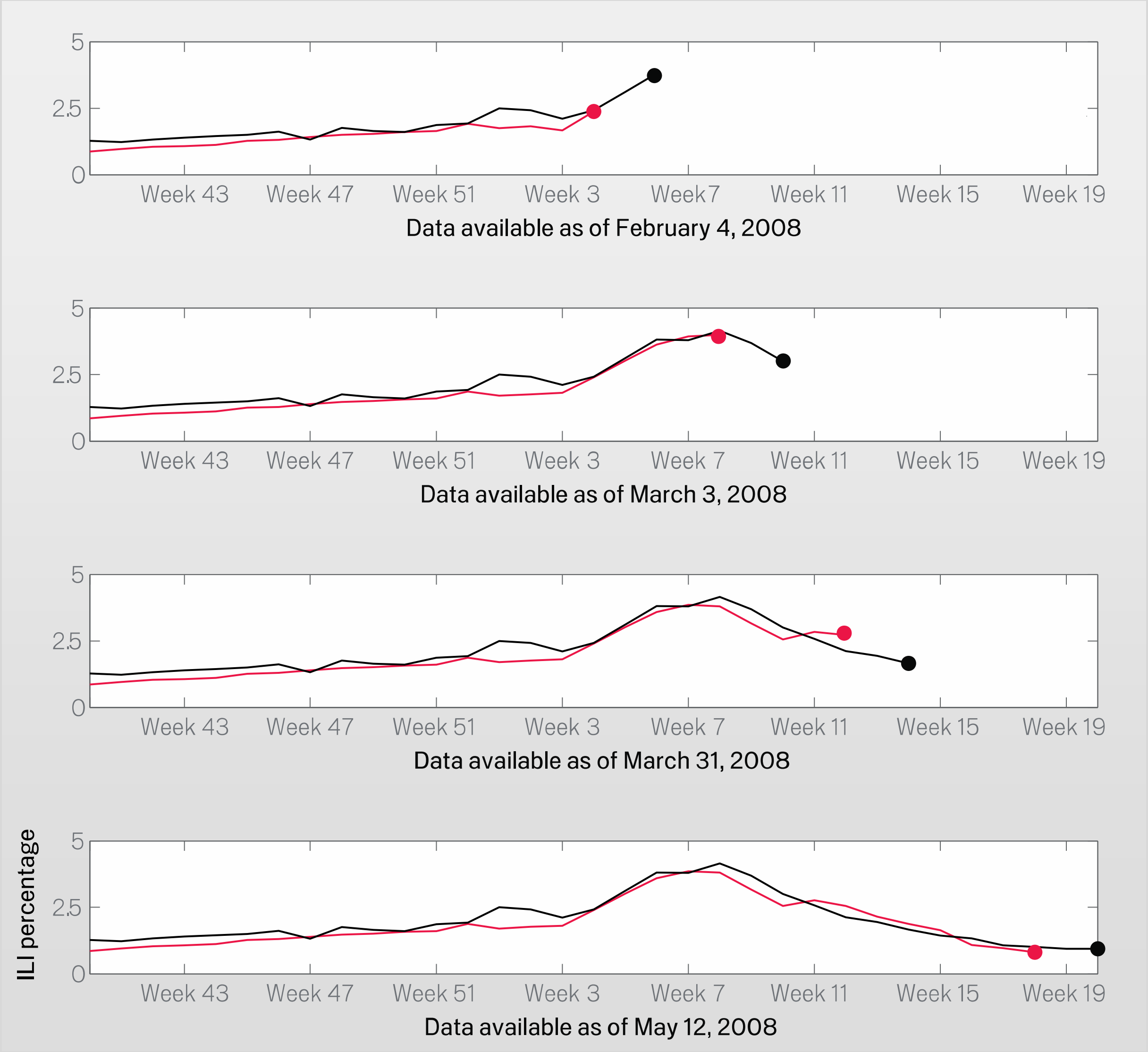
Searches on “bronchitis” vs Flu activity in South Africa



Scatter plot of searches vs flu

google.org Flu Trends

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



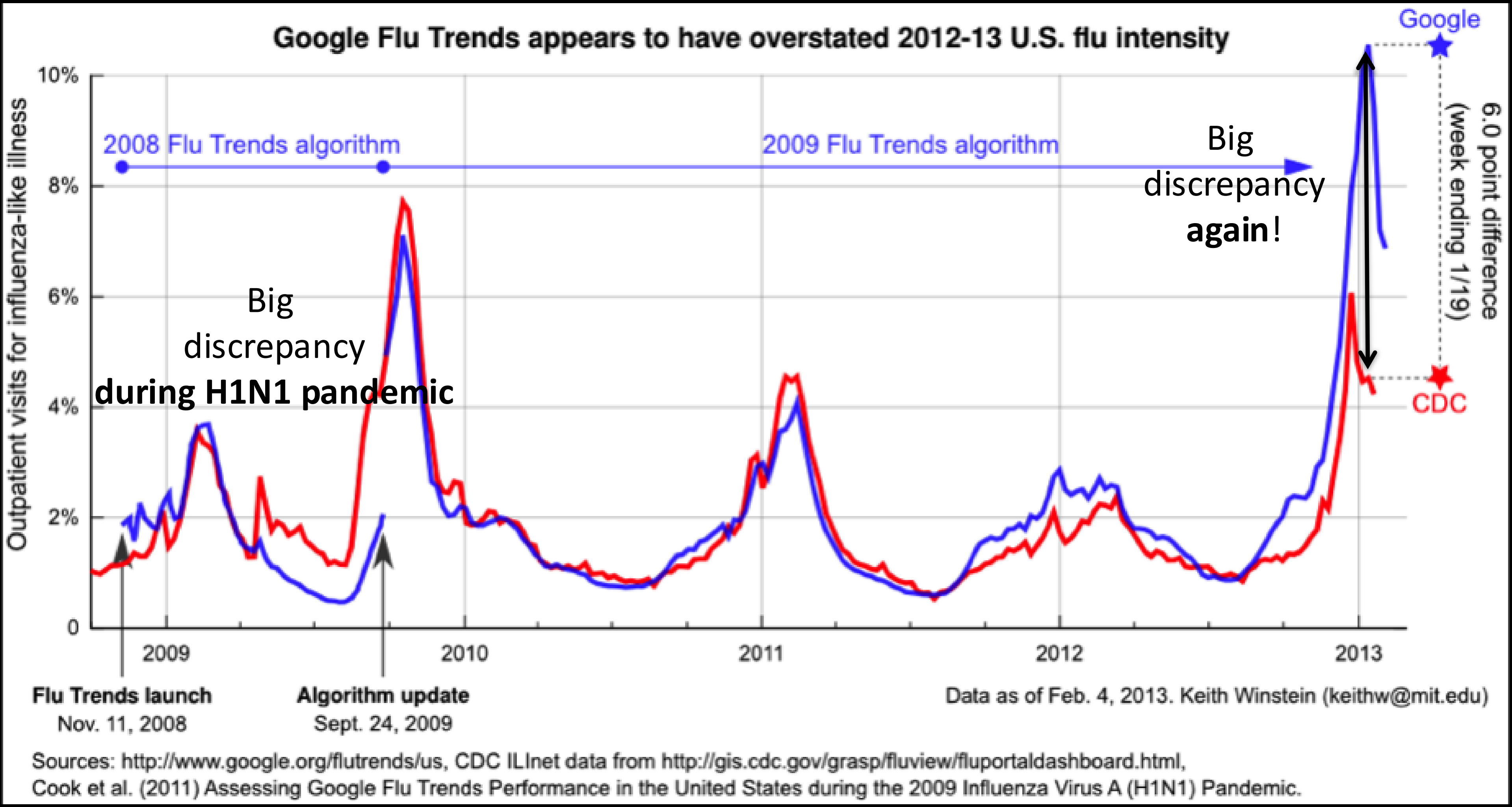
Data available as of February 4, 2008

Data available as of March 3, 2008

Data available as of March 31, 2008

Data available as of May 12, 2008

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


Google Flu Trends appears to have overstated 2012-13 U.S. flu intensity

Big discrepancy **during H1N1 pandemic**

Big discrepancy **again**!

Data as of Feb. 4, 2013. Keith Winstein (keithw@mit.edu)

Sources: http://www.google.org/flutrends/us, CDC ILInet data from http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html, Cook et al. (2011) Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic.

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

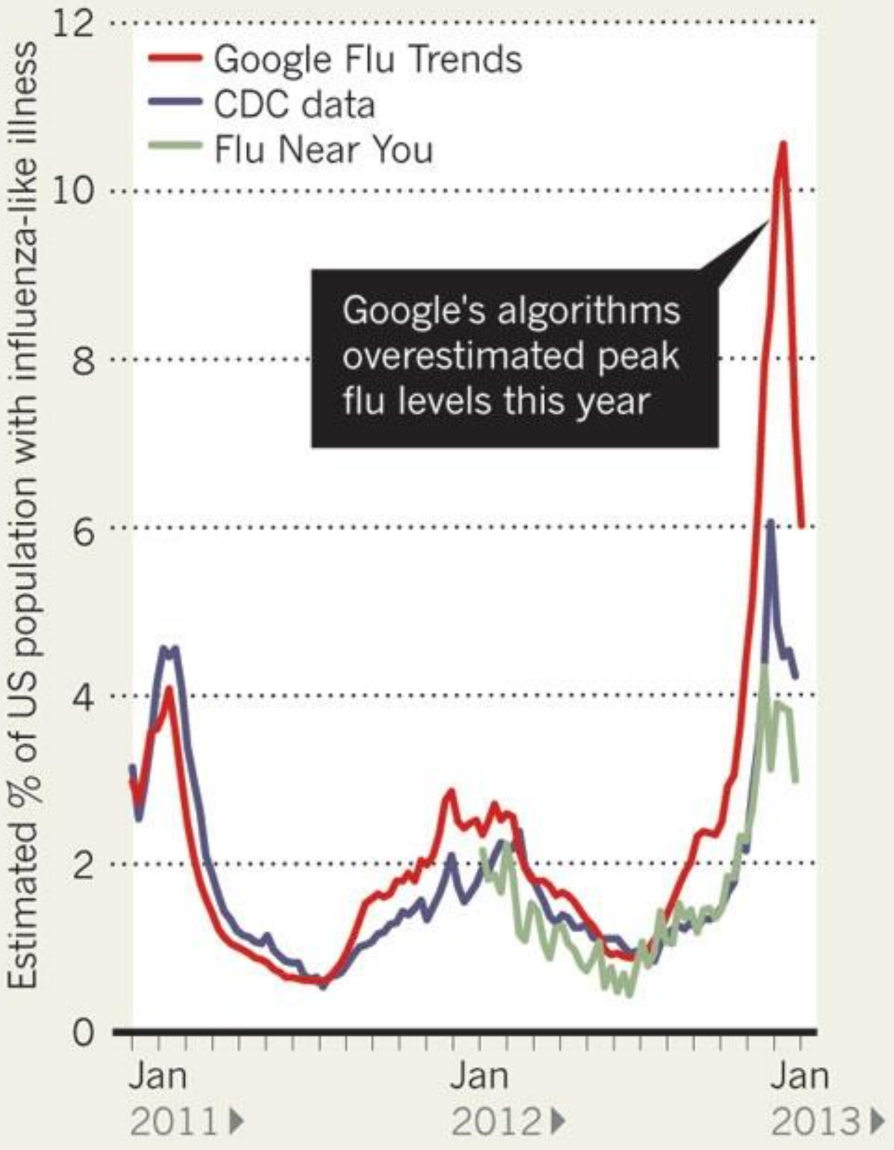
nature International weekly journal of science

When Google got flu wrong.

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

FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.



Google's algorithms overestimated peak flu levels this year

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

12 comments, 7 called-out

> *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 » 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

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

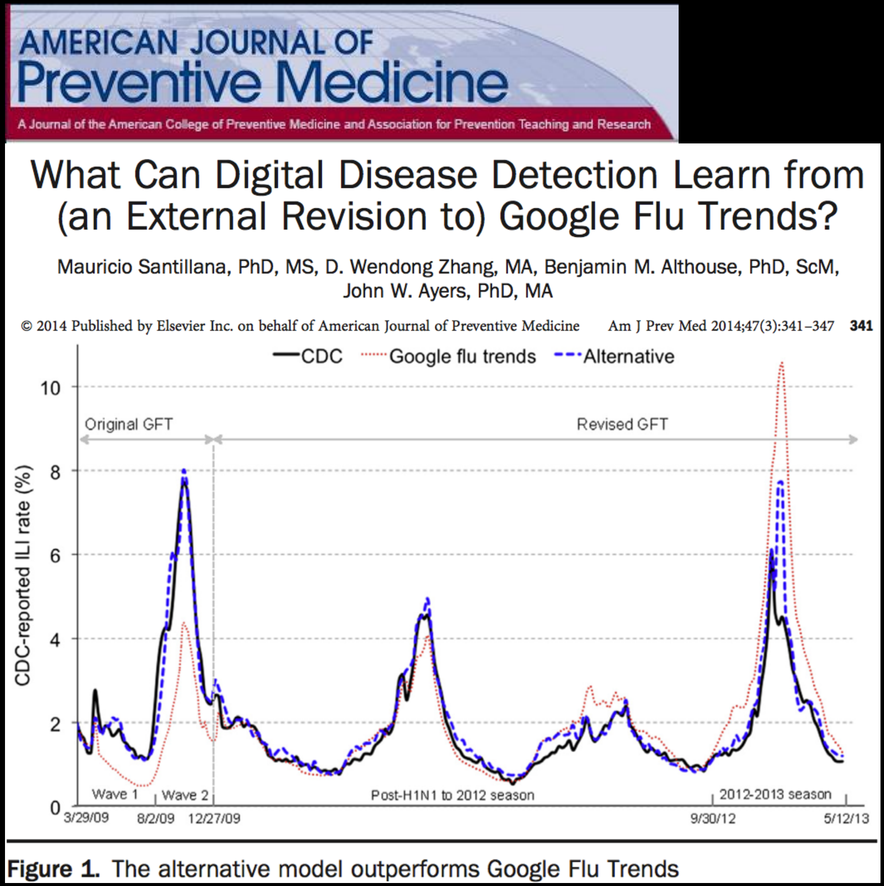
(with D. Wendong Zhang)

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.

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

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.


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

© 2014 Published by Elsevier Inc. on behalf of American Journal of Preventive Medicine Am J Prev Med 2014;47(3):341–347 341

Figure 1. The alternative model outperforms Google Flu Trends

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

Google Flu Trends gets a brand new engine

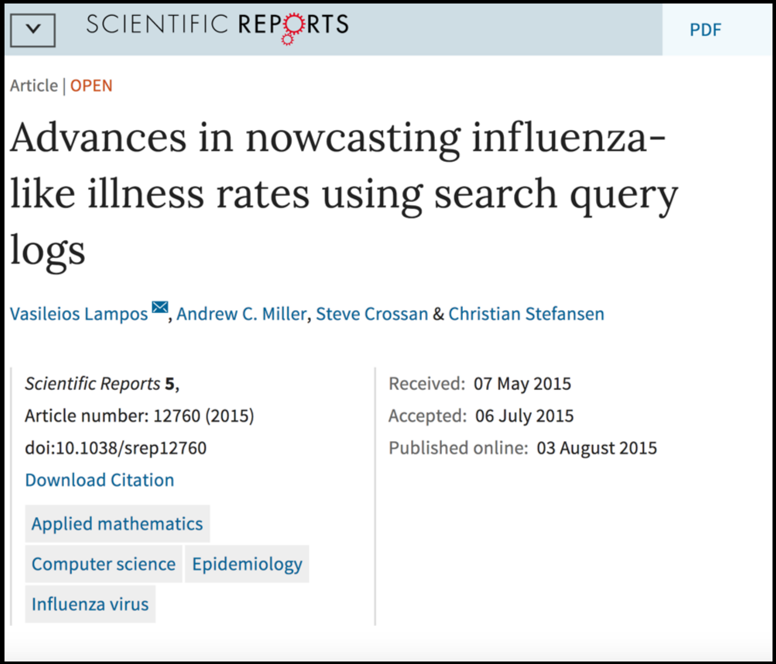
Posted: Friday, October 31, 2014

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.

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



Article | OPEN

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

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

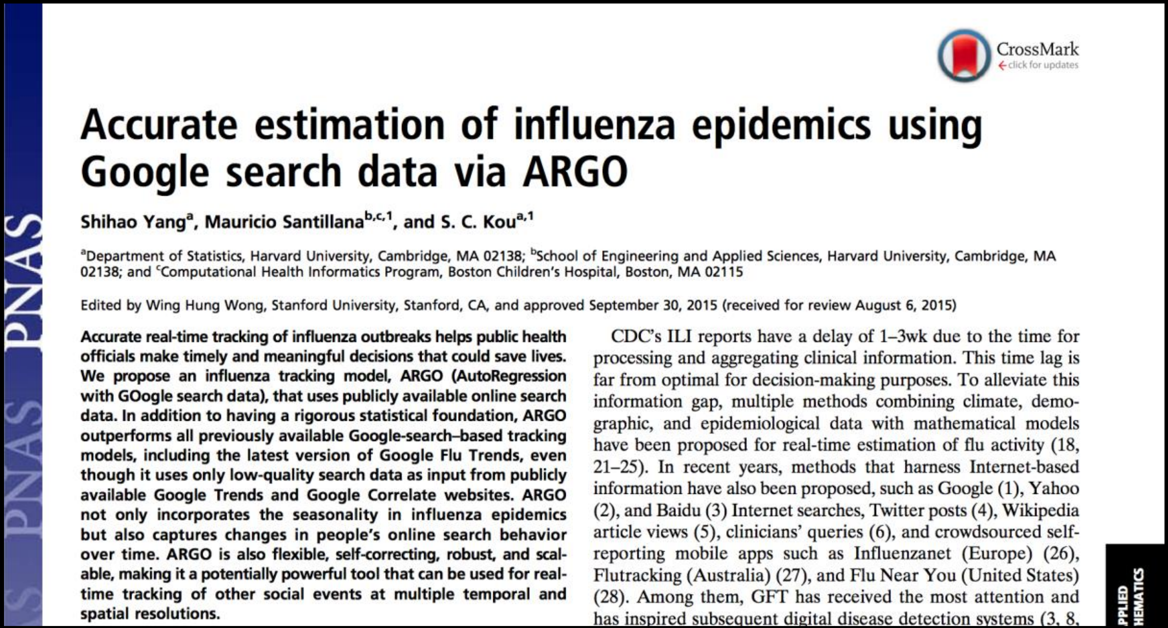
Scientific Reports **5**,
Article number: 12760 (2015)
doi:10.1038/srep12760
Download Citation

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



Accurate estimation of influenza epidemics using Google search data via ARGO

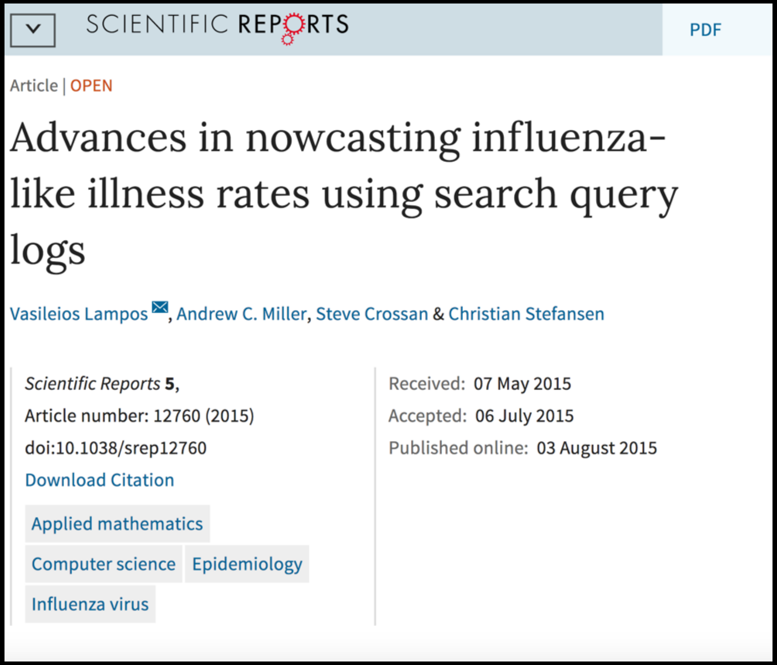
Shihao Yang[a], Mauricio Santillana[b,c,1], and S. C. Kou[a,1]

[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,



Article | OPEN

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

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

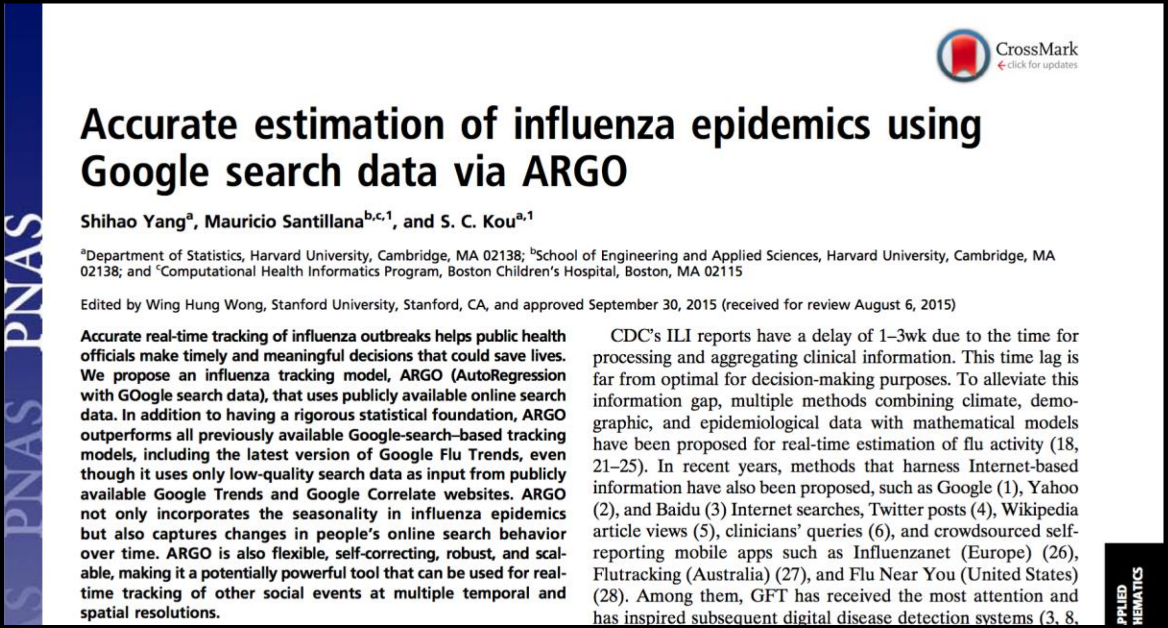
Scientific Reports **5**,
Article number: 12760 (2015)
doi:10.1038/srep12760
Download Citation

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



Accurate estimation of influenza epidemics using Google search data via ARGO

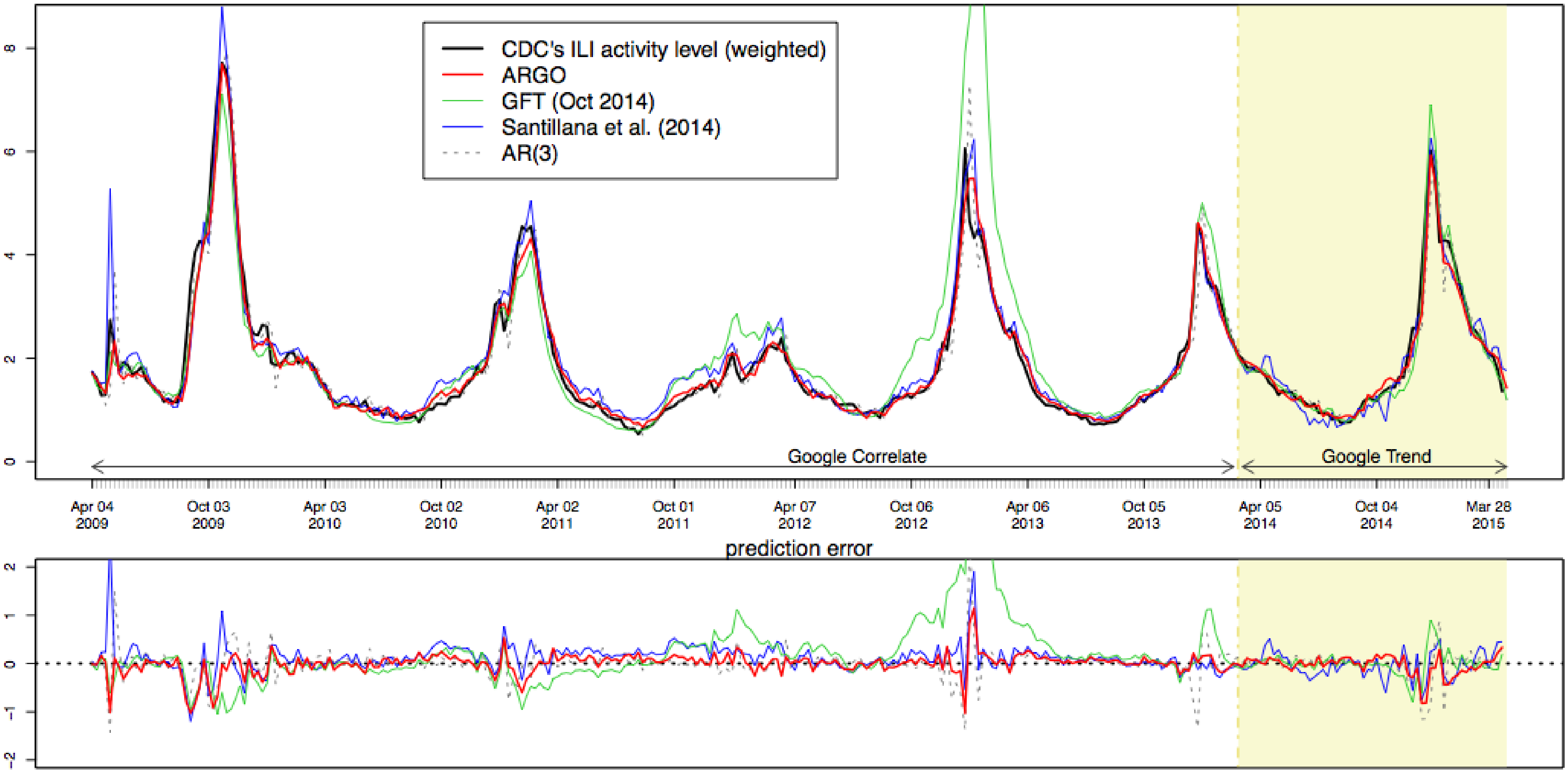
Shihao Yang[a], Mauricio Santillana[b,c,1], and S. C. Kou[a,1]

[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

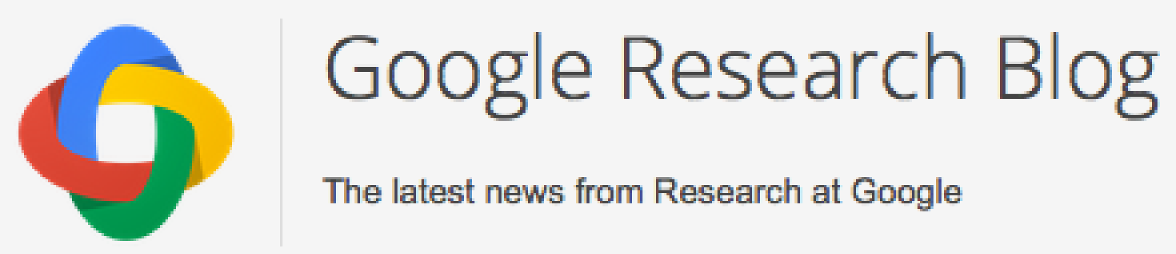
Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved September 30, 2015 (received for review August 6, 2015)

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,

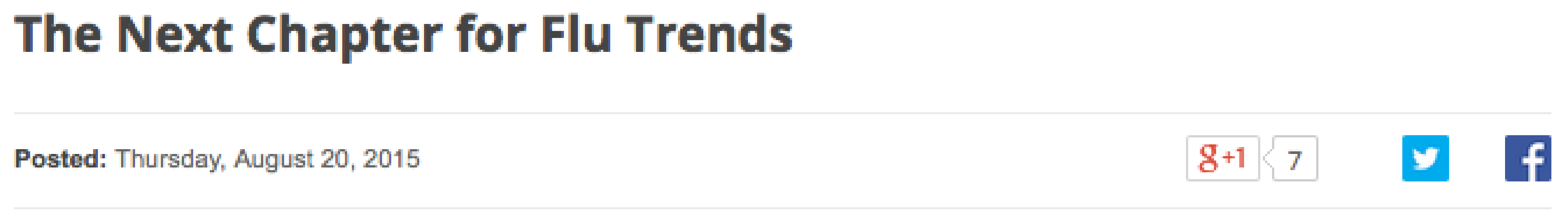

CDC's ILI activity level (weighted)
ARGO
GFT (Oct 2014)
Santillana et al. (2014)
AR(3)
0
2
4
6
8
Google Correlate
Google Trend
Apr 04 2009
Oct 03 2009
Apr 03 2010
Oct 02 2010
Apr 02 2011
Oct 01 2011
Apr 07 2012
Oct 06 2012
Apr 06 2013
Oct 05 2013
Apr 05 2014
Oct 04 2014
Mar 28 2015
prediction error
-2
-1
0
1
2

Google discontinues Flu Trends indefinitely!



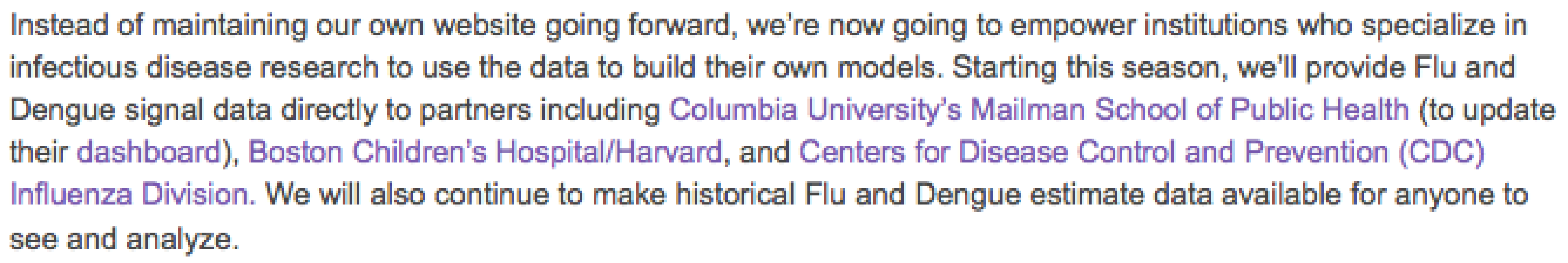
Google Research Blog

The latest news from Research at Google



The Next Chapter for Flu Trends

Posted: Thursday, August 20, 2015



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.

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

BIG DATA

Google discontinues Flu Trends, starts offering data to researchers

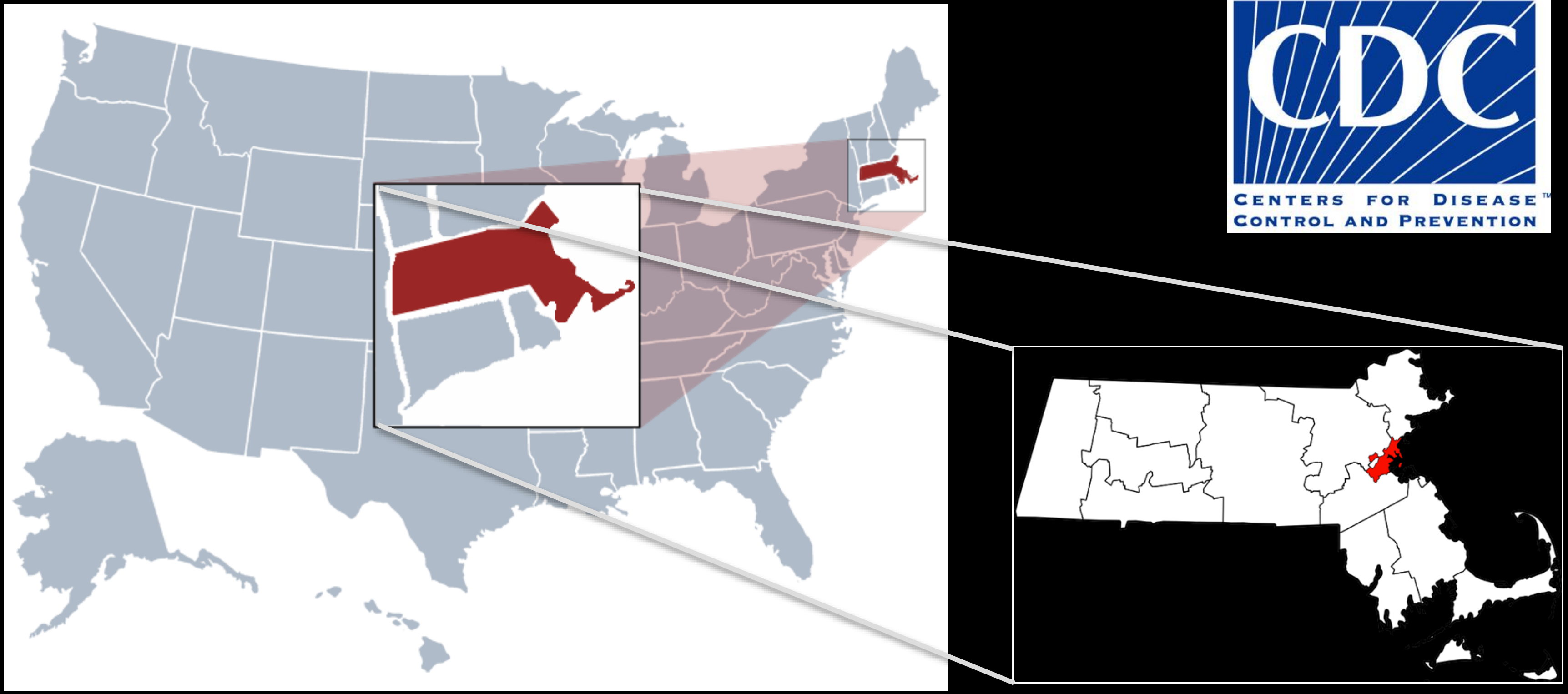
JORDAN NOVET AUGUST 20, 2015 12:17 PM

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*



CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Prepared by Mauricio Santillana

Beyond Google searches…

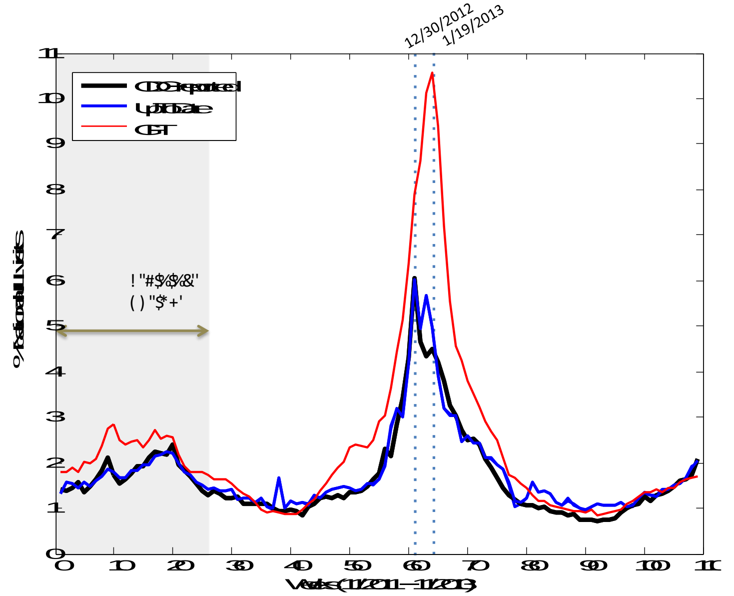
What are doctors searching for?

flu near you
do you have it in you?

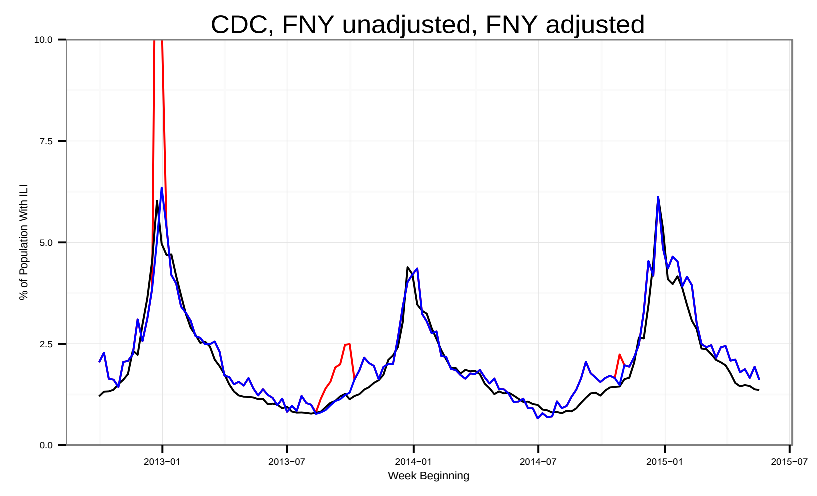
What are people tweeting? What are they reporting on crowd-sourced disease surveillance apps?

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

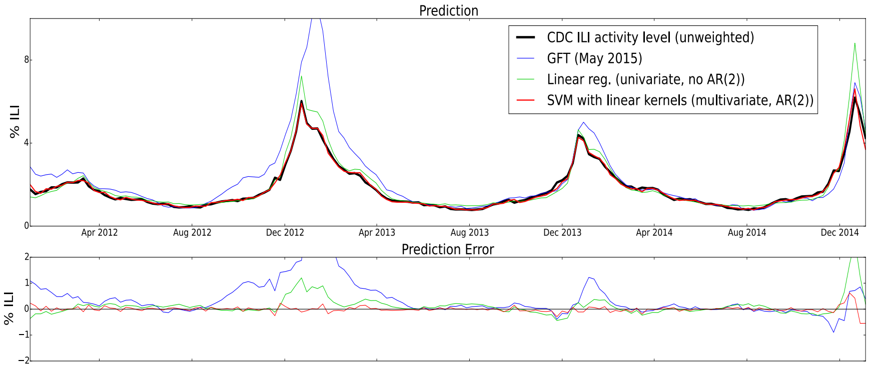
Beyond Google searches…



What are doctors searching for?

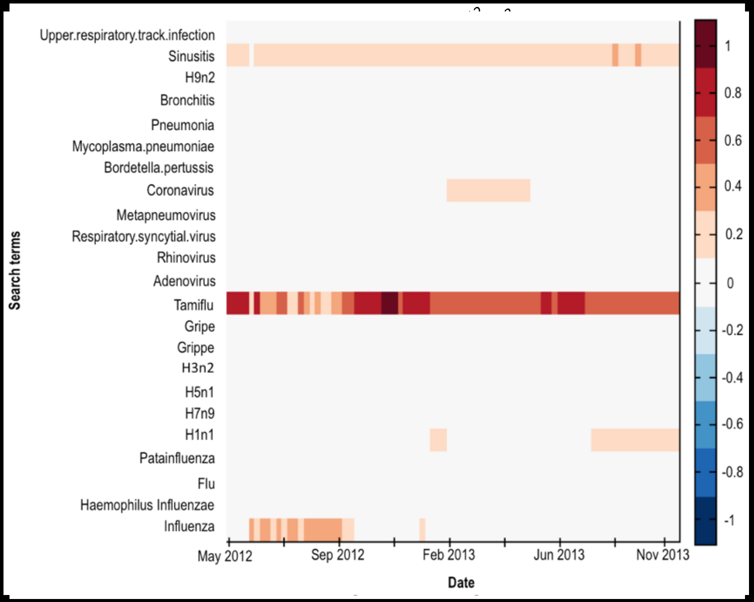


What are people tweeting? What are they reporting on crowd-sourced disease surveillance apps?



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

Beyond Google searches…



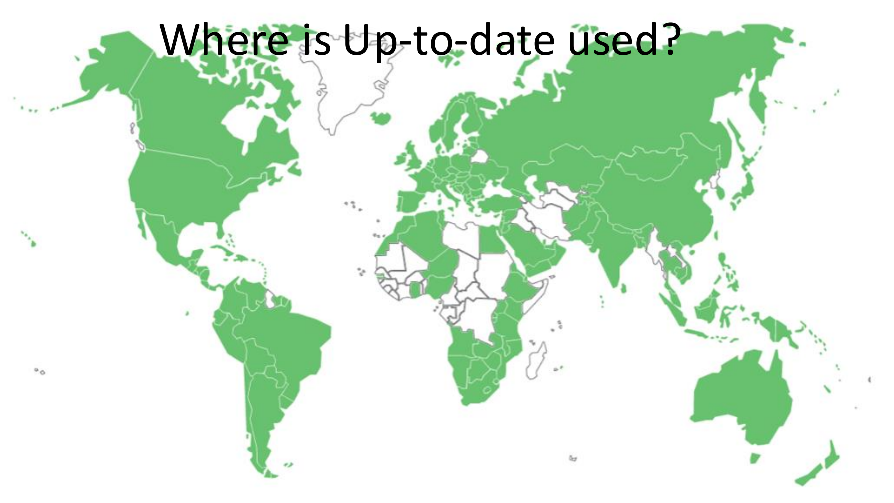
What are doctors searching for?

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…

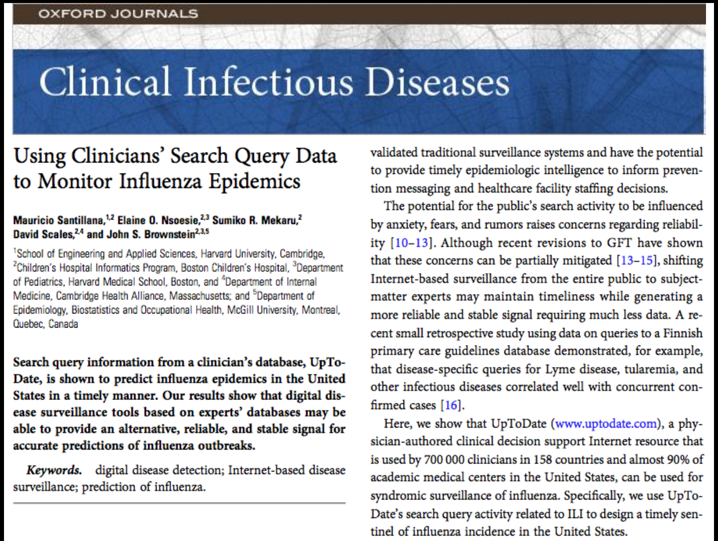

Where is Up-to-date used?

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…



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[2,3,5]

[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.

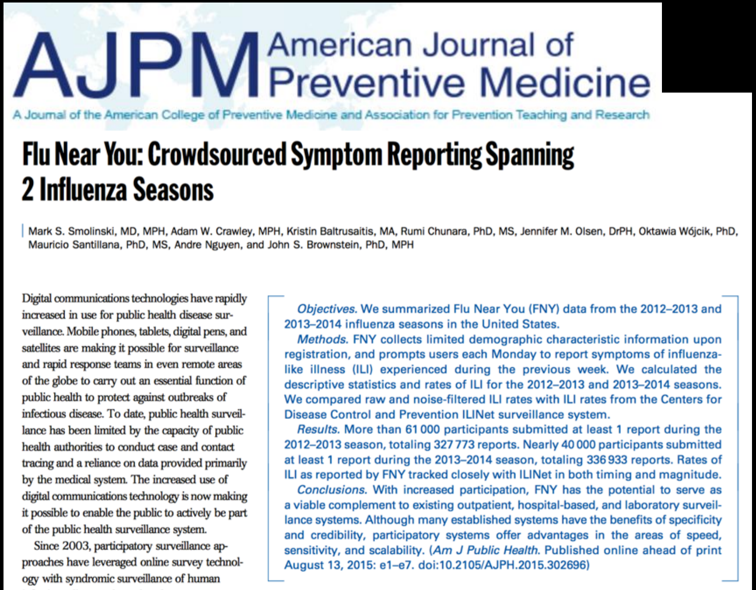
Keywords. digital disease detection; Internet-based disease surveillance; prediction of influenza.

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

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.

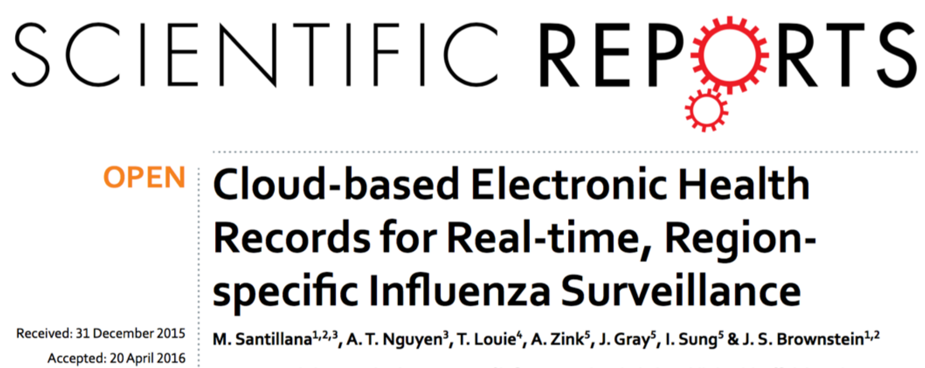
Results. More than 61 000 participants submitted at least 1 report during the 2012–2013 season, totaling 327 773 reports. Nearly 40 000 participants submitted at least 1 report during the 2013–2014 season, totaling 336 933 reports. Rates of ILI as reported by FNY tracked closely with ILINet in both timing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

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

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**

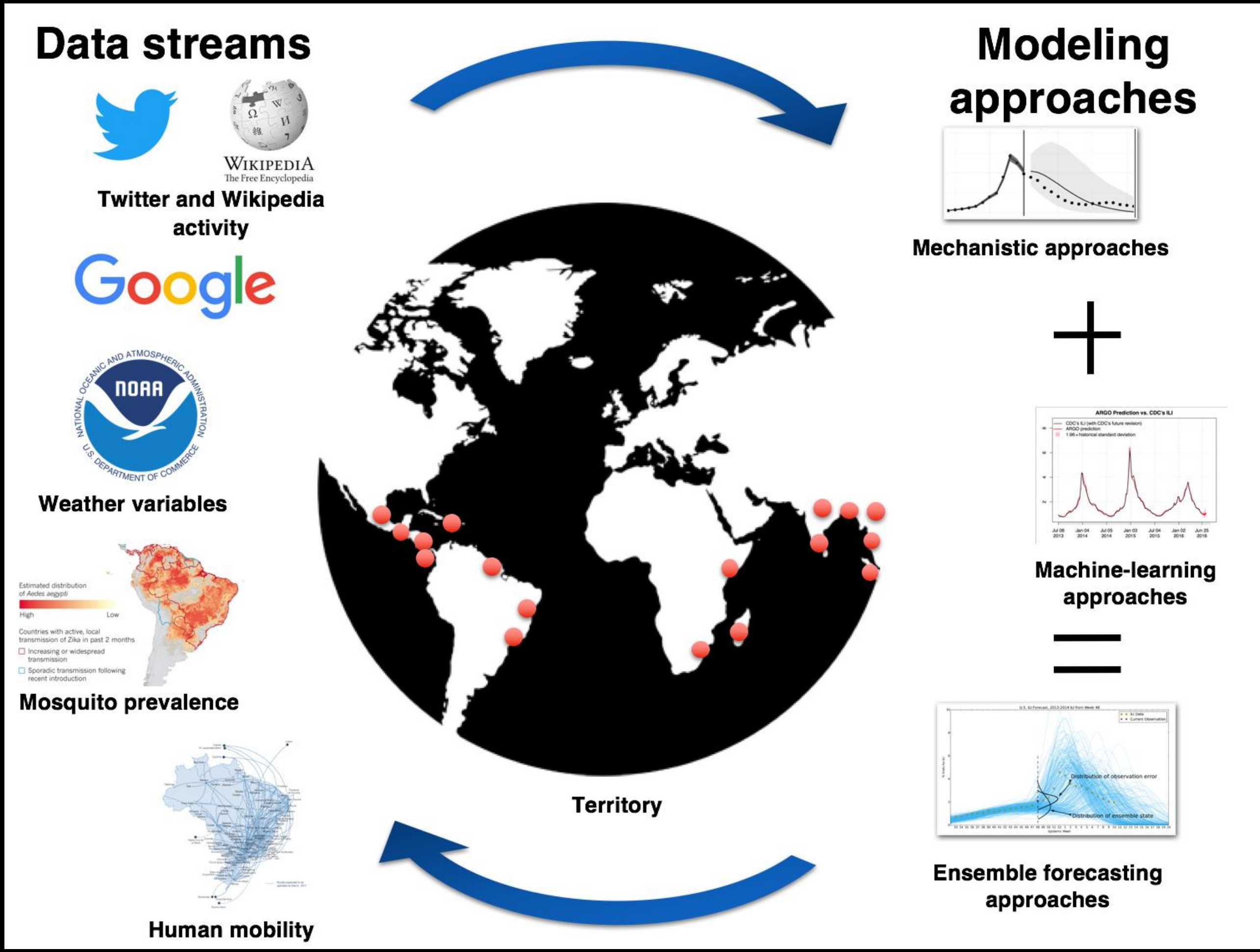
Location: **United States** (Data rich, wealthy country)

Spatial resolution: **Country**

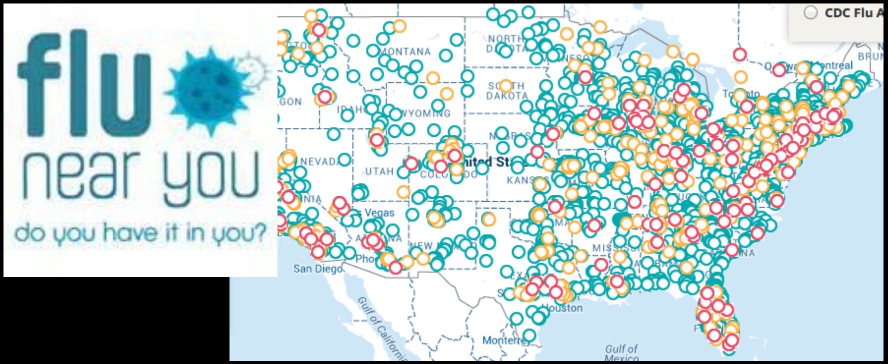
Method: **Machine learning**

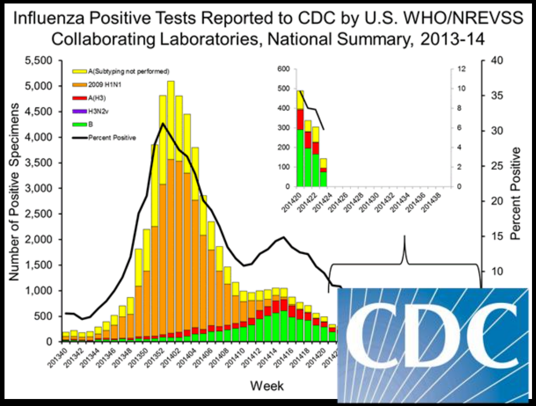
Input data sources:

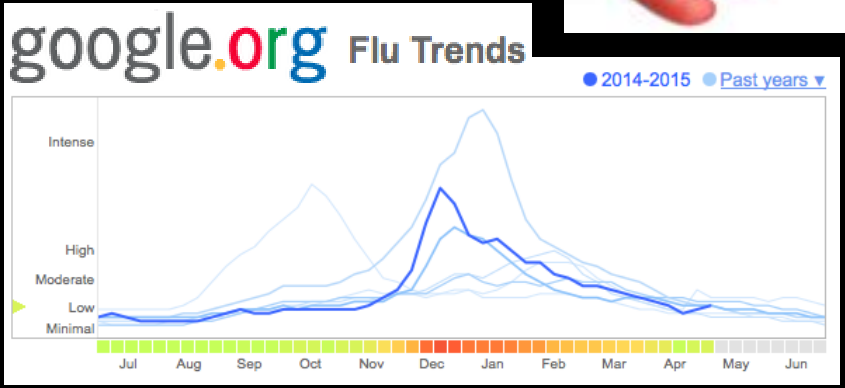
- **Historical flu activity**
- **Google search activity**
- **Electronic Health records**
- **Crowd sourced information**

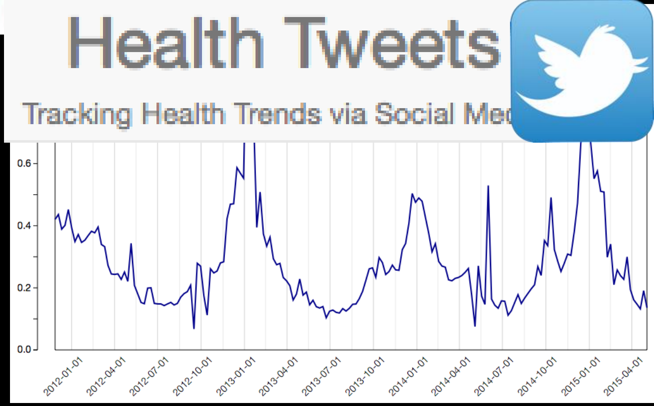


Ensemble approaches yield more accurate and more robust real-time and forecast flu estimates

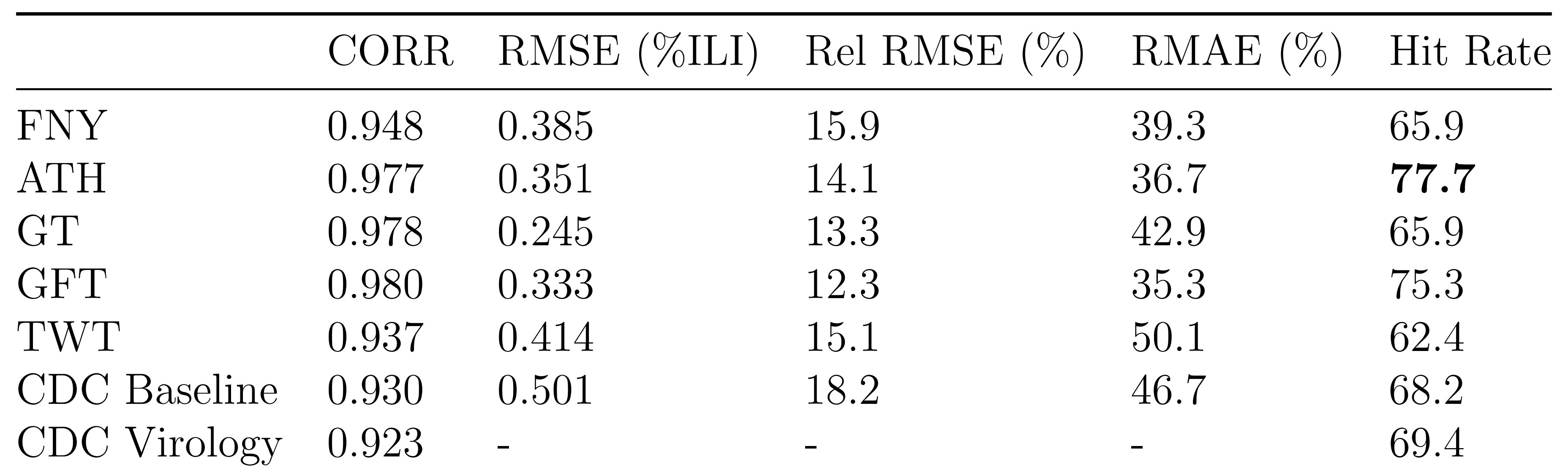








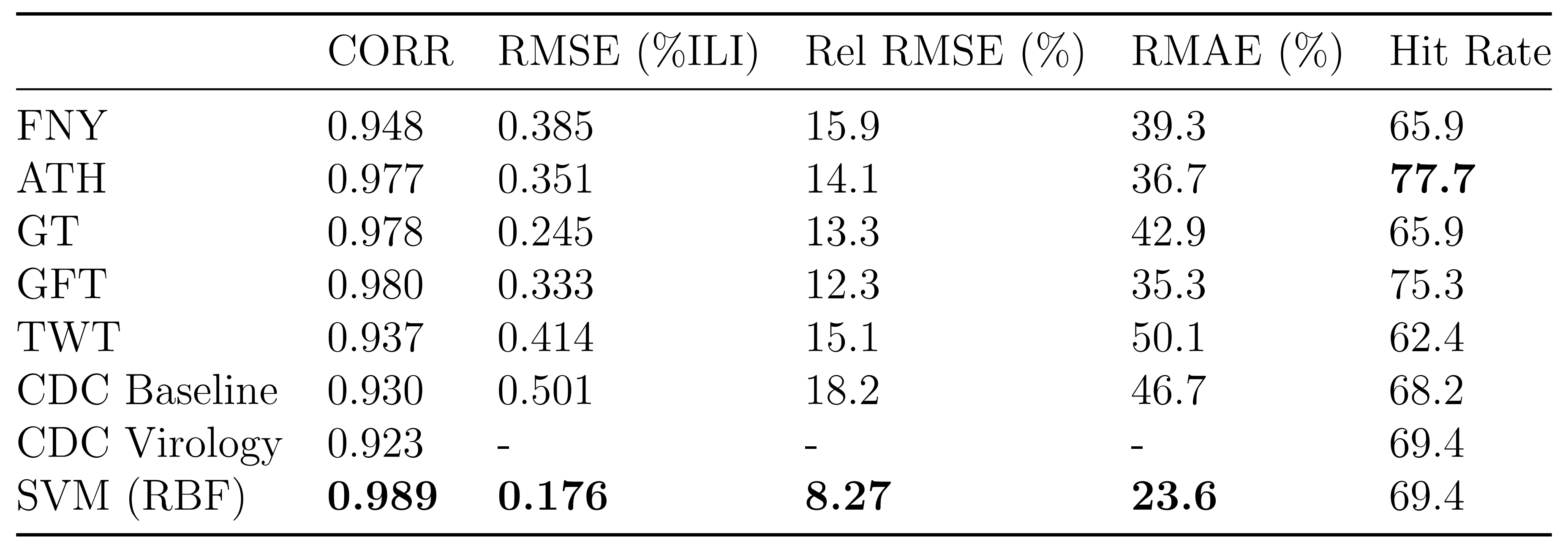
Performance of individual data sources



	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

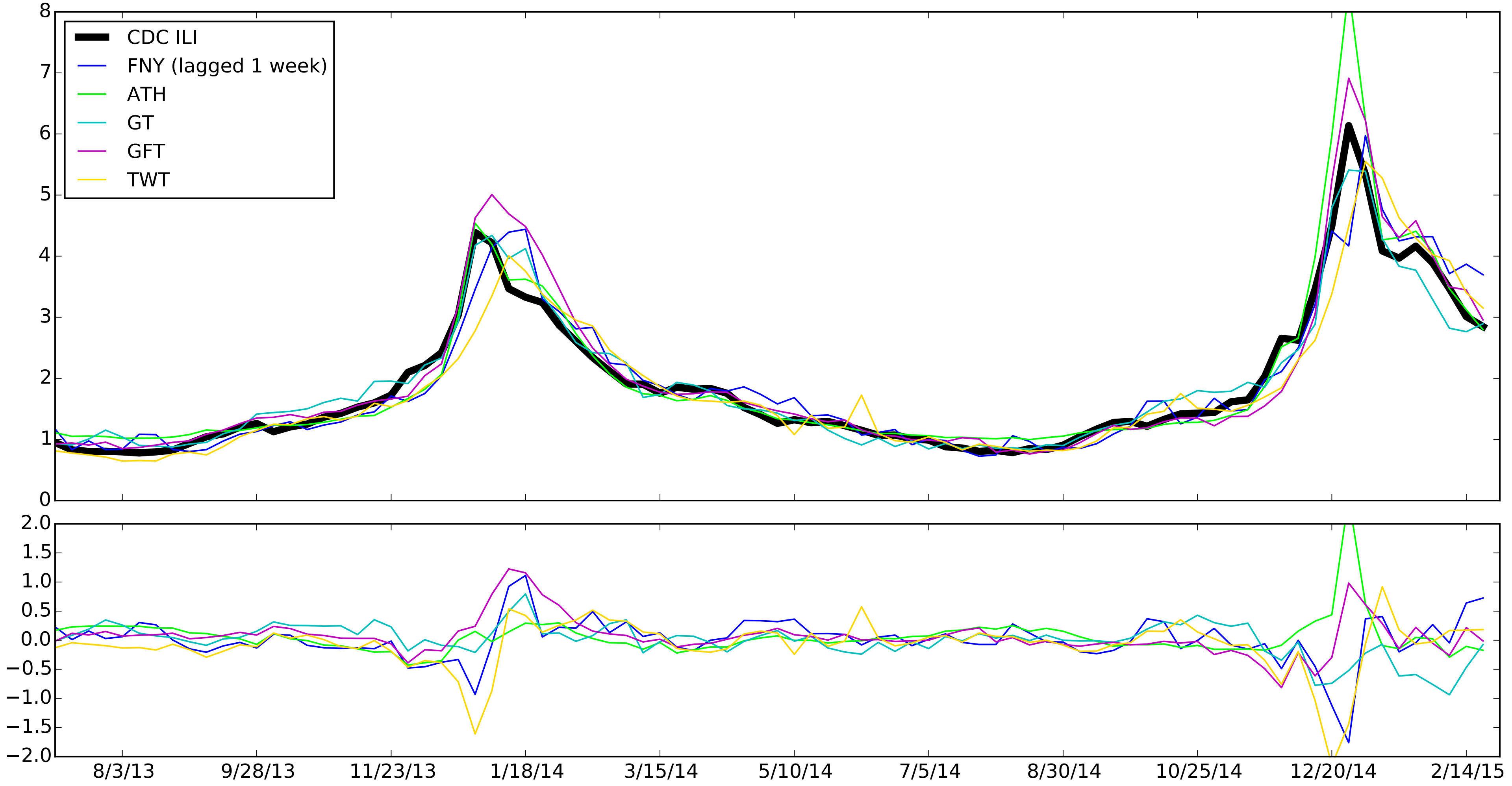
Performance ensemble



	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

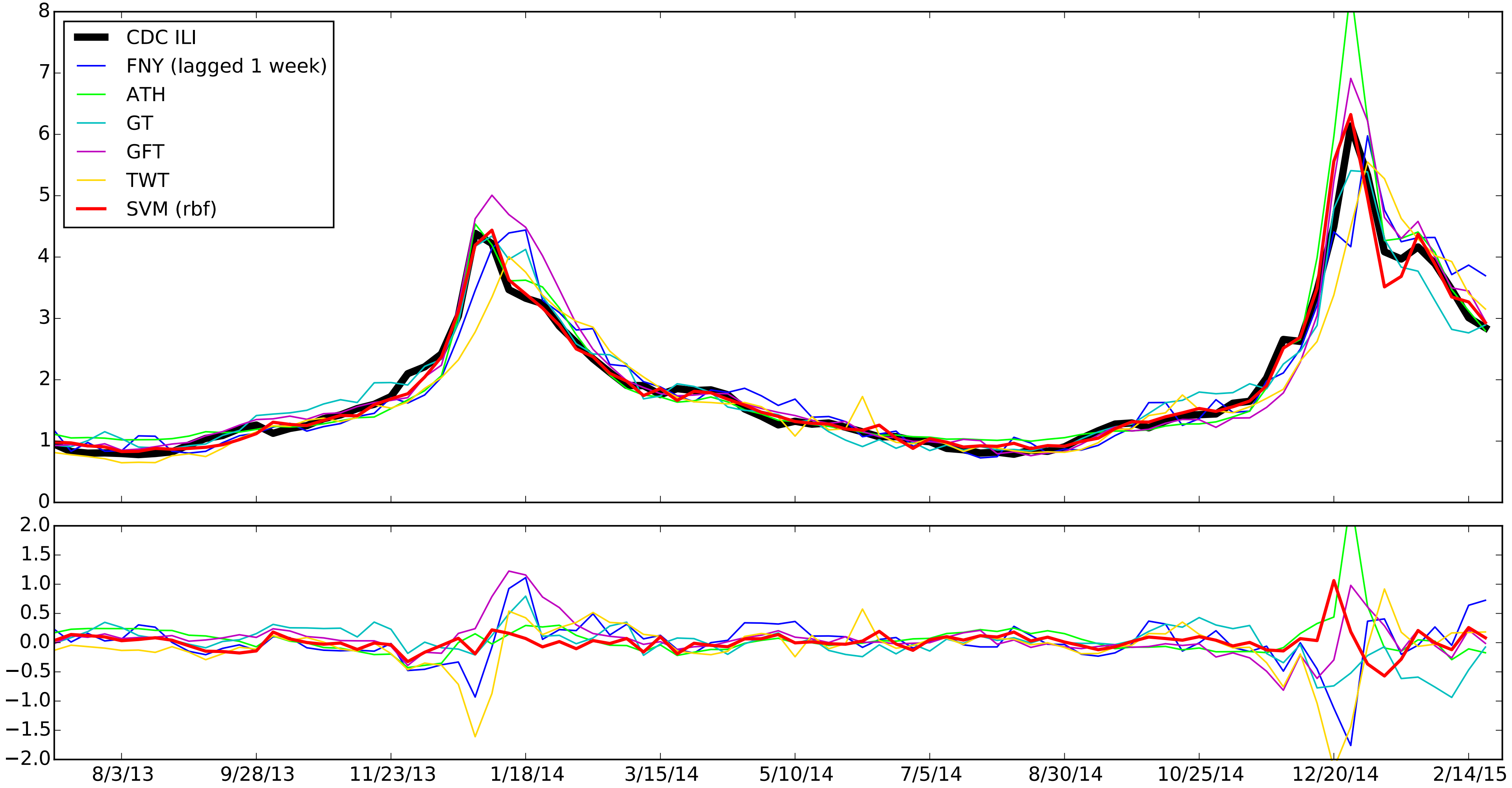
Santillana et al. PLoS Computational Biology, 2015

Performance of individual data sources



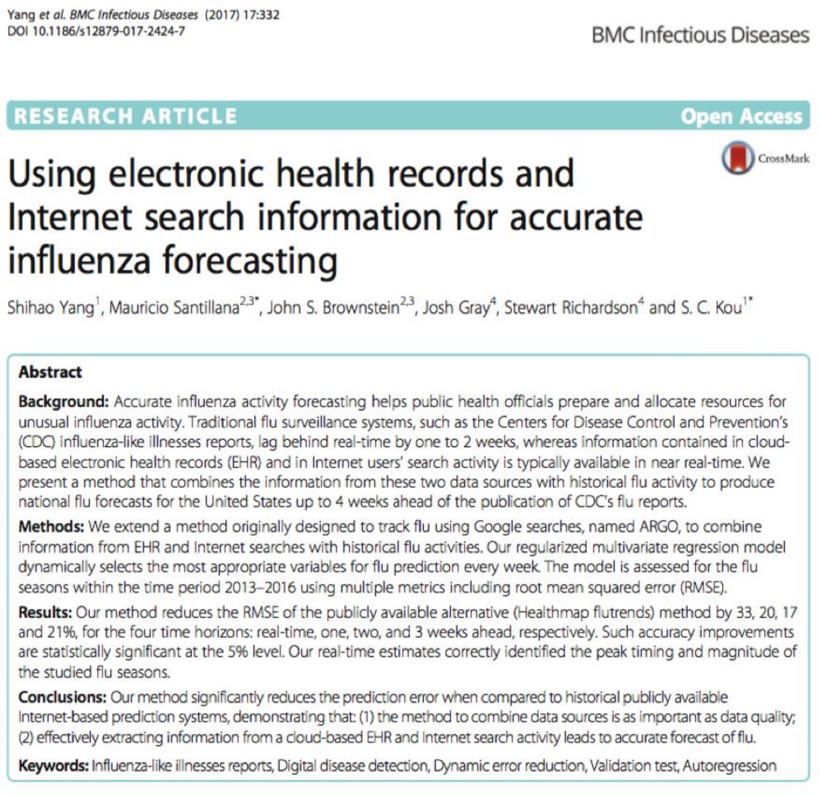
Santillana et al. PLoS Computational Biology, 2015

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

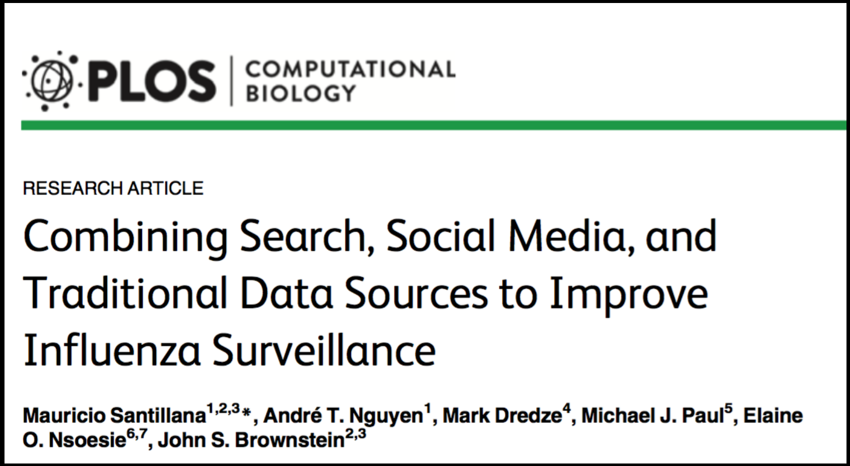
Background: Accurate influenza activity forecasting helps public health officials prepare and allocate resources for unusual influenza activity. Traditional flu surveillance systems, such as the Centers for Disease Control and Prevention's (CDC) influenza-like illnesses reports, lag behind real-time 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

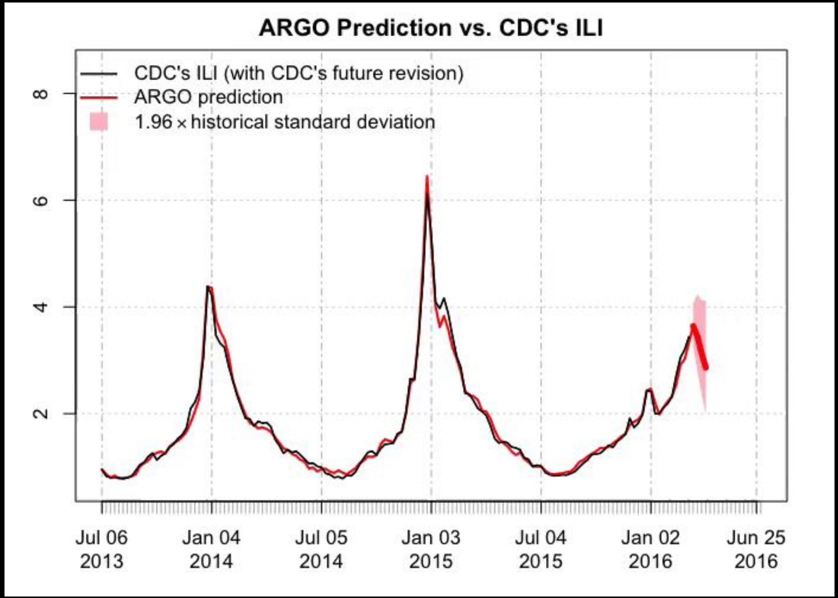


PLOS | COMPUTATIONAL BIOLOGY

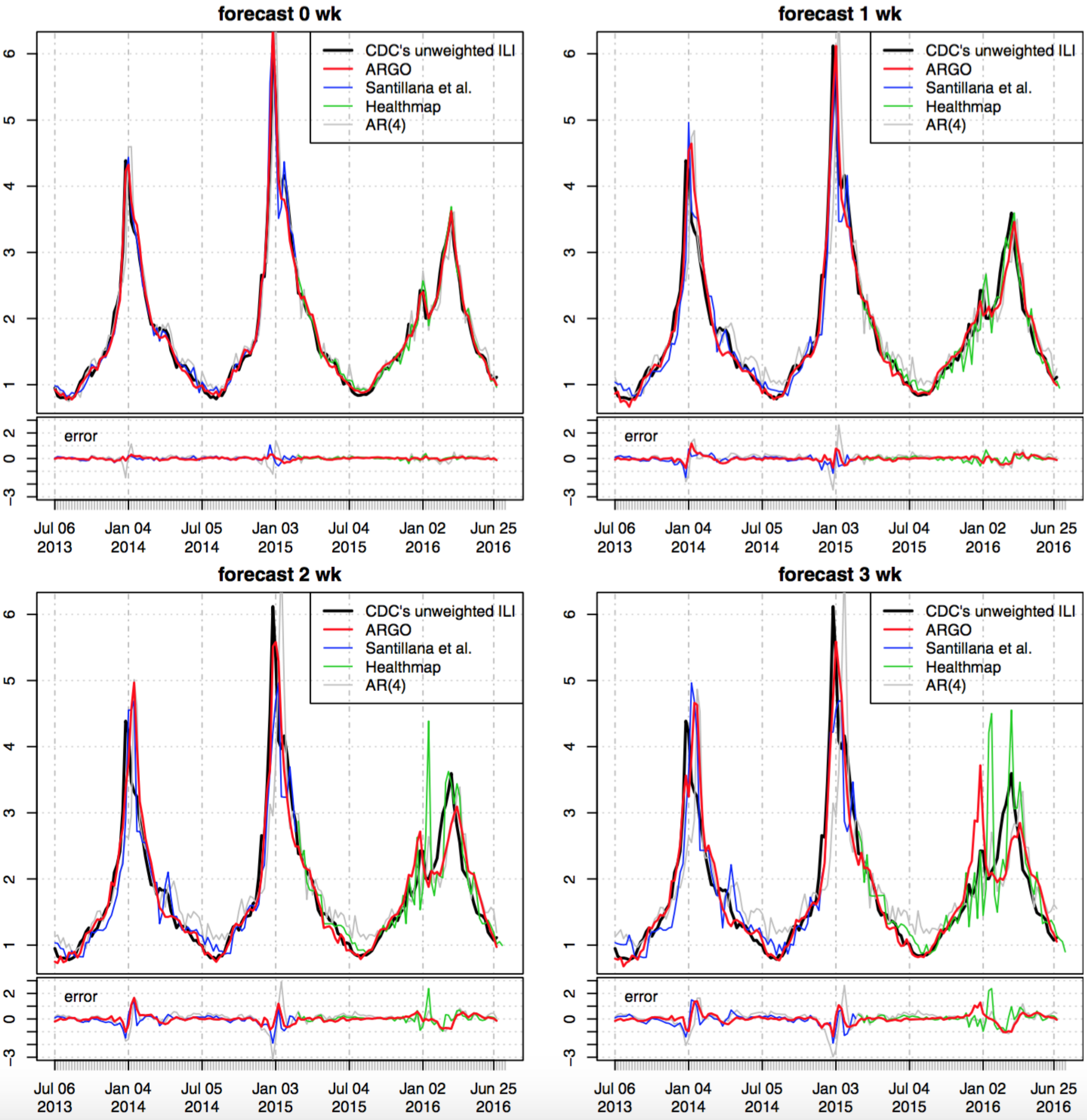
RESEARCH ARTICLE

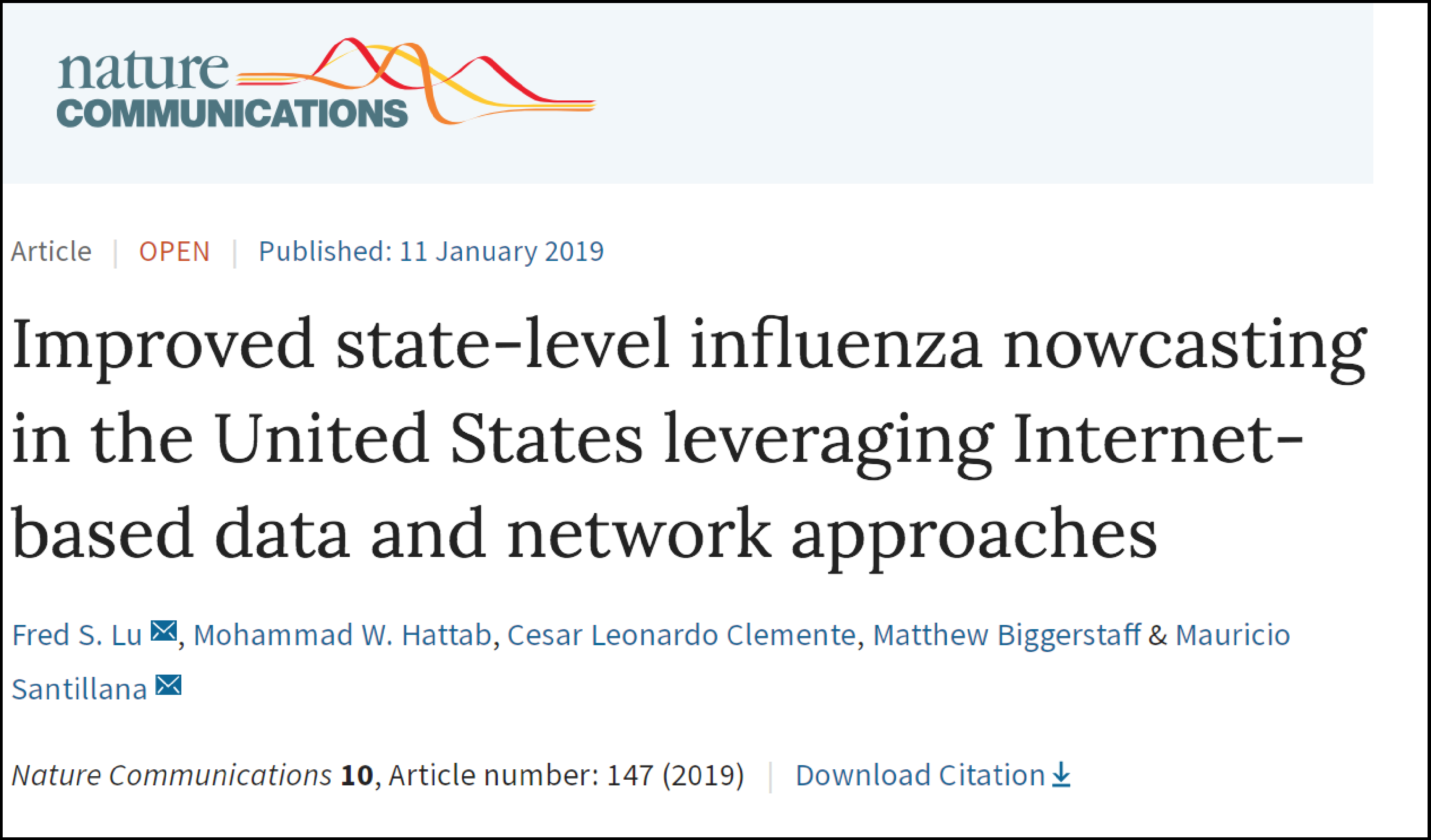
Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance

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]



ARGO Prediction vs. CDC's ILI


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

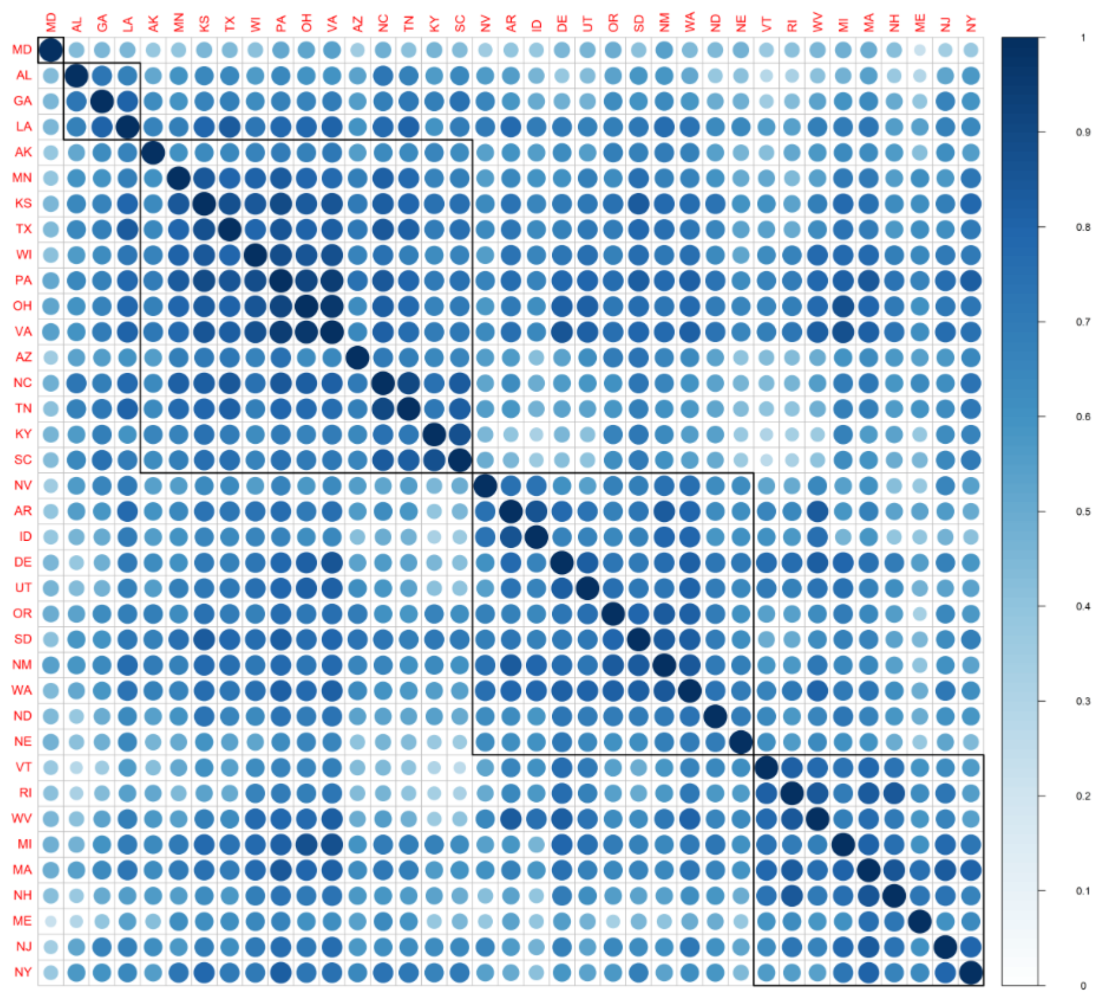
Spatial-temporal synchronicities

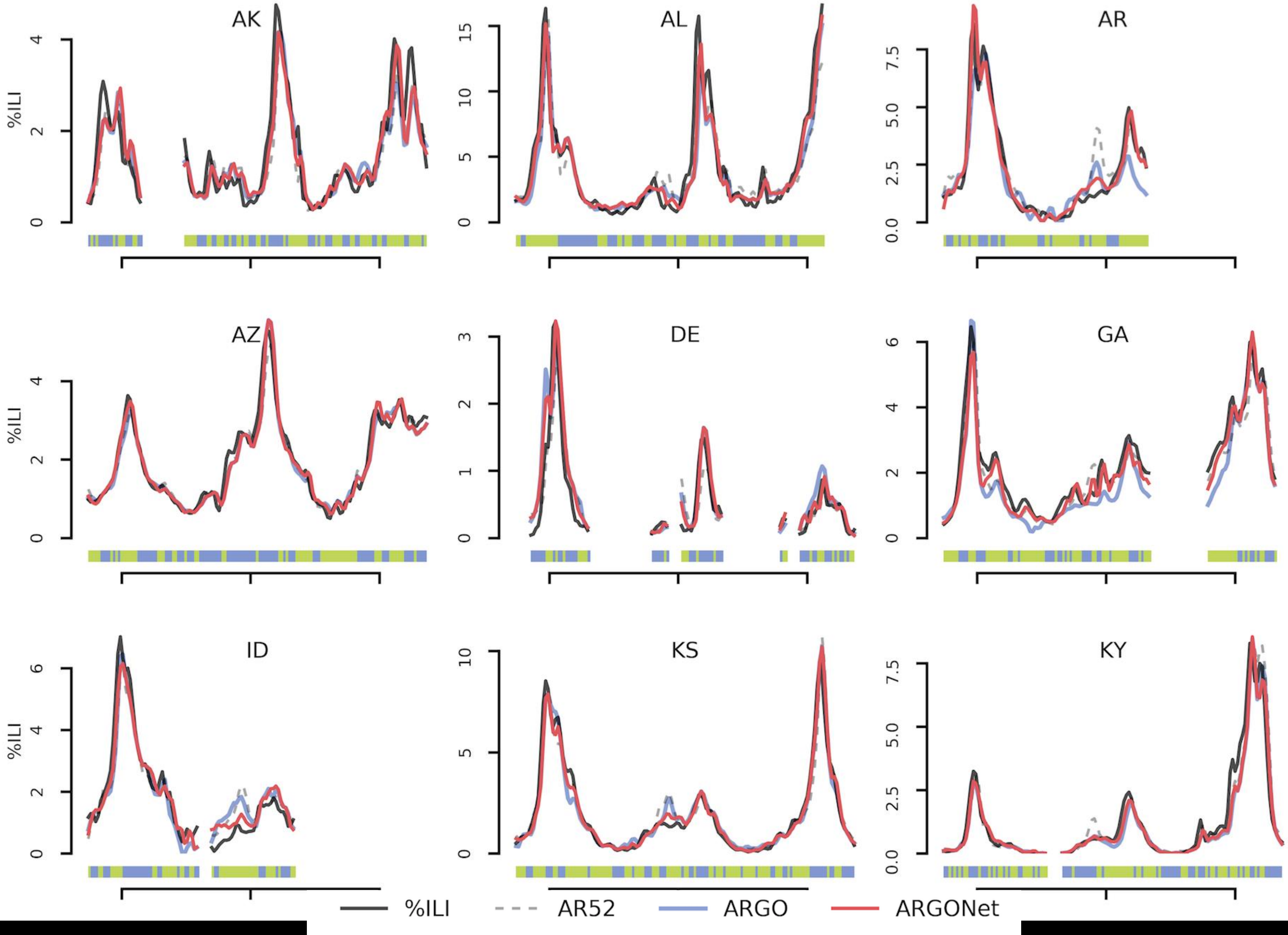
Flu-related Google search information

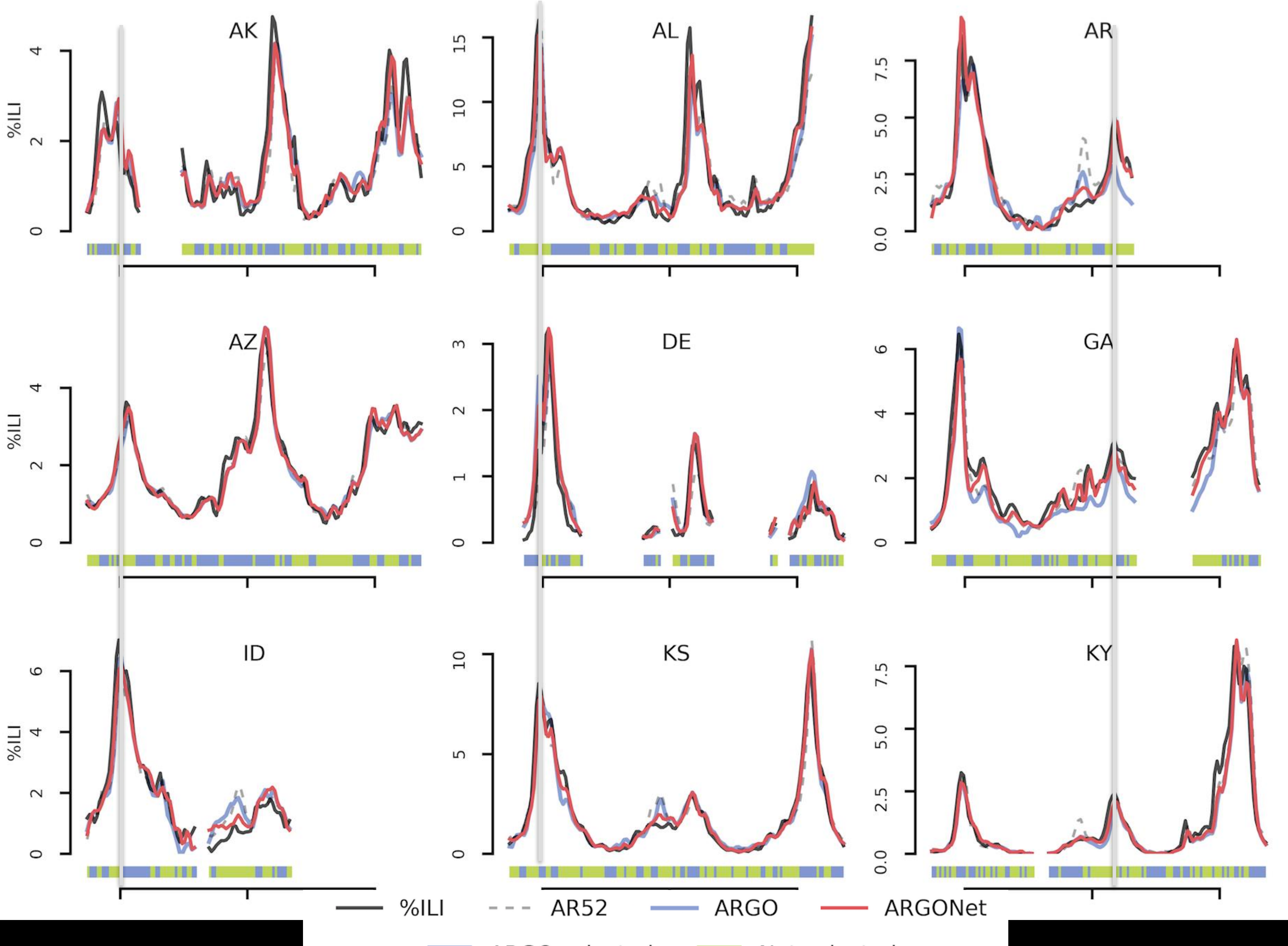
+

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
0
2
4
6
5
10
15
0.0
2.5
5.0
7.5
1
3
%ILI
AR52
ARGO
ARGONet
ARGO selected
Net selected


AK
AL
AR
AZ
DE
GA
ID
KS
KY
%ILI
%ILI
%ILI
0
2
4
0
5
10
15
0.0
2.5
5.0
7.5
0
2
4
0
1
2
3
0
2
4
6
0
2
4
6
0
5
10
0.0
2.5
5.0
7.5
%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

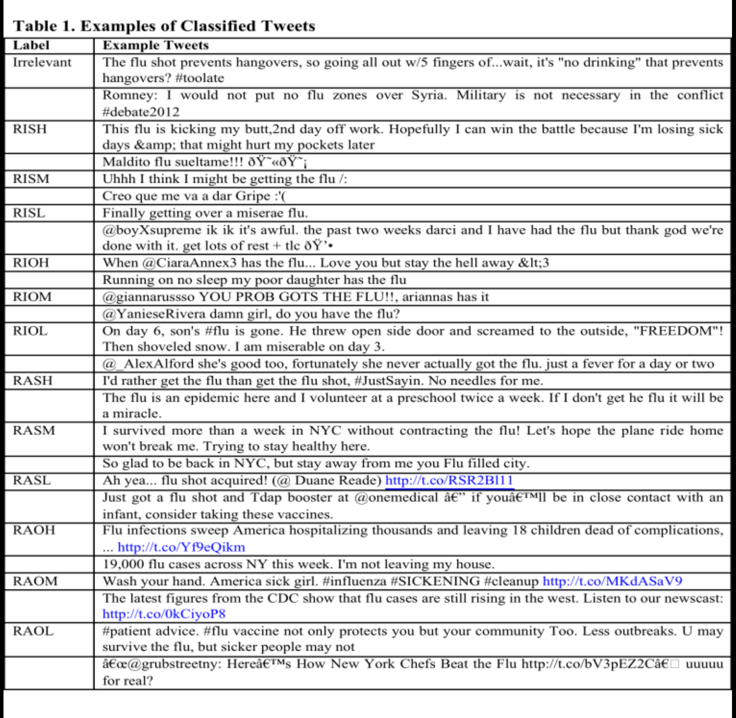
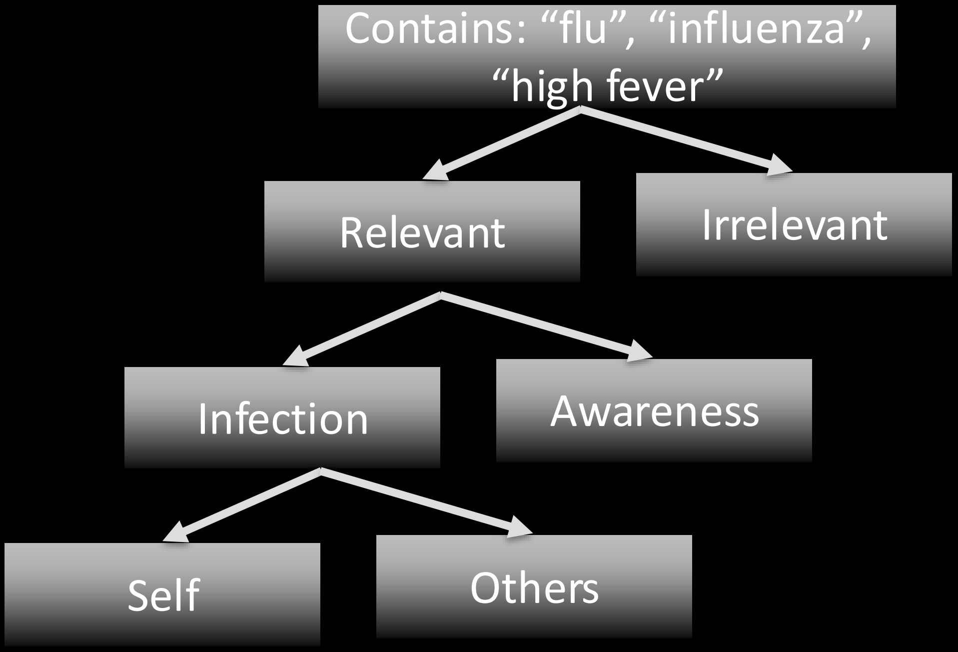


Table 1. Examples of Classified Tweets

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/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
	â€œ@grubstreetny: Hereâ€™s How New York Chefs Beat the Flu http://t.co/bV3pEZ2Câ€□ uuuuu for real?

Categories



Contains: “flu”, “influenza”, “high fever”

Relevant

Irrelevant

Infection

Awareness

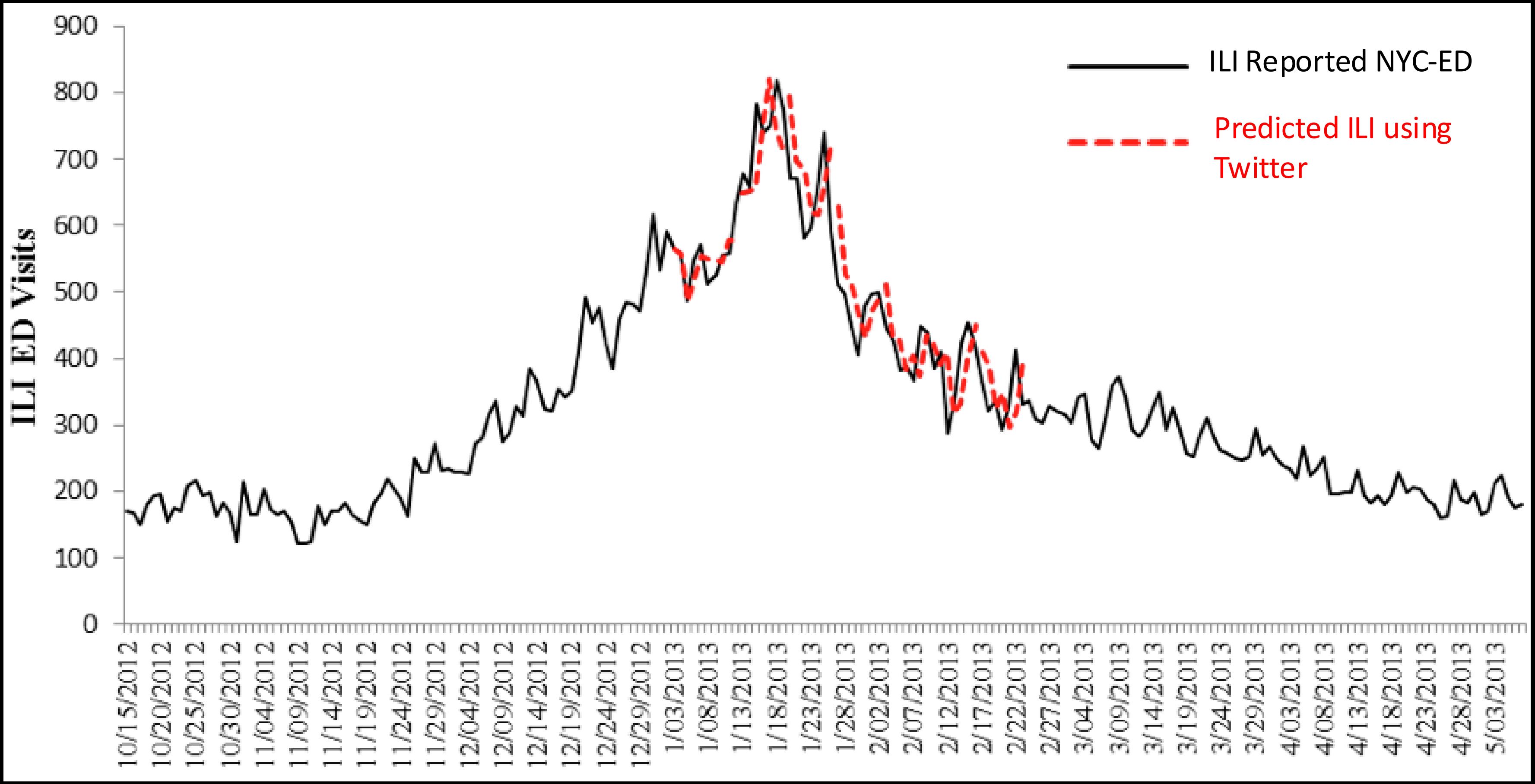
Self

Others

First experiment: was done by hand…

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

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

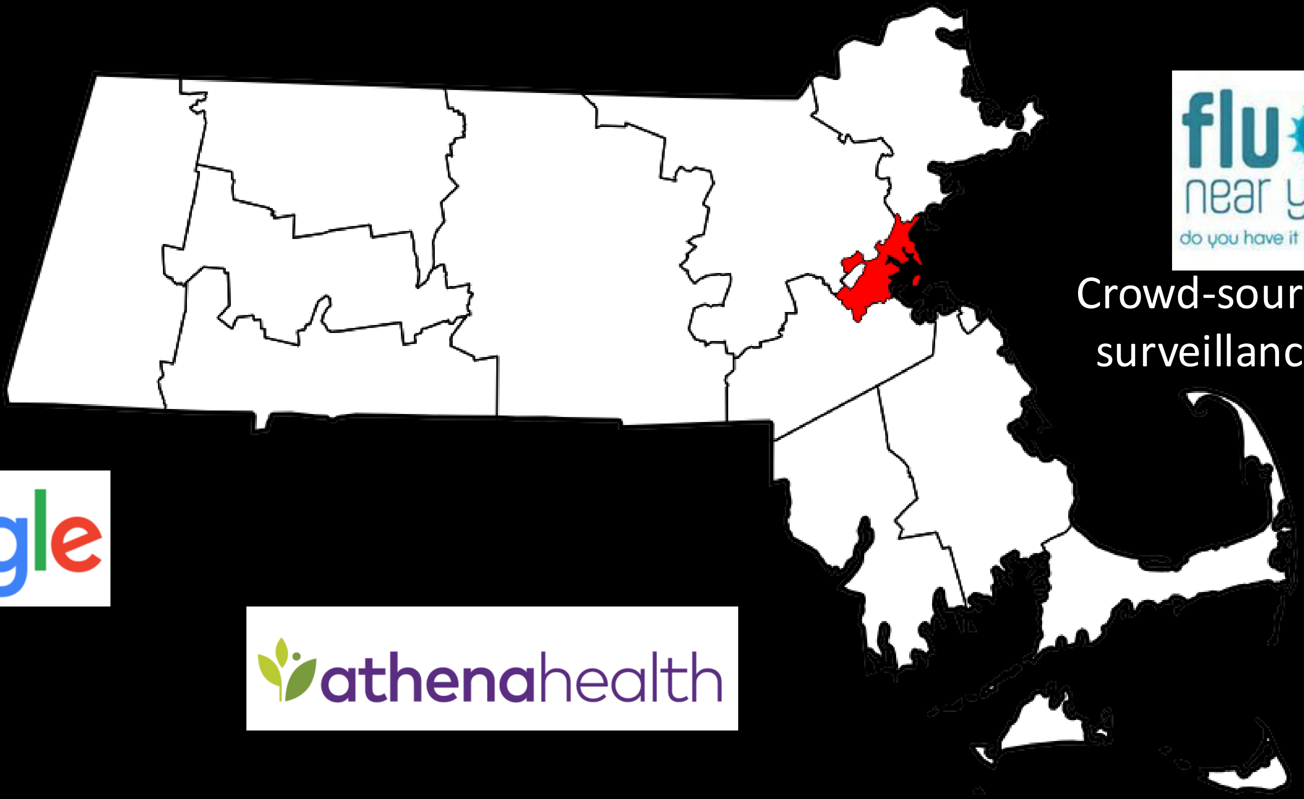


Nagar et al. (2014) Journal of Medical Internet Research. In press

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

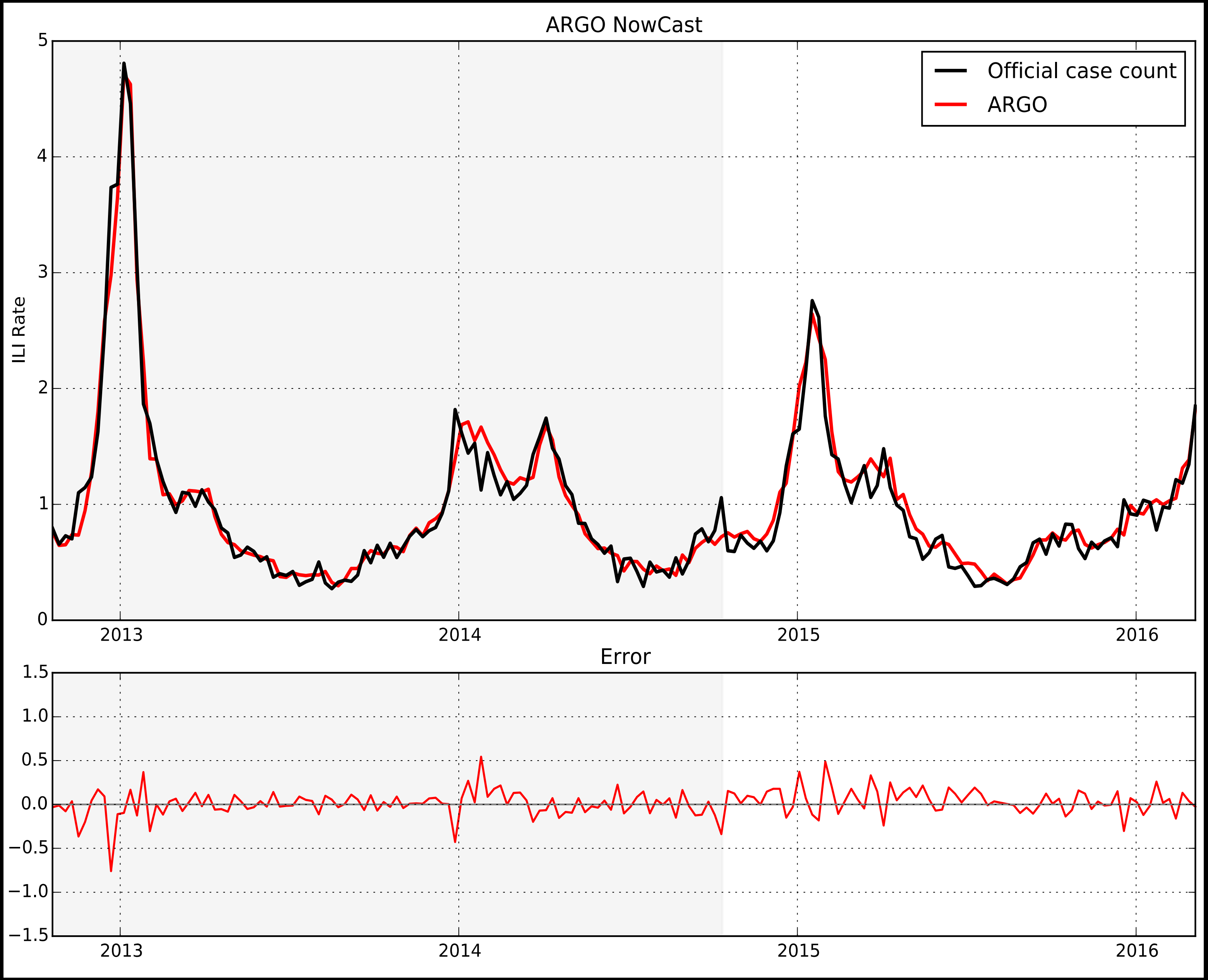
Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis

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

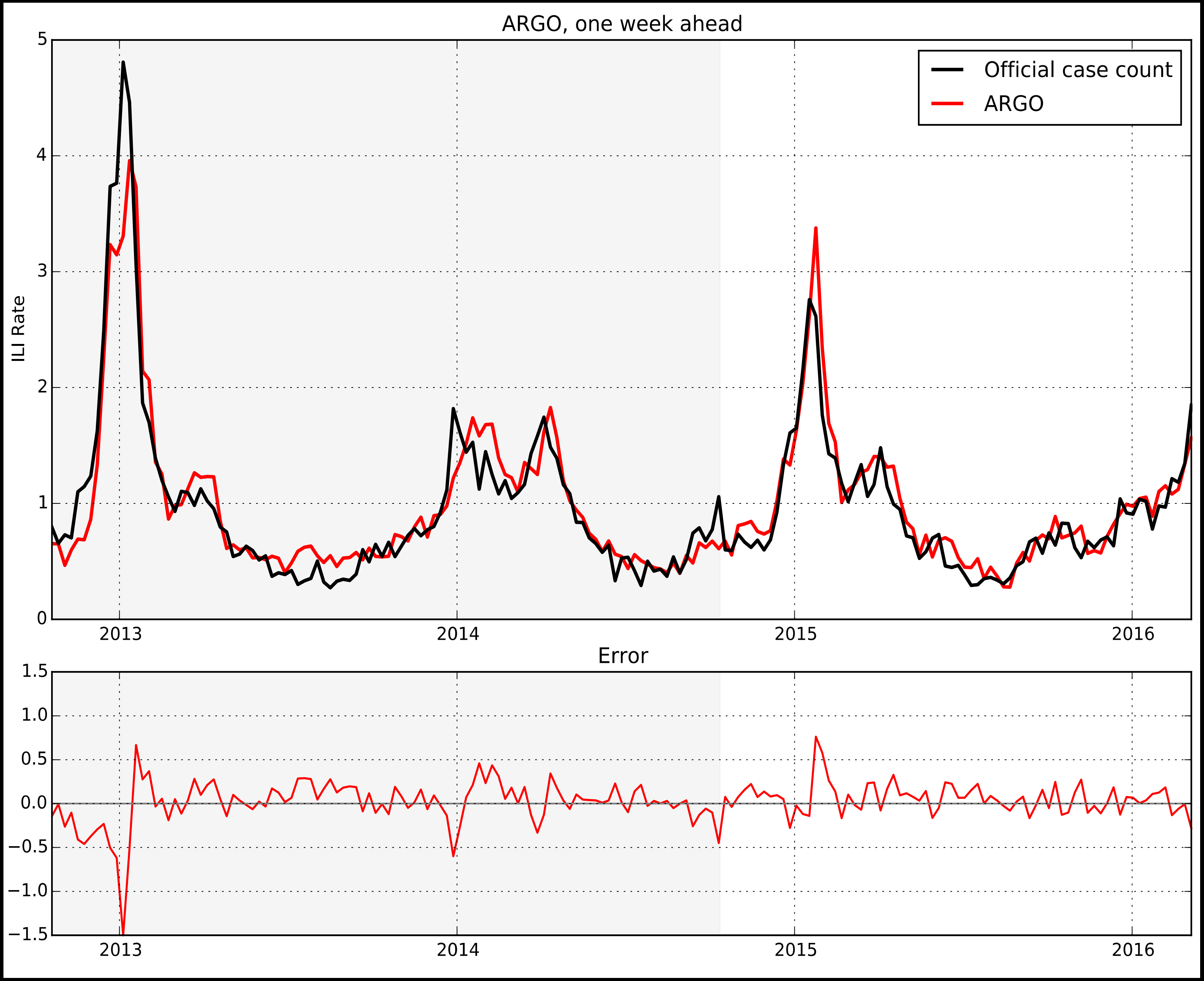
Corresponding Author:
Mauricio Santillana, MS, PhD

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



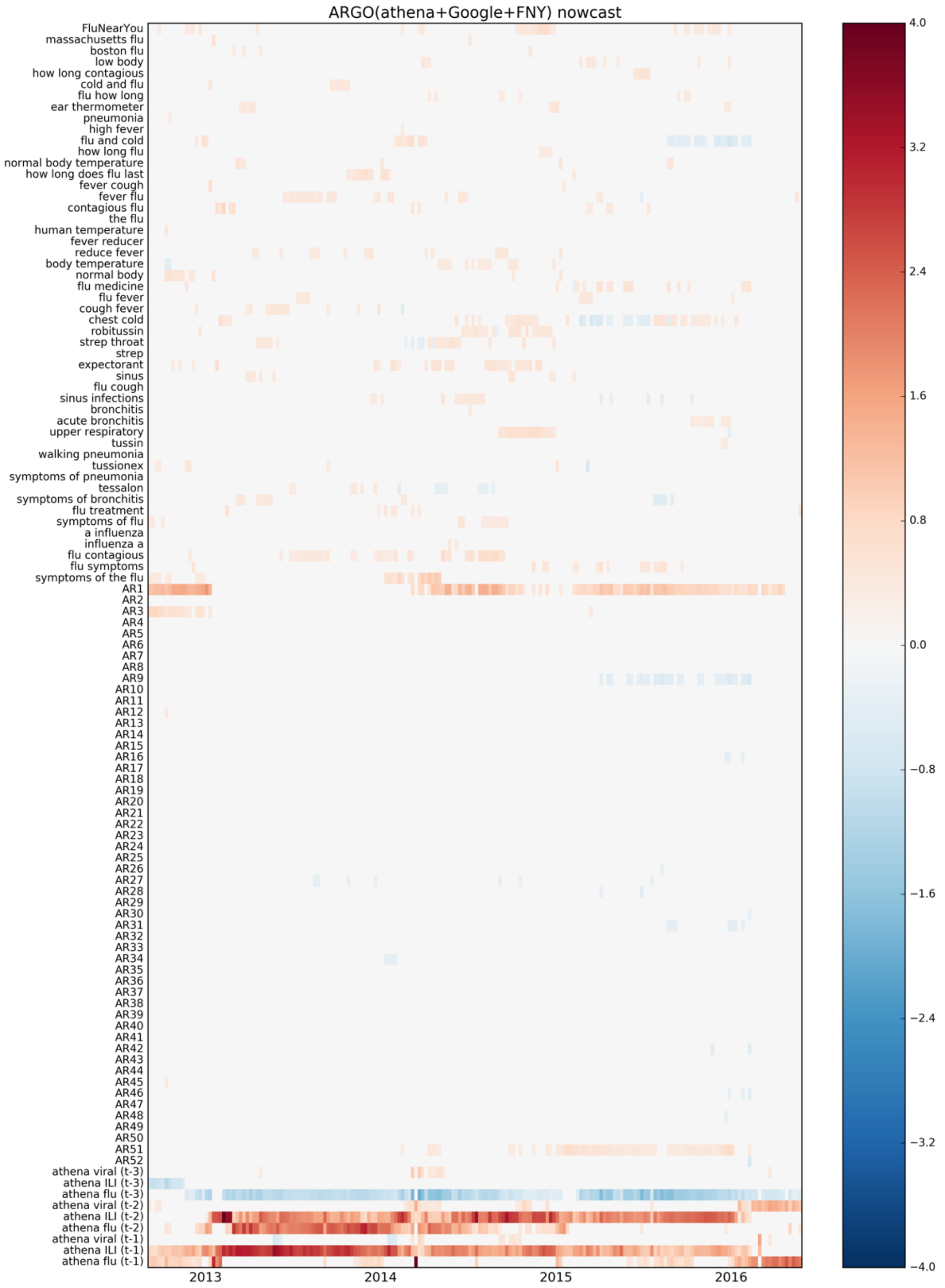
Plots produced by Fred Lu and Suqin Hou

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



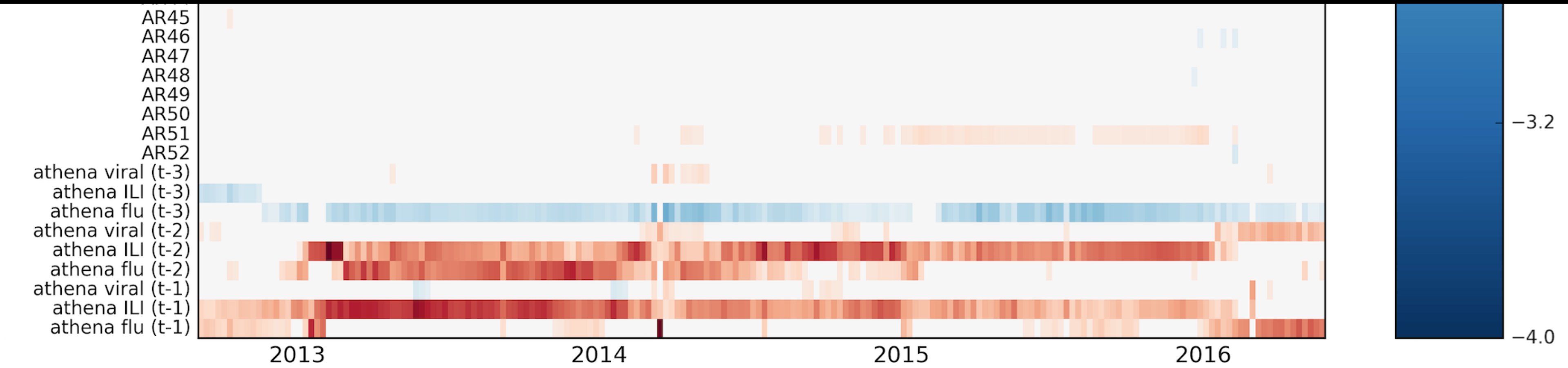
Plots produced by Fred Lu and Suqin Hou

When combined, what are the strongest predictors?



When combined, what are the strongest predictors?

When combined, what are the strongest predictors?

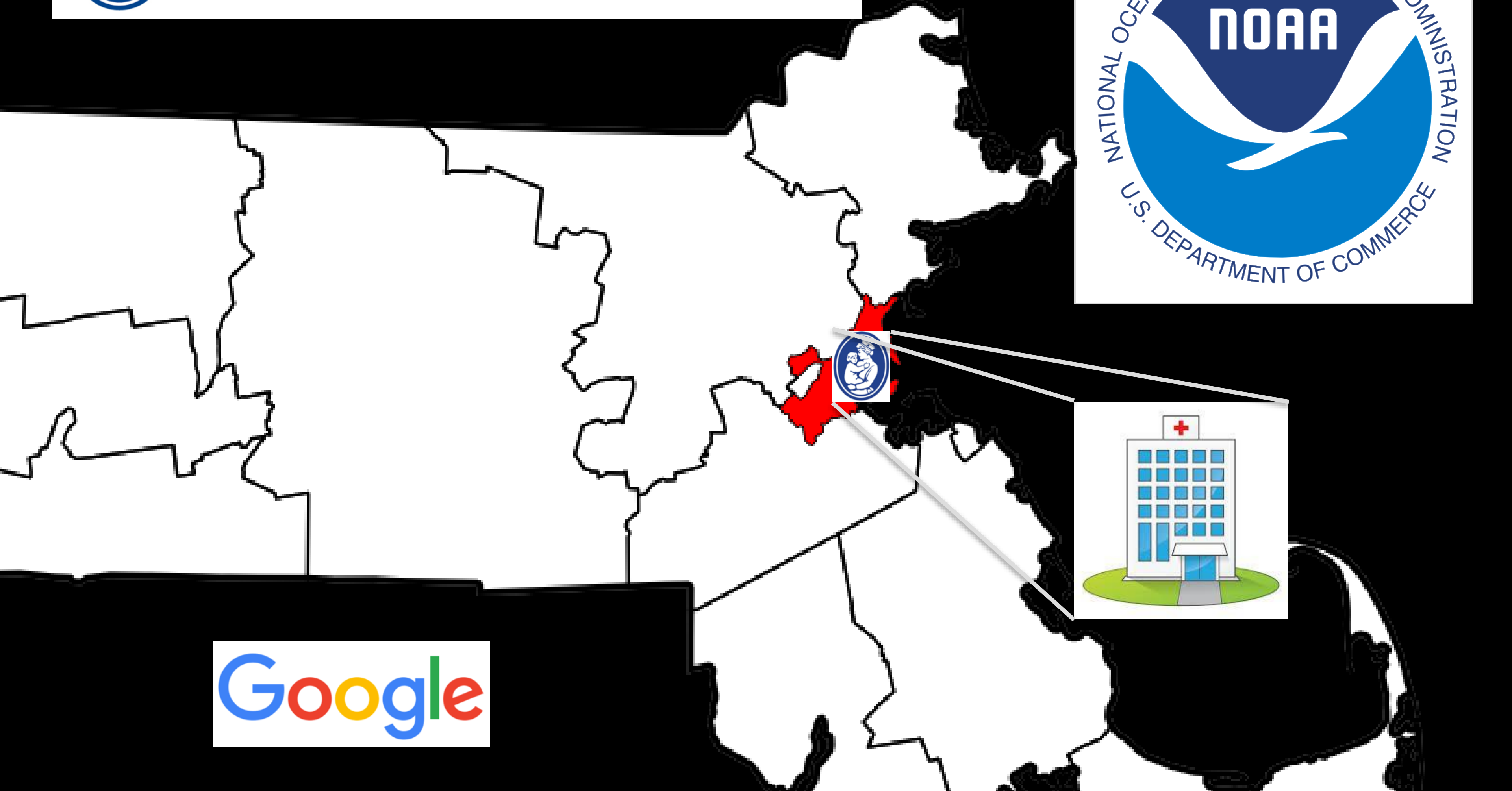


Hyper-local predictions

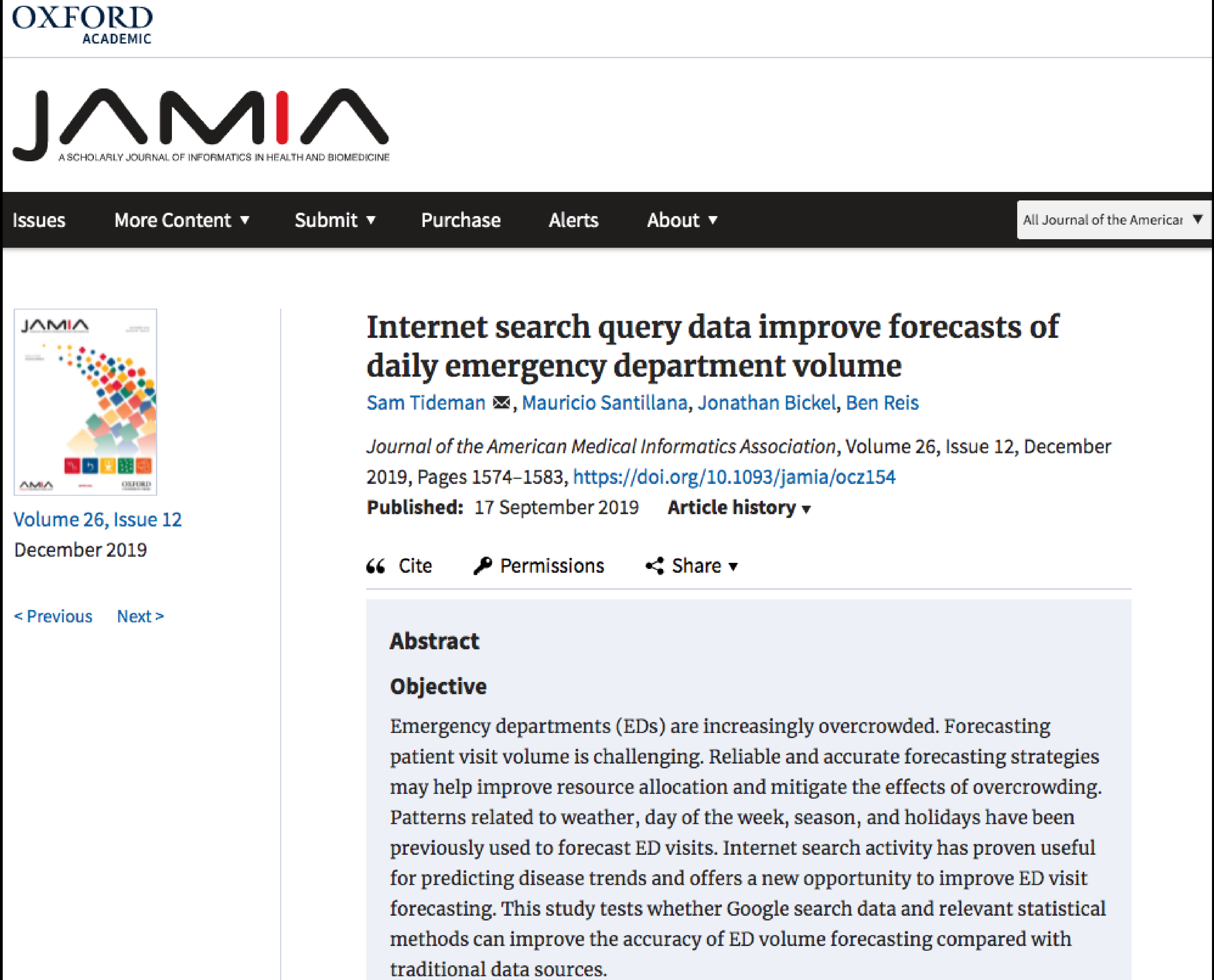
Can we predict daily emergency department visits in a hospital?

Boston Children's Hospital
Until every child is well℠

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NOAA
U.S. DEPARTMENT OF COMMERCE



Google



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

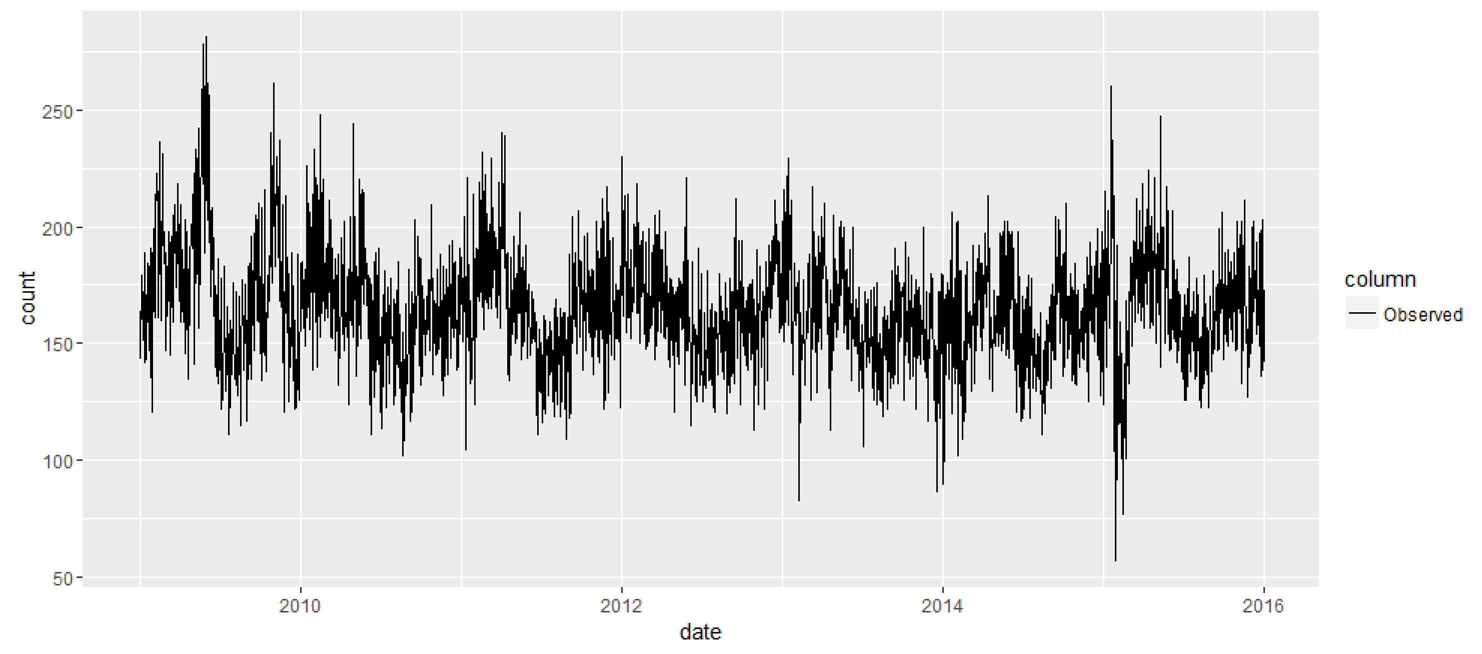
Published: 17 September 2019 **Article history** ▾

Abstract

Objective

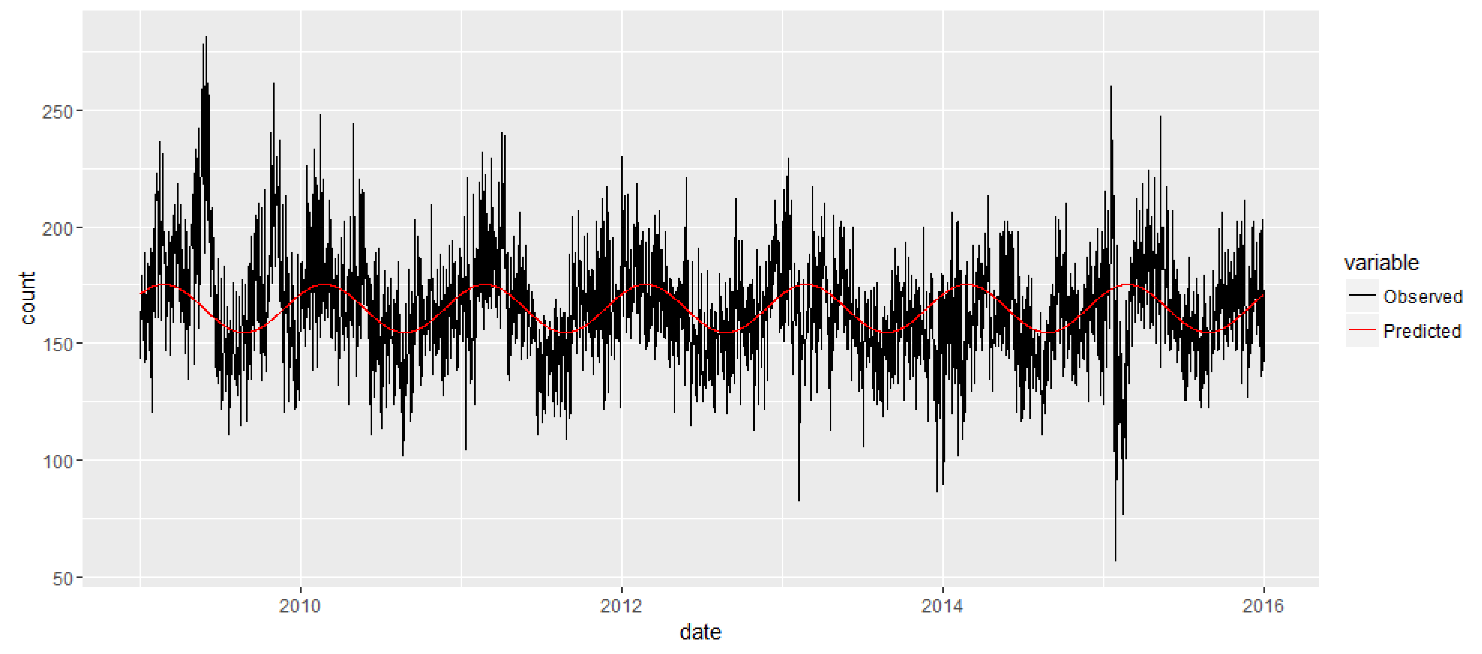
Emergency departments (EDs) are increasingly overcrowded. Forecasting patient visit volume is challenging. Reliable and accurate forecasting strategies may help improve resource allocation and mitigate the effects of overcrowding. Patterns related to weather, day of the week, season, and holidays have 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



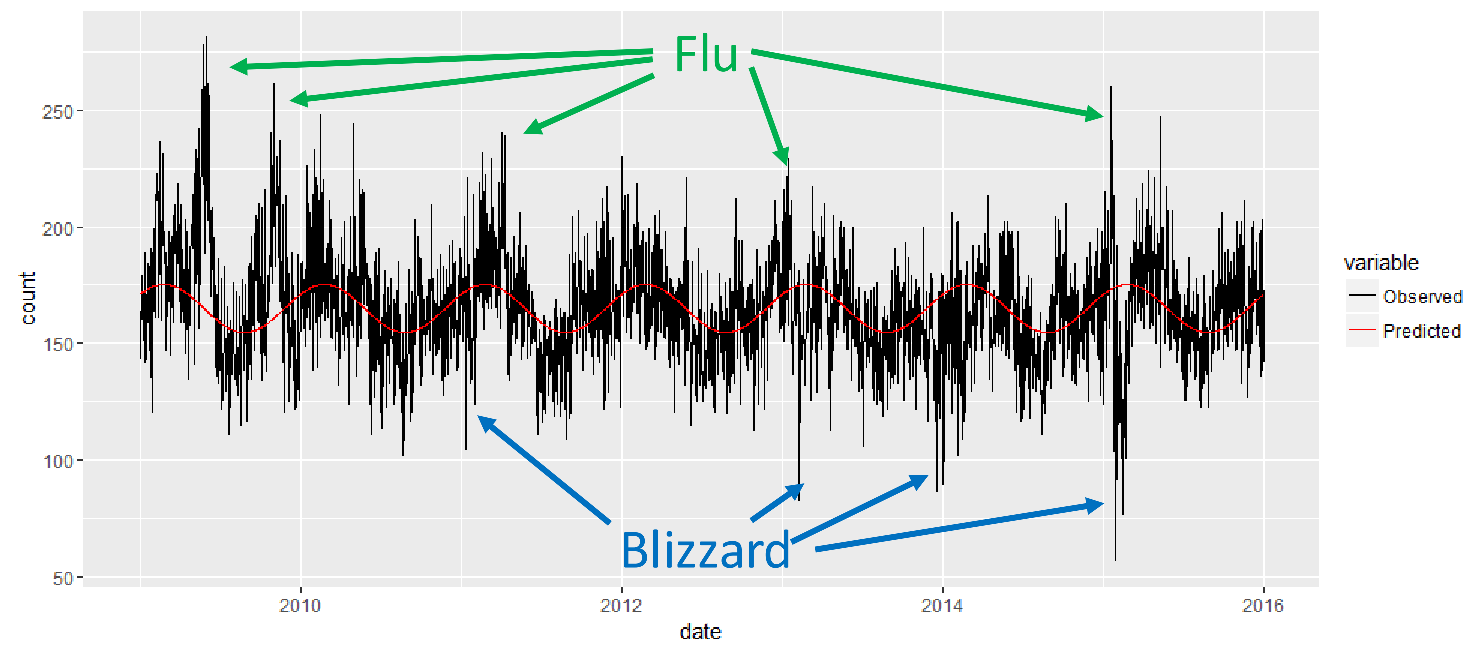
In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Seasonal Trend



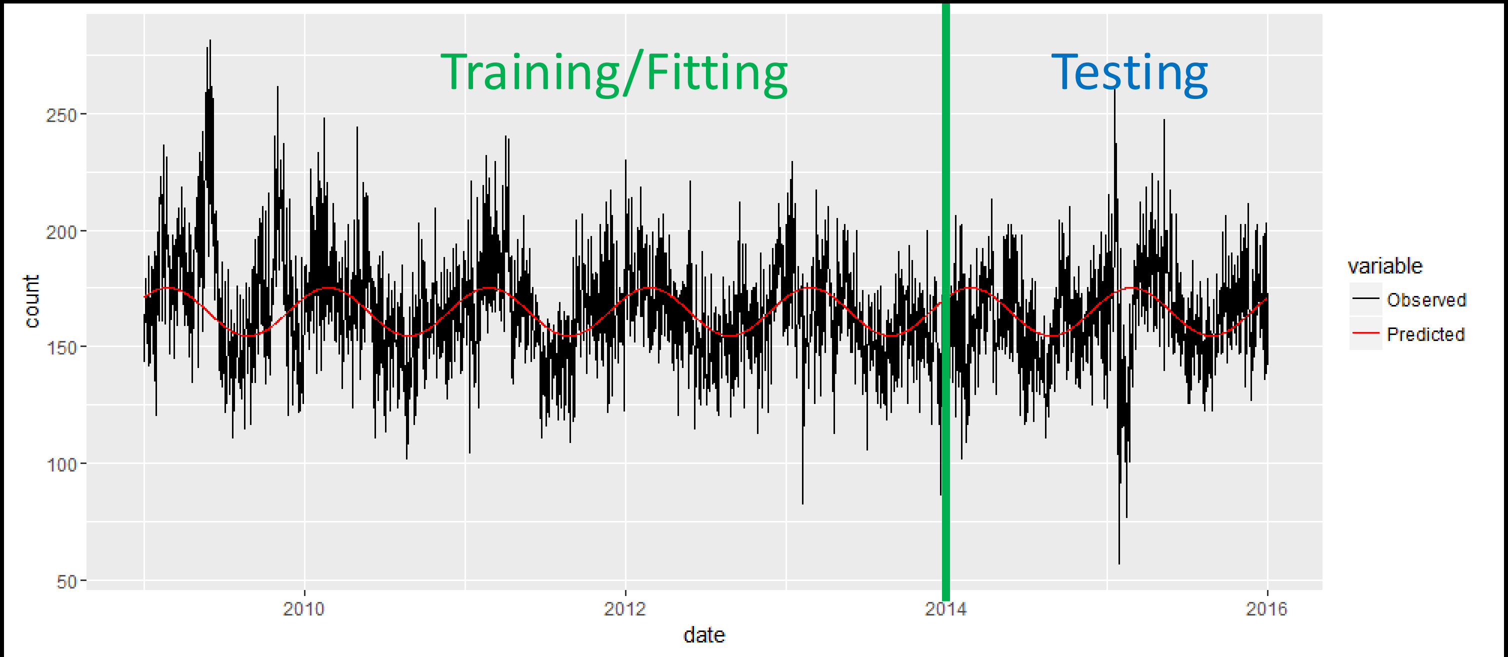
In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Noticeable Events



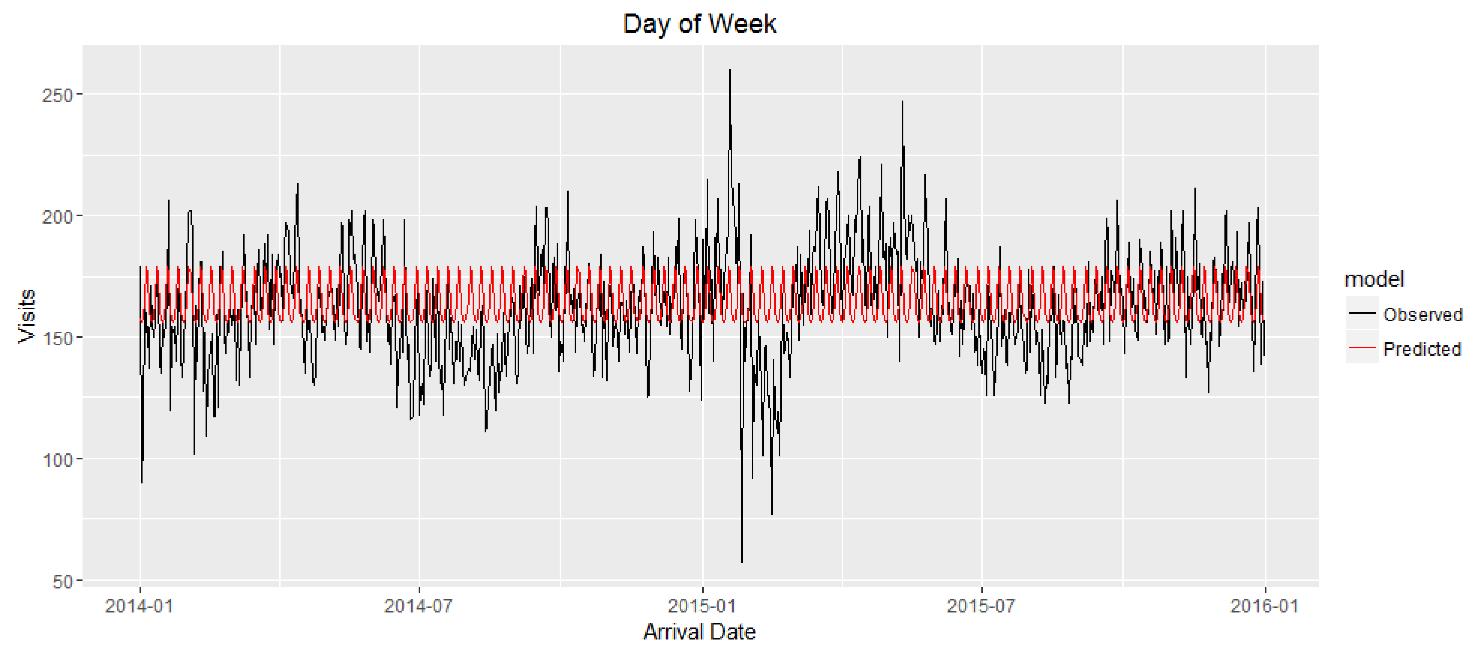
In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Split data for modeling



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

Current Staffing model = Day of Week

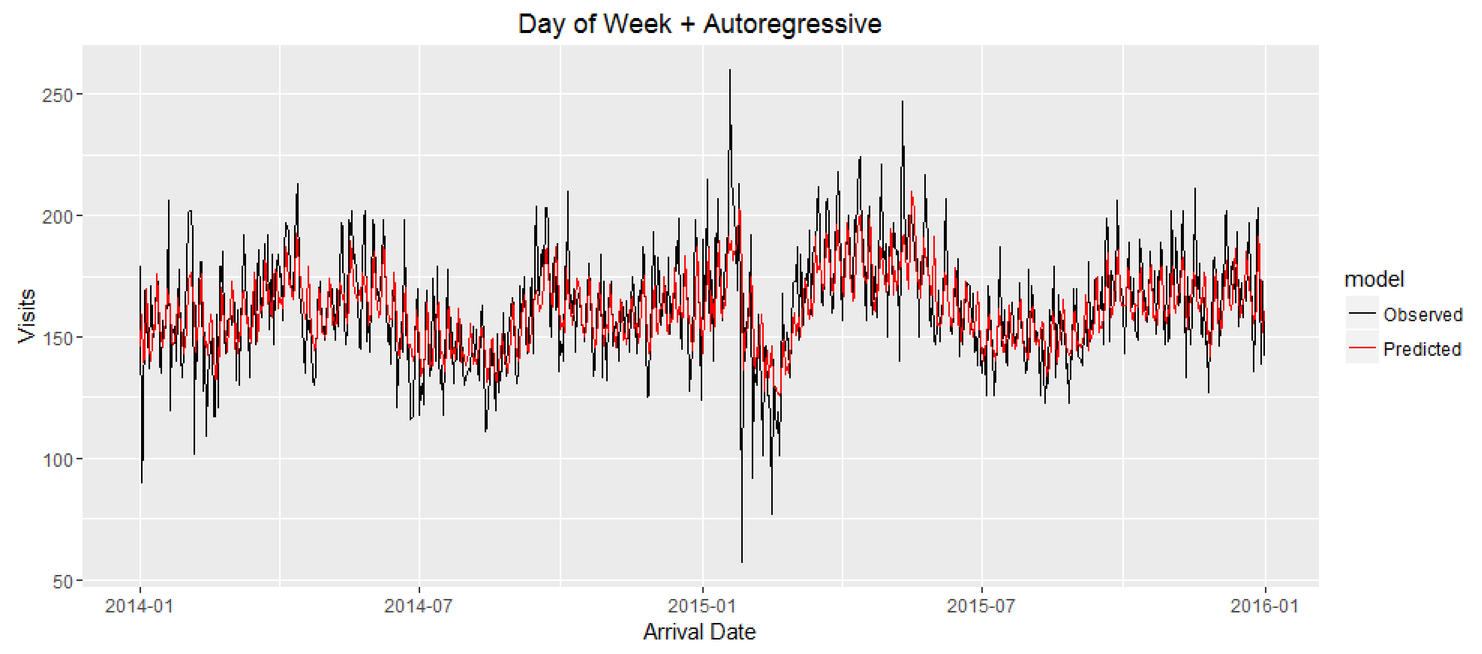


MAPE = 11.0%

Percent of days with bad staffing= 11.2%

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

Add in Auto regression

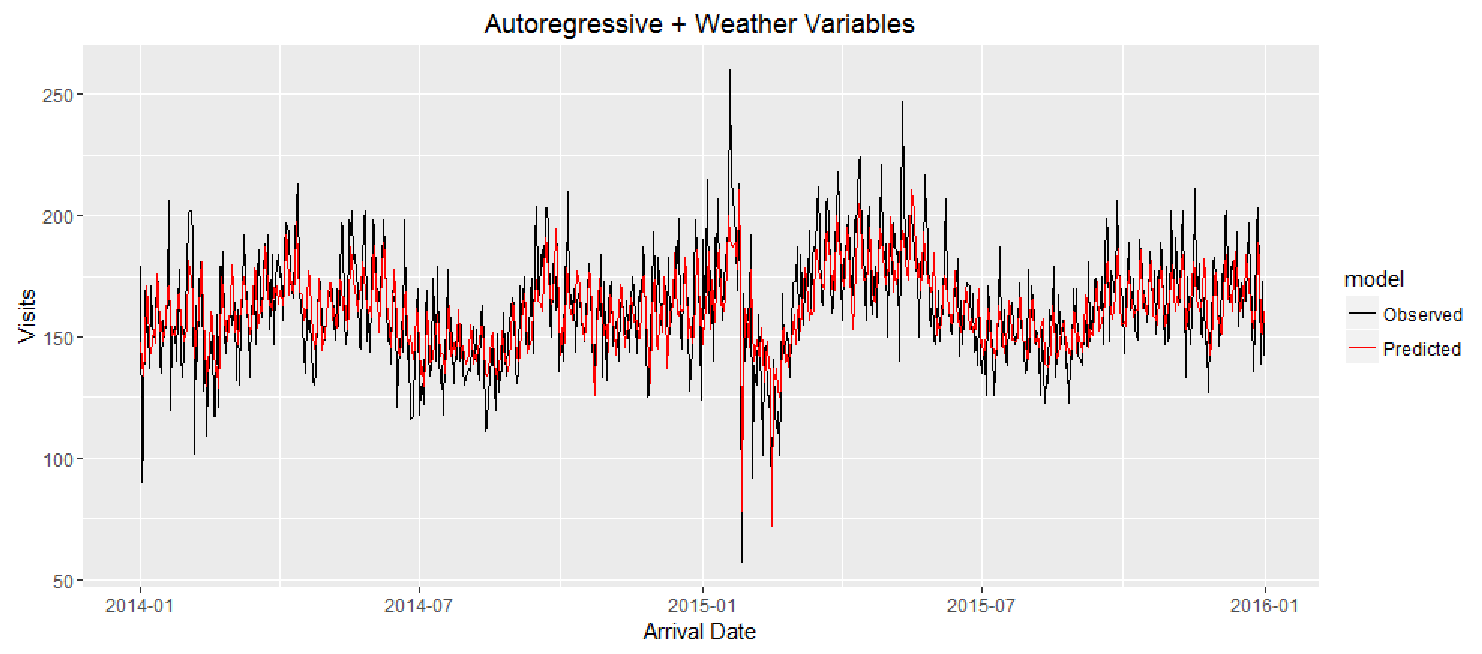


MAPE = 8.4%

Percent of days with bad staffing= 4.9%

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

Add in Weather Data

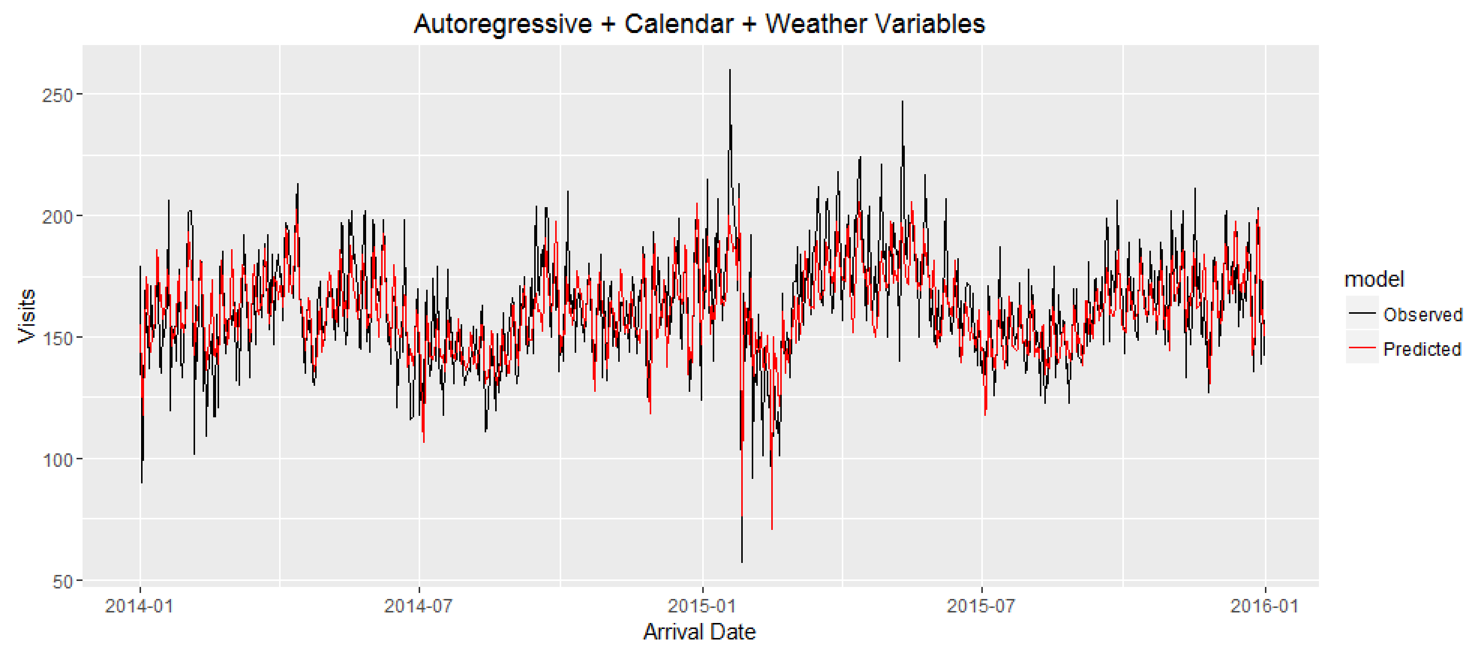


MAPE = 7.9%

Percent of days with bad staffing= 4.8%

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

Add in Calendar Data

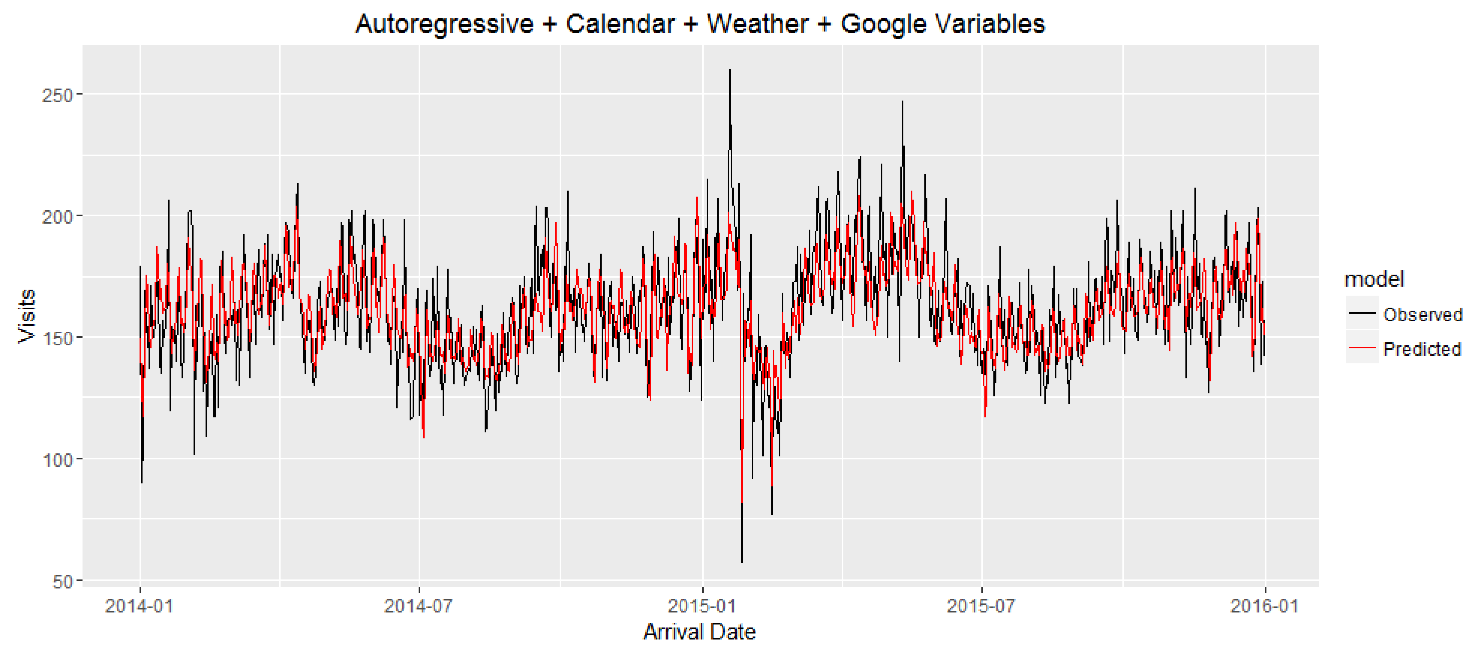


MAPE = 7.7%

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

Article | Open Access | Published: 20 April 2022

Machine learning approaches to predicting no-shows in pediatric medical appointment

Dianbo Liu, Won-Yong Shin, Eli Sprecher, Kathleen Conroy, Omar Santiago, Gal Wachtel & Mauricio Santillana

npj Digital Medicine **5**, Article number: 50 (2022) | Cite this article

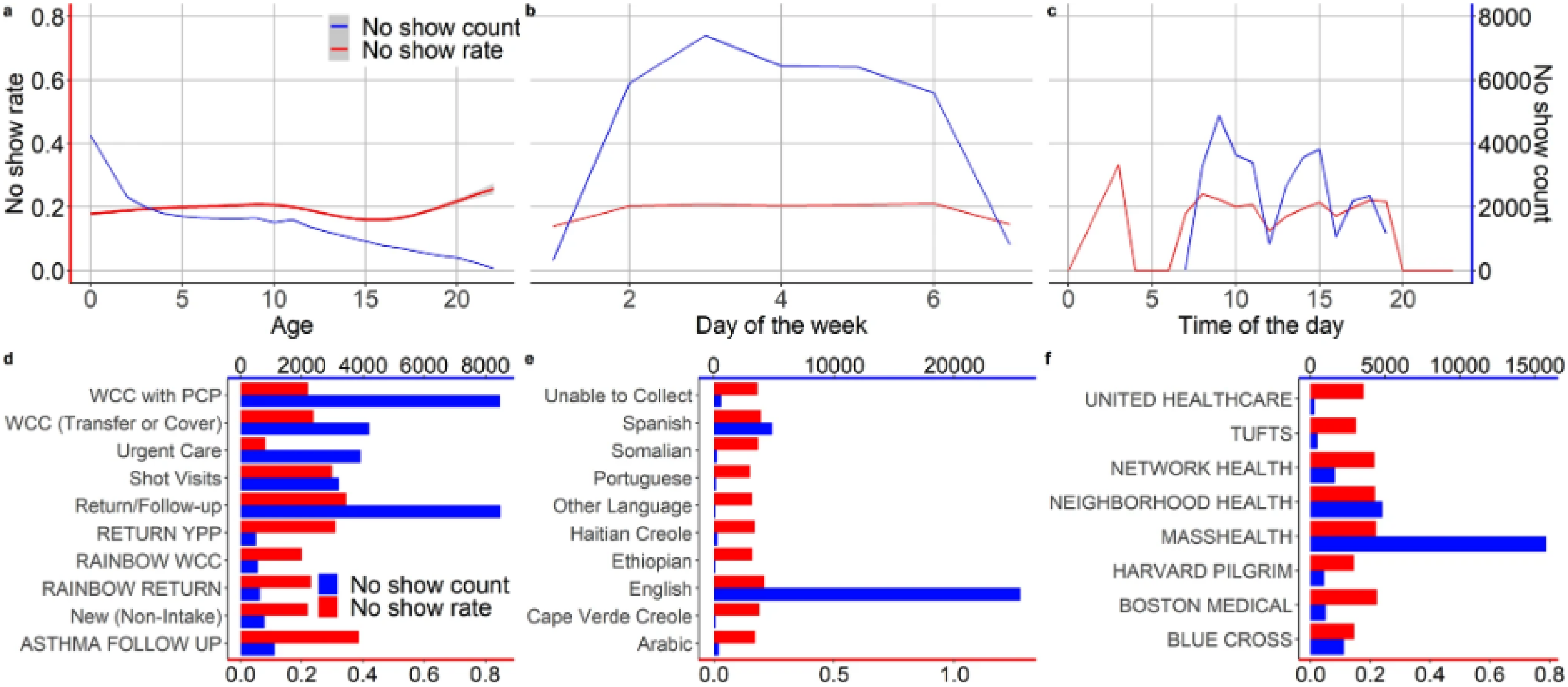
3675 Accesses | **39** Altmetric | Metrics

Abstract

Patients’ no-shows, scheduled but unattended medical appointments, have a direct negative impact on patients’ health, due to discontinuity of treatment and late presentation to care. They also lead to inefficient use of medical resources in hospitals and clinics. The ability to predict a likely no-show in advance could enable the design and implementation of interventions to reduce the risk of it happening, thus improving patients’ care and clinical resource allocation. In this study, we develop a new interpretable deep learning-based approach for predicting the risk of no-shows at the time when a medical appointment is first scheduled. The

Fig. 1: Summary statistics of no-show rates and counts.

From: [Machine learning approaches to predicting no-shows in pediatric medical appointment]()



a Distribution of medical appointment no-show rates and counts over ages at a pediatric primary care clinic in a major academic Children's Hospital in the U.S. (patients above age 18 were excluded from our analysis). **b** Distribution of no-show rates and counts over days of the week. **c** Distribution of no-show rates and counts over hours of day. **d** Distribution of no-show rates and counts over visit types. **e** Distribution of no-show rates and counts over spoken languages (languages were excluded from the predictive model training). **f** Distribution of no-show rates and counts over health insurance types.

Fig. 2: Feature importance in predicting no-shows.

From: Machine learning approaches to predicting no-shows in pediatric medical appointment

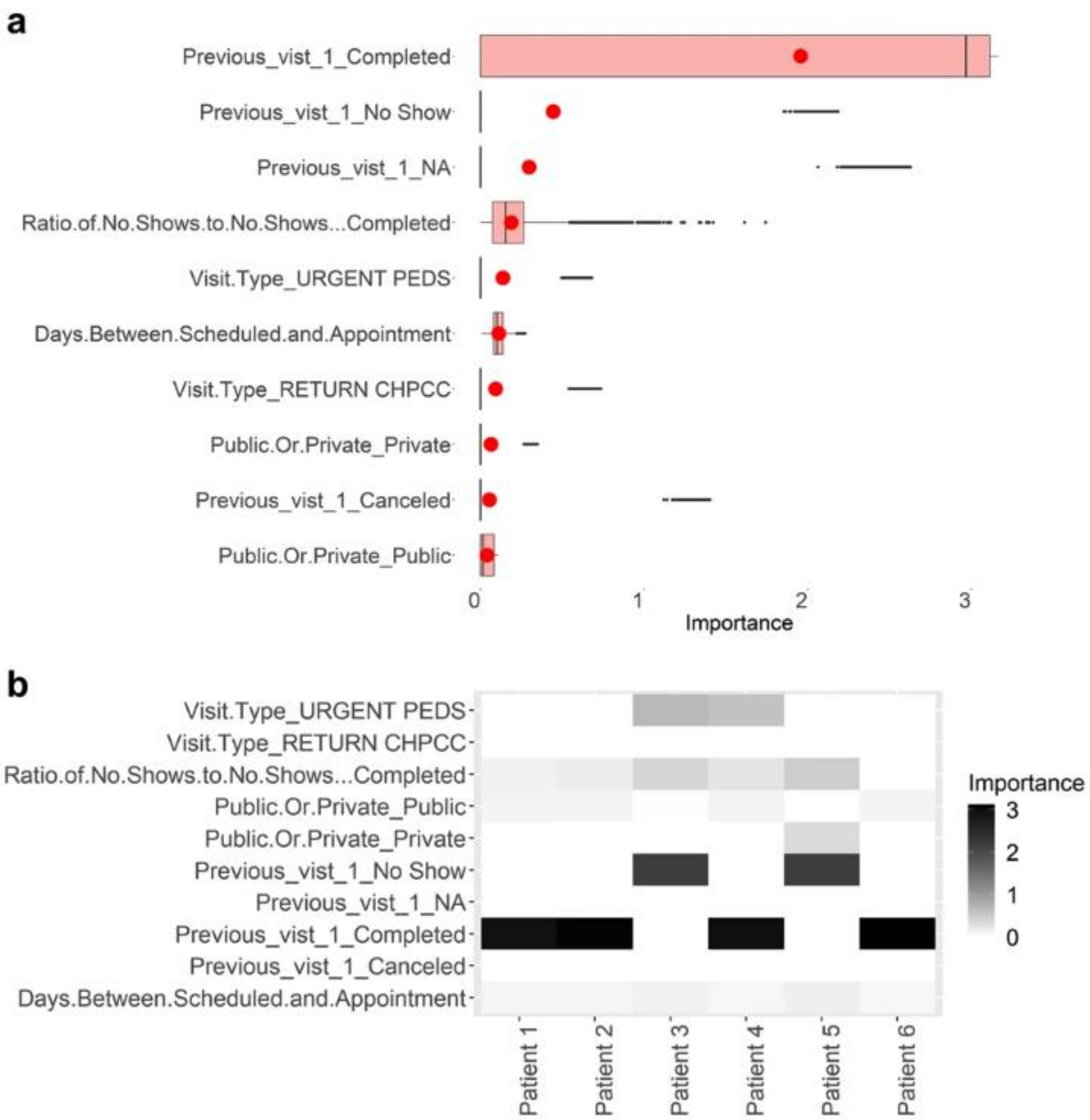
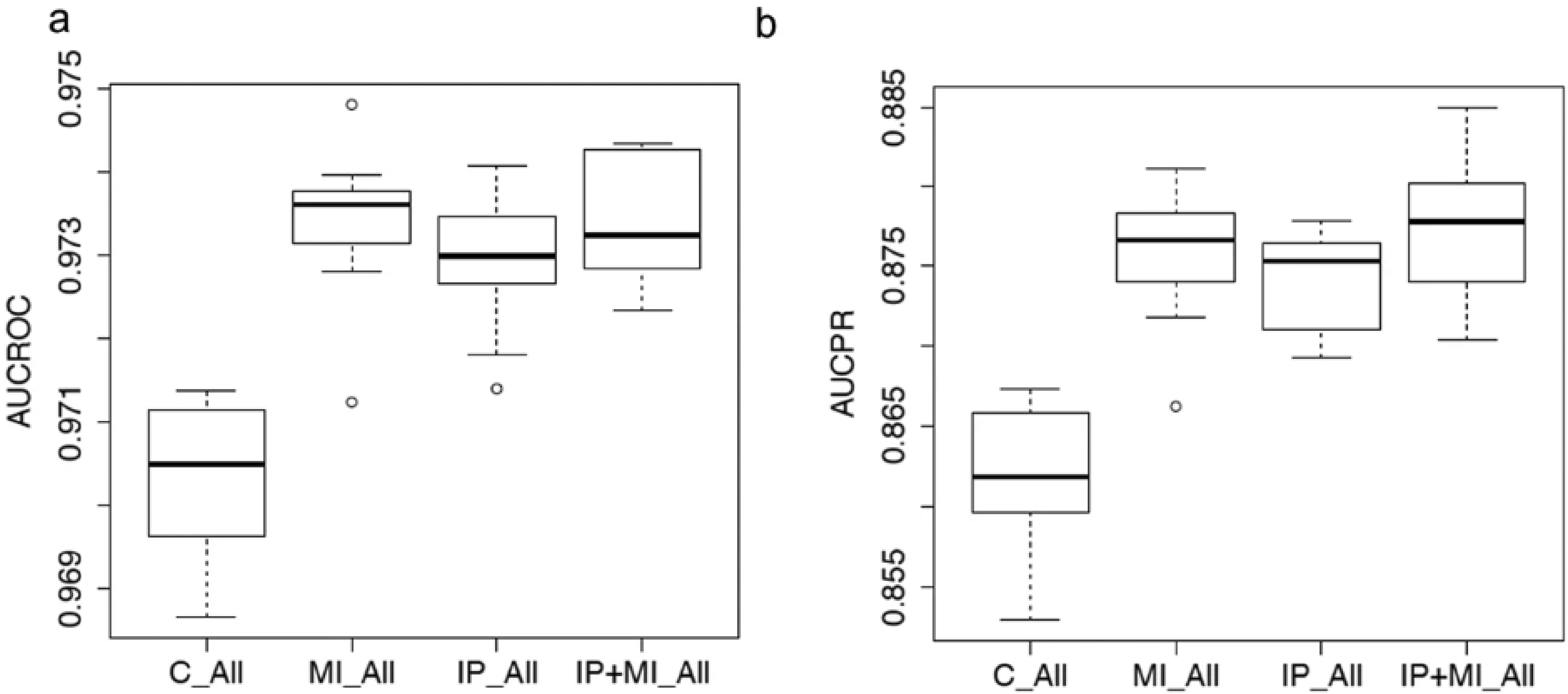
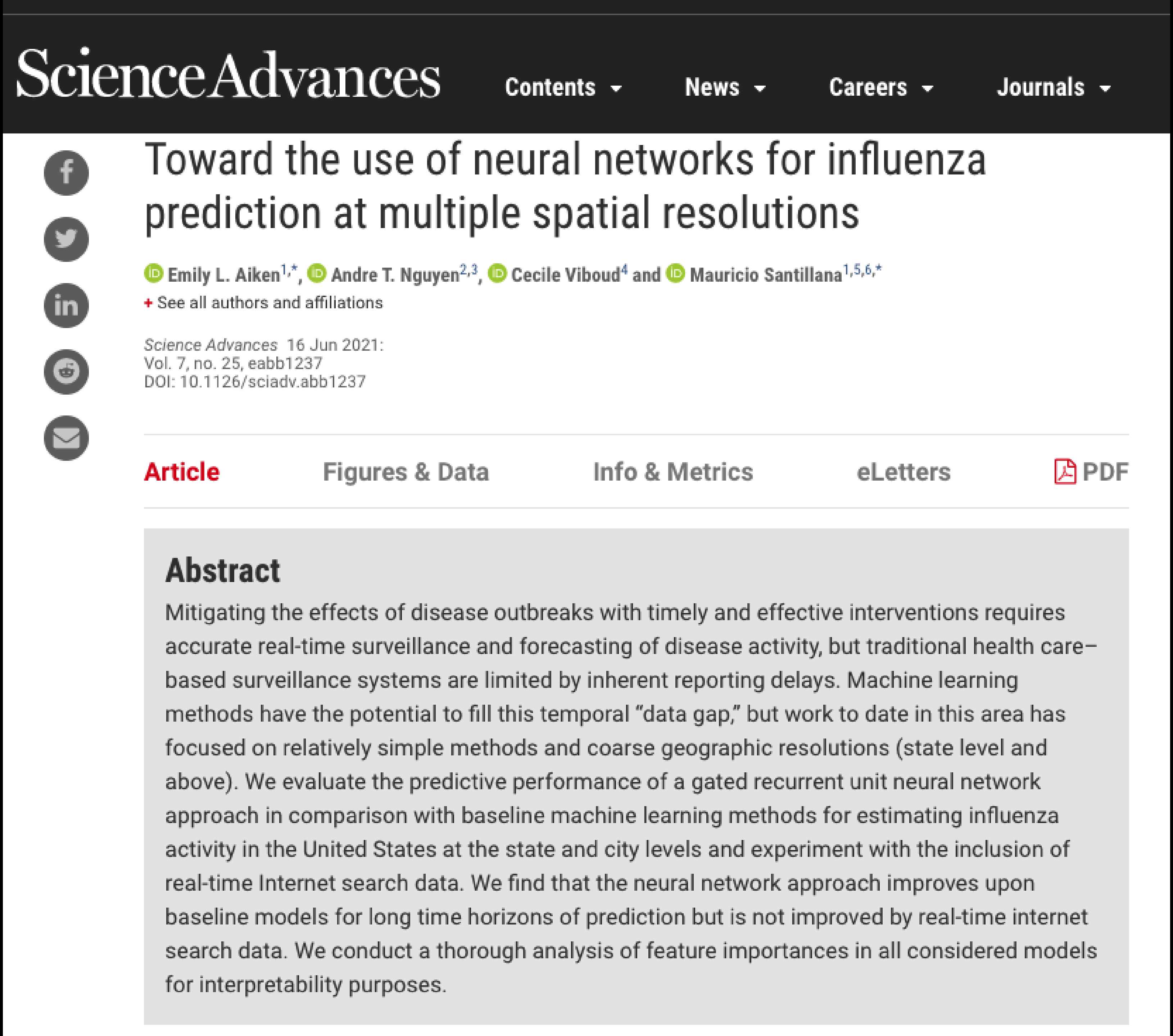


Fig. 4: Performance of deep learning models predicting no-shows with different approaches to handle missing information.

From: Machine learning approaches to predicting no-shows in pediatric medical appointment



AUROC and AUPRC of our prediction model trained on the imputed data (IP) and both the imputed data and the missing data indicator (IP + MI). Here, "C" indicates data without missing information; "All" indicates data including all patients with or without missing information; "C_All" represents the model trained only on complete data without missing information and tested on data including all patients; and other models such as "MI_All", "IP_All", and "IP + MI_All" are defined in a similar fashion.



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

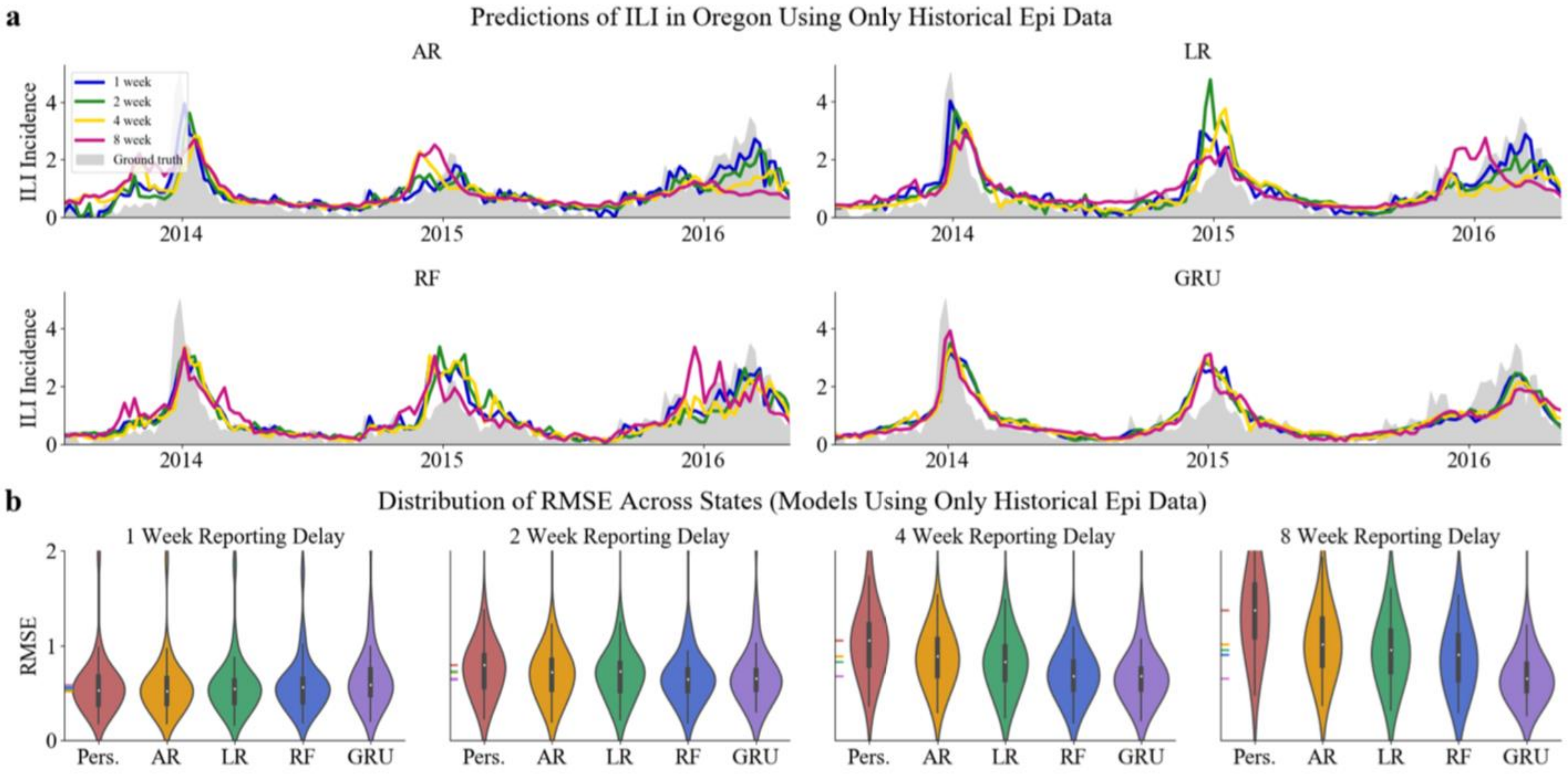
Emily L. Aiken[1,*], Andre T. Nguyen[2,3], Cecile Viboud[4] and Mauricio Santillana[1,5,6,*]
+ See all authors and affiliations

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

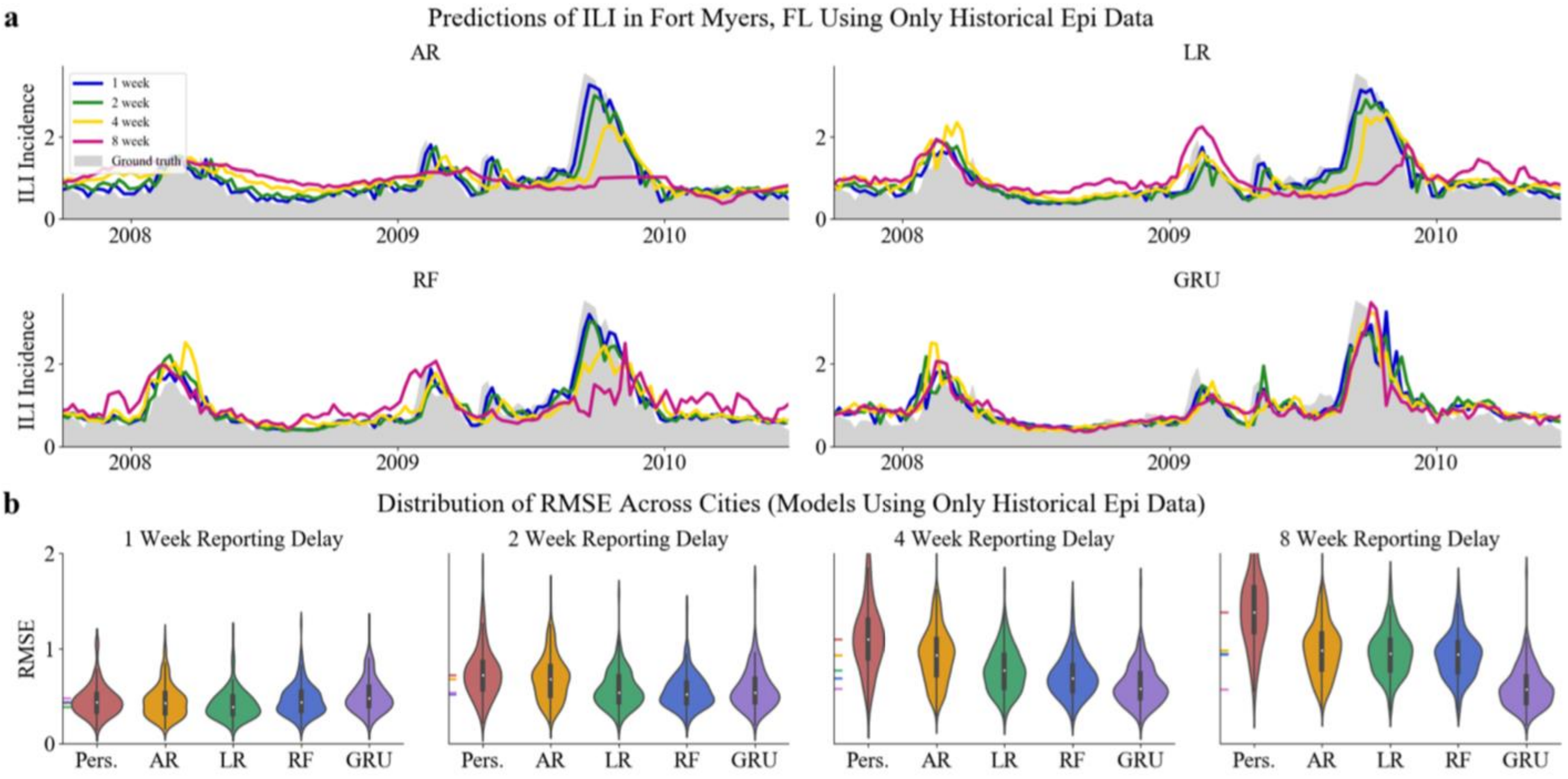
Abstract

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

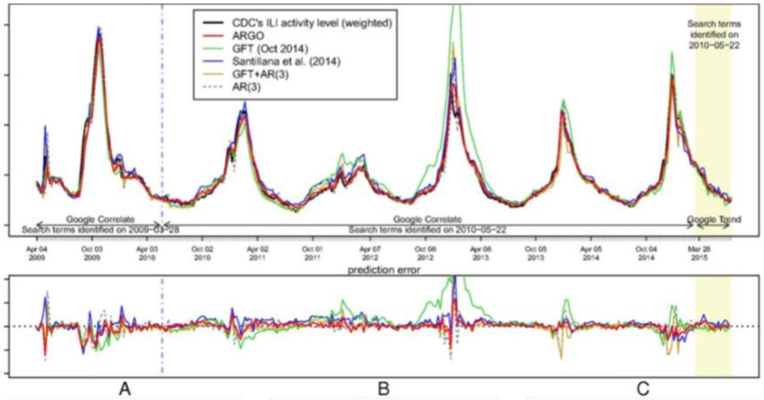


Super-ensemble Seasonal Influenza Forecasting with Machine Learning Augmented Mechanistic Models

Xinyue Xiong, Qian Zhang, Fred S. Lu, Mauricio Santillana, Alessandro Vespignani

BACKGROUND

Statistical Machine Learning Models



Statistical machine learning models that harness internet-based information have exhibited desirable prediction accuracy of now-casting.

Mechanistic Models

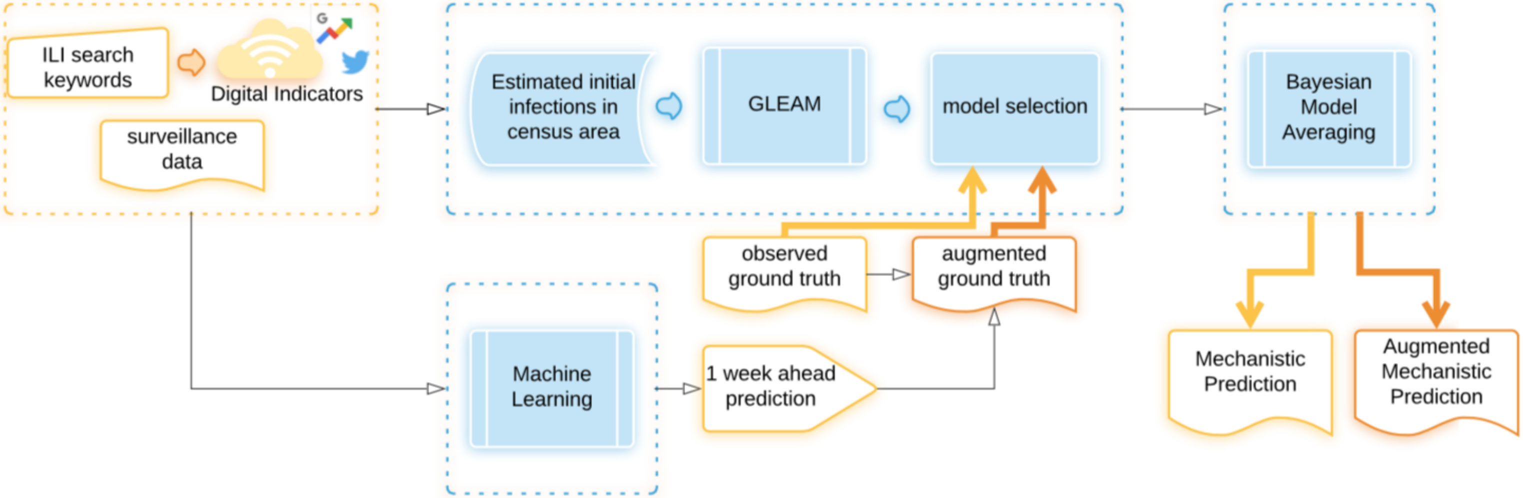
The mechanistic model Global Epidemic And Mobility (GLEAM) have demonstrated the strong power of inferring key epidemiological parameters.

To improve the accuracy and reliability of our GLEAM model without losing the power of predicting epidemiological characteristics, we introduce a super-ensemble forecasting framework.

Super-ensemble Seasonal Influenza Forecasting with Machine Learning Augmented Mechanistic Models

Xinyue Xiong, Qian Zhang, Fred S. Lu, Mauricio Santillana, Alessandro Vespignani

APPROACH



ILI search keywords

Digital Indicators

surveillance data

Estimated initial infections in census area

GLEAM

model selection

Bayesian Model Averaging

observed ground truth

augmented ground truth

Machine Learning

1 week ahead prediction

Mechanistic Prediction

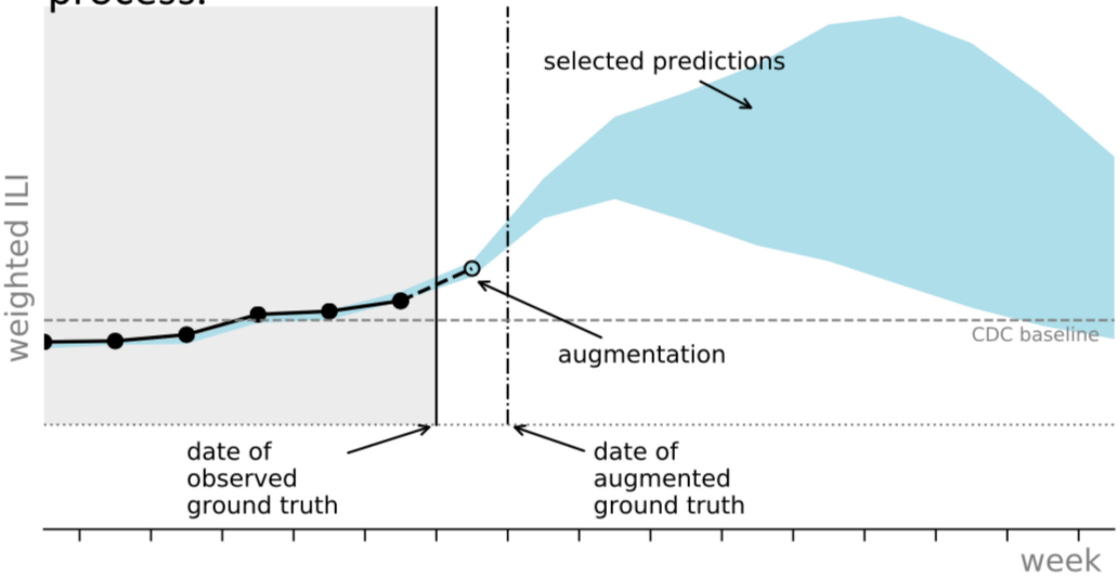
Augmented Mechanistic Prediction

Forecasting Framework: We collect surveillance data and digital surrogates first to estimate the initial infections in each census area. Next, the initial infection are seeded into the GLEAM simulator to generate a corpus of epidemic models with predicted epidemic profiles. Meanwhile, the initial data are fed into Machine Learning model to generate 1-wlp for ILI, which is used to augment the ground truth ILI rate. The original and augmented ground truth are plugged separately into the model selection to generate two corpus of epidemic models, which are further assembled by Bayesian Model Averaging method to be a Mechanistic super-ensemble and a Augmented Mechanistic super-ensemble.

Super-ensemble Seasonal Influenza Forecasting with Machine Learning Augmented Mechanistic Models

Xinyue Xiong, Qian Zhang, Fred S. Lu, Mauricio Santillana, Alessandro Vespignani

Augmentation: The observed ground truth of weekly ILI rate (dotted black curve) is appended by the 1-wlp (hollow black circle) by ARGO (dashed curve). This augmented ILI is used to compare with all the epidemic profile predictions generated by GLEAM. The ones that are closer to the augmented ground truth are selected (blue shadow) to form an augmented ensemble prediction as the output of model selection process.



GLEAM: A spatial, stochastic and individual based epidemic model utilizing human mobility infrastructures.

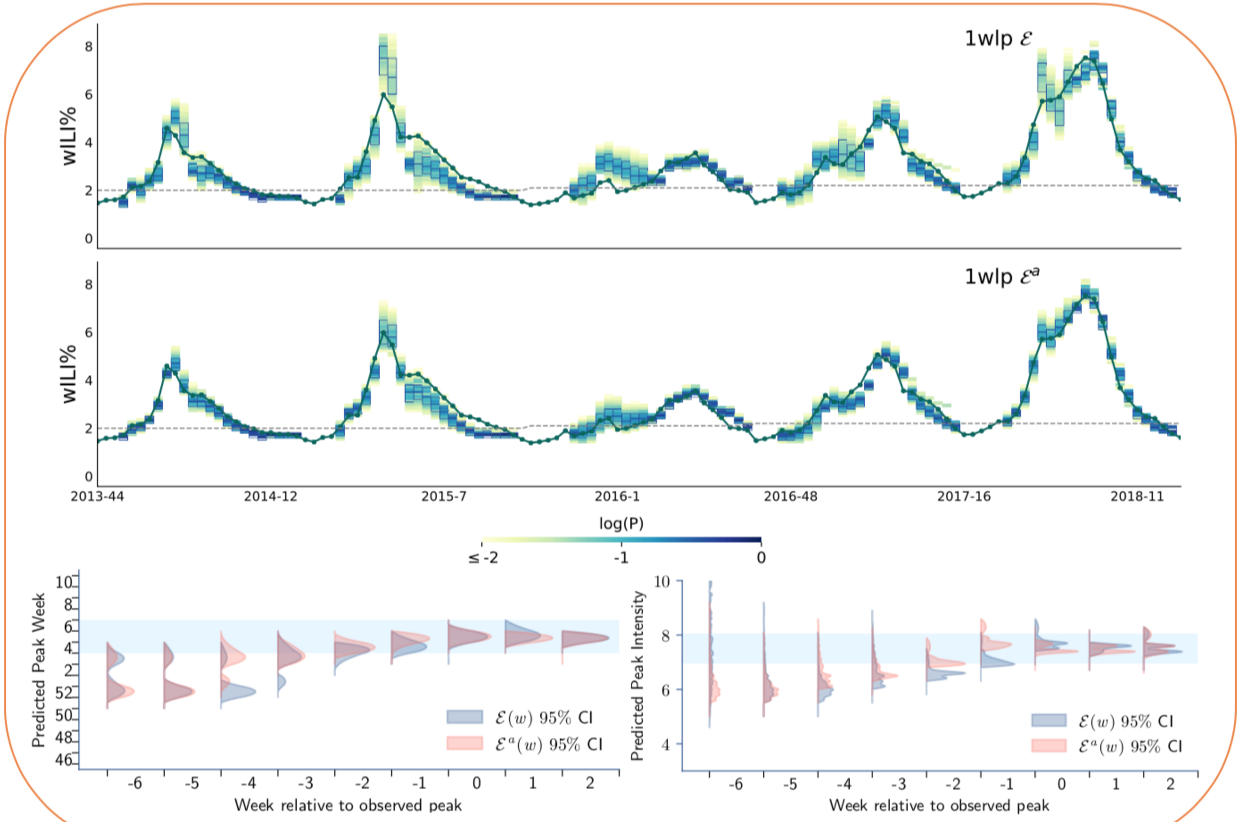
ARGO: A machine learning algorithm to predict short-term time series for seasonal influenza.

Ensemble: We use Bayesian Model Averaging model with Markov Chain Monte Carlo sampling method.

Super-ensemble Seasonal Influenza Forecasting with Machine Learning Augmented Mechanistic Models

Xinyue Xiong, Qian Zhang, Fred S. Lu, Mauricio Santillana, Alessandro Vespignani

RESULTS



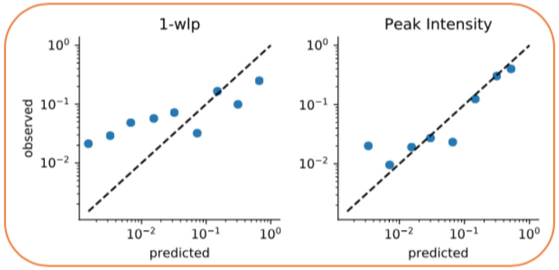
National Short & Long Term Forecasting: The upper/lower panel compares the short/long-term forecast by non-augmented ensemble model $\mathcal{E}(w)$ and augmented model $\mathcal{E}^a(w)$. Obviously the augmented prediction is more concise and accurate in short-term targets, and converges faster in long-terms.

Super-ensemble Seasonal Influenza Forecasting with Machine Learning Augmented Mechanistic Models

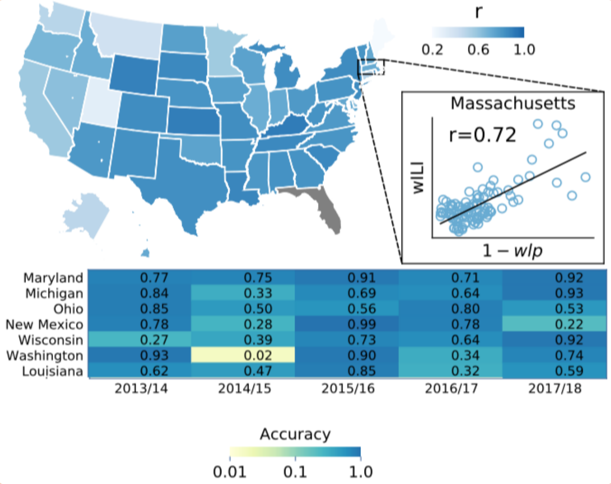
Xinyue Xiong, Qian Zhang, Fred S. Lu, Mauricio Santillana, Alessandro Vespignani

season	2013/14	2015/16	2017/18
Effective reproduction number R_0^{eff}	1.36 (1.35, 1.38)	1.11 (1.10, 1.11)	1.15 (1.13, 1.16)
Residual immunity r	0.34 (0.30, 0.38)	0.21 (0.18, 0.23)	0.15 (0.12, 0.17)
Average infectious time μ^{-1}	4.2 (4.0, 4.3)	2.7 (2.5, 2.8)	3.2 (2.9, 3.5)
ILI Patient visit rate γ	0.37 (0.31, 0.42)	0.73 (0.69, 0.77)	0.71 (0.67, 0.76)

Epidemic Parameters: estimated by $\mathcal{E}^a(w)$ at the end of each season.



Calibration: For the 1-wlp, the predictions (blue dots) with probability less than 0.1 are above the reference line (black dashed line), implying over-confidence. While the peak intensity shows well-calibrated.

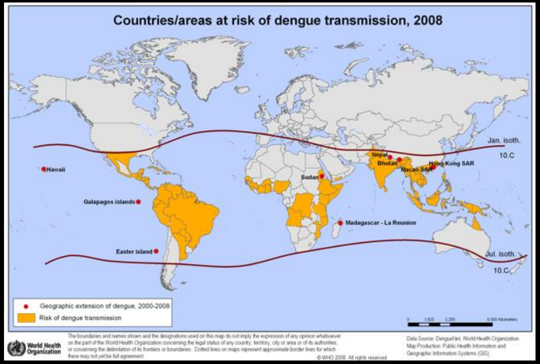


	2013/14	2014/15	2015/16	2016/17	2017/18
Maryland	0.77	0.75	0.91	0.71	0.92
Michigan	0.84	0.33	0.69	0.64	0.93
Ohio	0.85	0.50	0.56	0.80	0.53
New Mexico	0.78	0.28	0.99	0.78	0.22
Wisconsin	0.27	0.39	0.73	0.64	0.92
Washington	0.93	0.02	0.90	0.34	0.74
Louisiana	0.62	0.47	0.85	0.32	0.59

State Forecasting: Performance of $\mathcal{E}^a(w)$ on state level. The upper map shows the Pears- on correlation between the 1-wlp and the observed wILI incidences, while the lower table shows the accuracy of peak week prediction half-month prior to the actual peaks.

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

Dengue, Zika, and Flu



Countries/areas at risk of dengue transmission, 2008

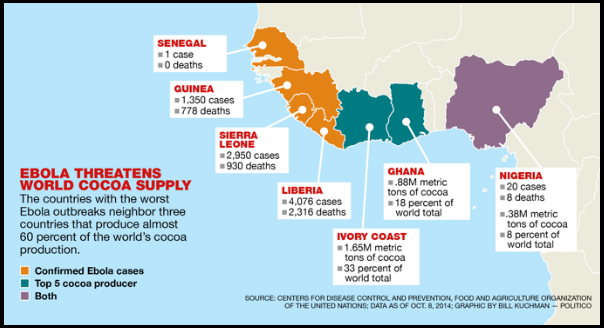
The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted lines on maps represent approximate border lines for which there may not yet be full agreement.

© WHO 2008. All rights reserved

Data Source: DengueNet, World Health Organization
Map Production: Public Health Information and Geographic Information Systems (GIS)

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

Ebola



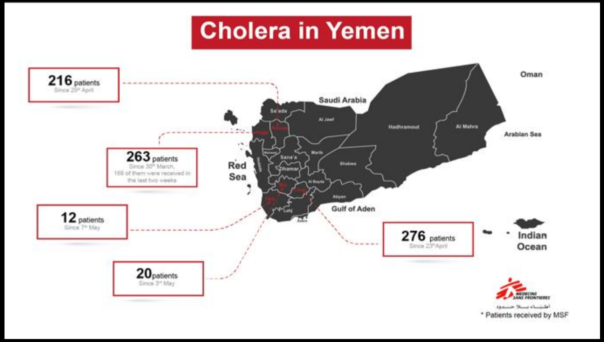
EBOLA THREATENS WORLD COCOA SUPPLY

The countries with the worst Ebola outbreaks neighbor three countries that produce almost 60 percent of the world's cocoa production.

SOURCE: CENTERS FOR DISEASE CONTROL AND PREVENTION, FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS; DATA AS OF OCT. 8, 2014; GRAPHIC BY BILL KUCHMAN — POLITICO

- West Africa

Cholera



Cholera in Yemen

* Patients received by MSF

- 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

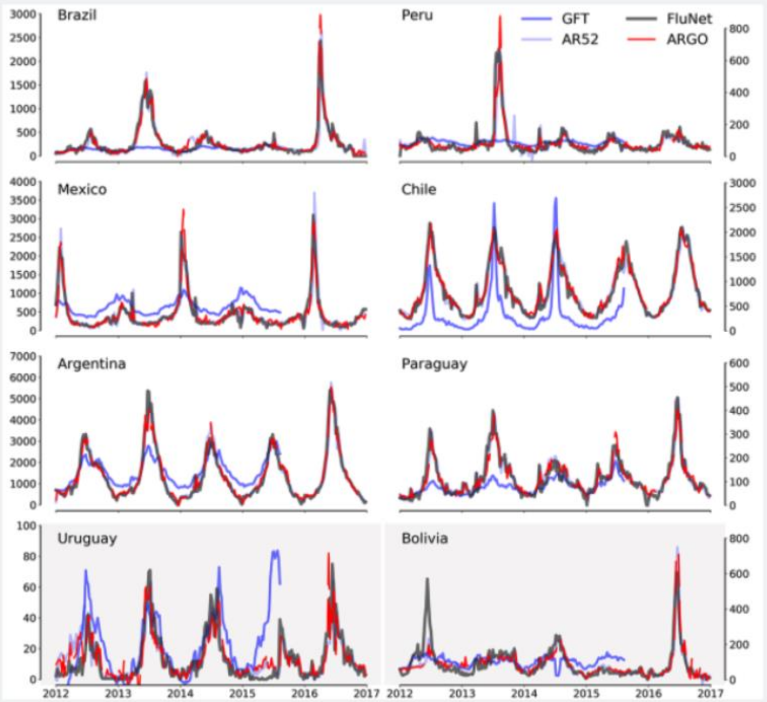


Figure 1. Graphical representation of the number of processed specimens (NPSs) as reported by WHO's FluNet (black), along with the NPS estimates generated by ARGO (red), AR (light blue), and Google Flu Trends (GFT; blue), over the whole study period of January 1, 2012 to December 25, 2016.

View this figure

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]

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

Abstract

Background:

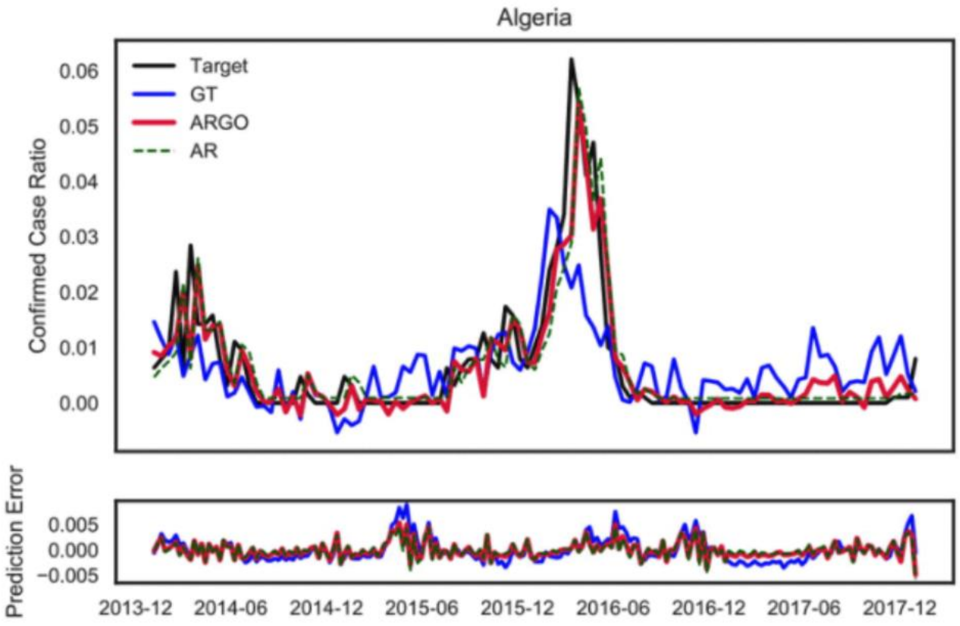
Novel influenza surveillance systems that leverage Internet-based real-time data sources including Internet search frequencies, social-network information, and crowd-sourced flu surveillance tools have shown improved accuracy over the past few years in data-rich countries like the United States. These 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



RESEARCH ARTICLE

Leveraging Google search data to track influenza outbreaks in Africa [version 1; peer review: 1 approved, 1 not approved]

Karla Mejia [1], Cecile Viboud[2], Mauricio Santillana[3,4]

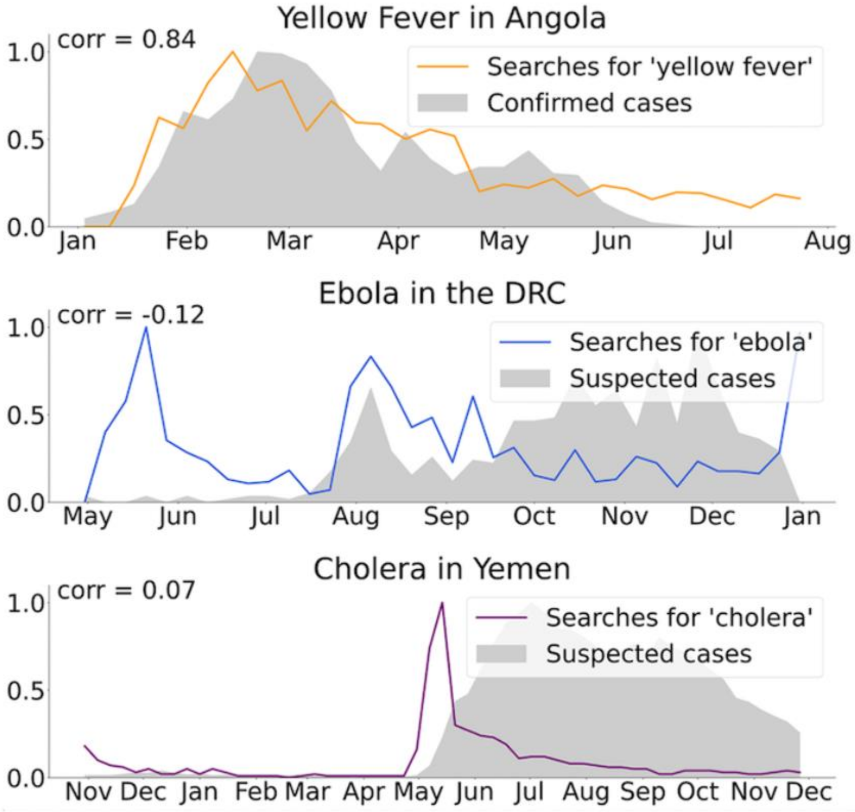
Abstract

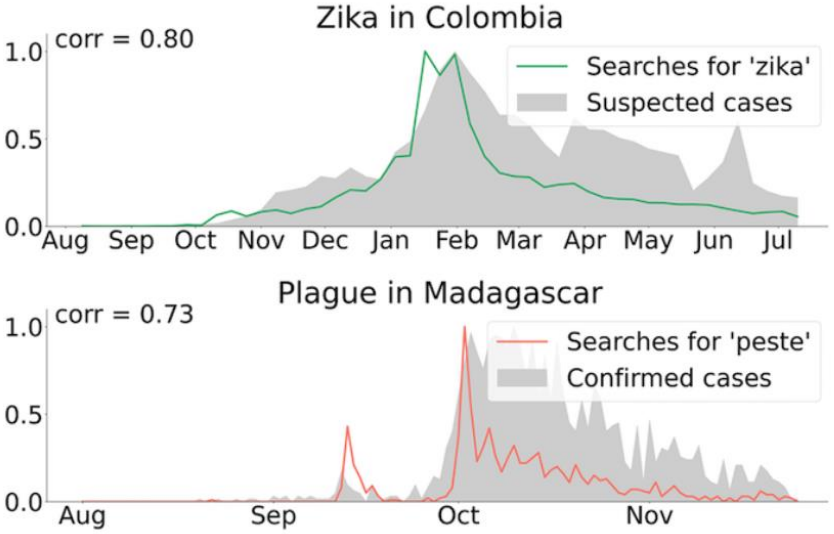
Background: Traditionally, public health agencies track seasonal influenza activity by collecting information from clinics, hospitals, and laboratories. The inherent slowness of the processes used to collect influenza activity data limits the ability of public health agencies to adapt to unexpected changes in influenza activity in near real-time. In recent years, new influenza surveillance methods that use nontraditional data sources, such as Google searches, have been proposed to successfully estimate influenza activity in near real-time. However, most of these methods have been designed for and implemented in high-income countries even though influenza disease burden remains high in low- to middle-income countries. Here, we seek to predict influenza activity in near real-time in Africa using machine learning models that combine Google searches with traditional epidemiological data.

Methods: We extend the AutoRegression with Google search data (ARGO) model to track influenza activity in near-real-time in Africa. The ARGO model, which was originally designed to predict influenza activity in the United States, combines influenza-related Google searches with

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

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

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

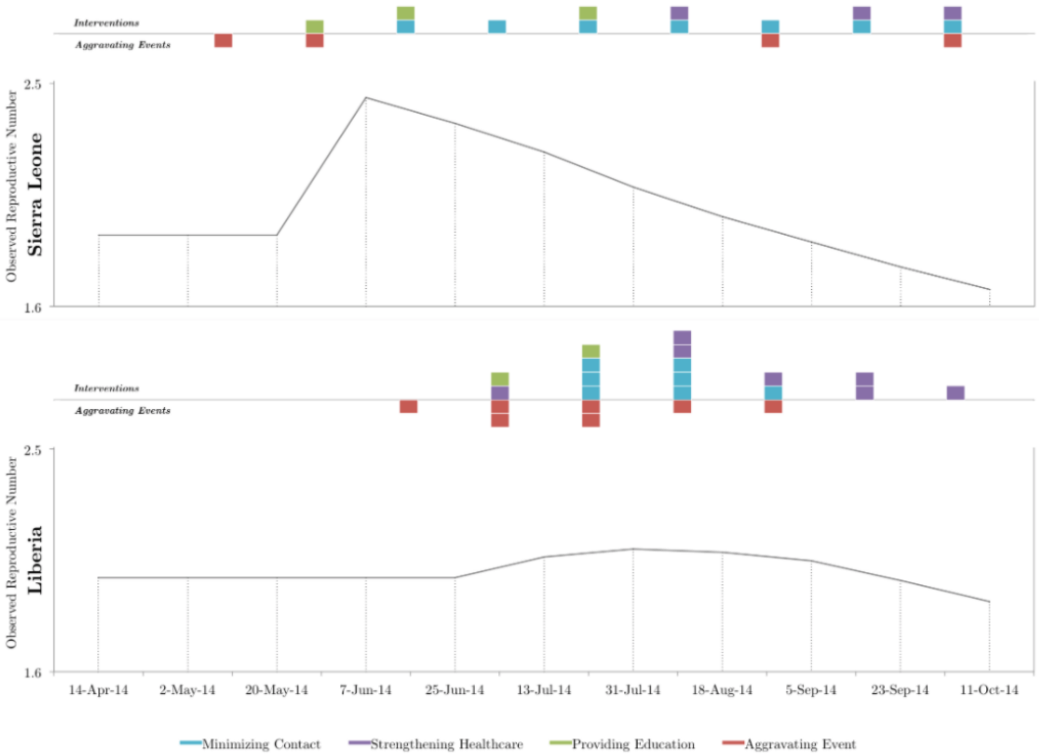
Yes, news alerts related to the 2014 Ebola outbreak in Western Africa foreshadowed changes in the local reproductive number

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

APRIL 28, 2015 · DISCUSSION

AUTHORS

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



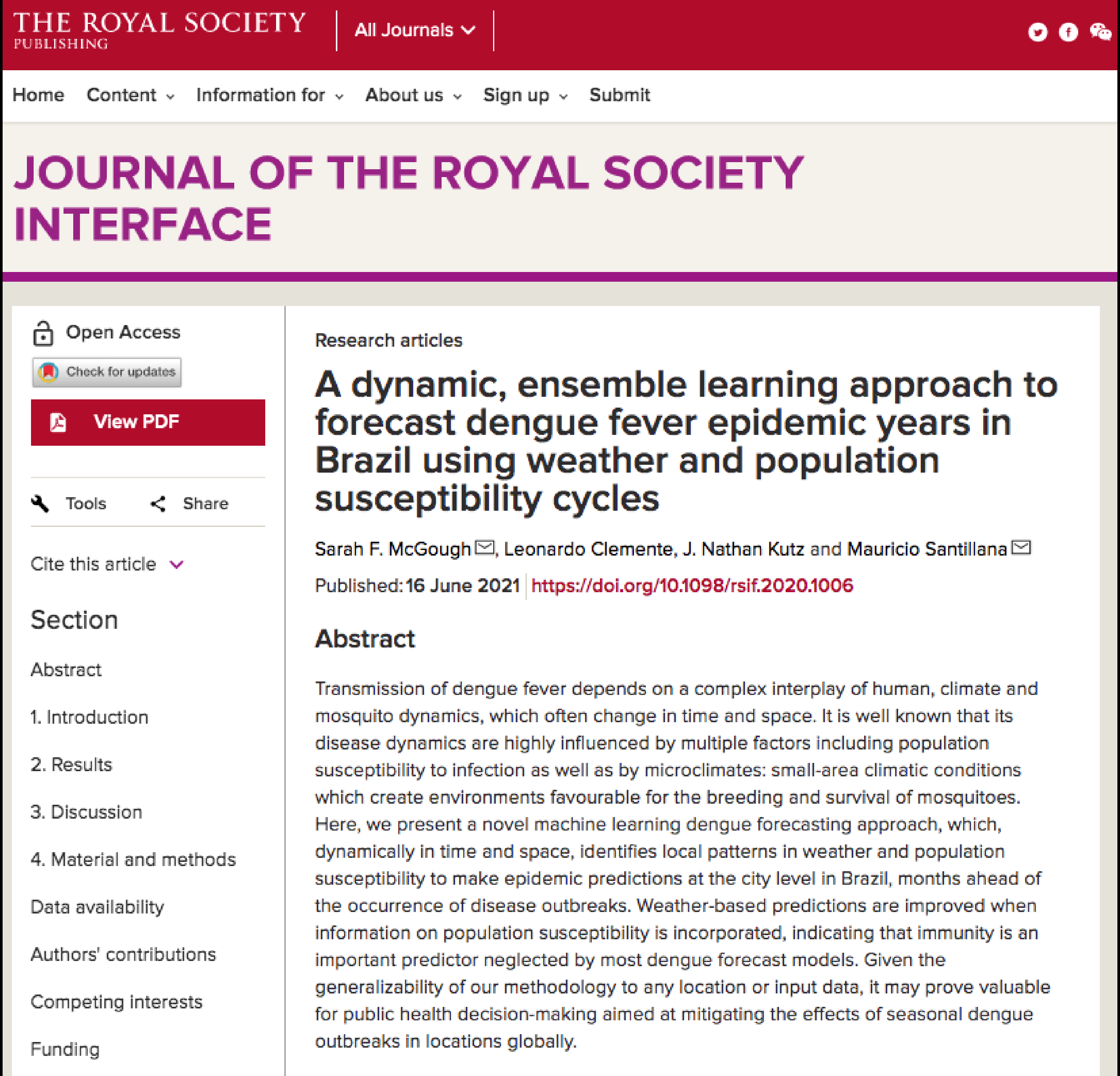
https://currents.plos.org/outbreaks/index.html%3Fp=50634.html

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



JOURNAL OF THE ROYAL SOCIETY INTERFACE

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

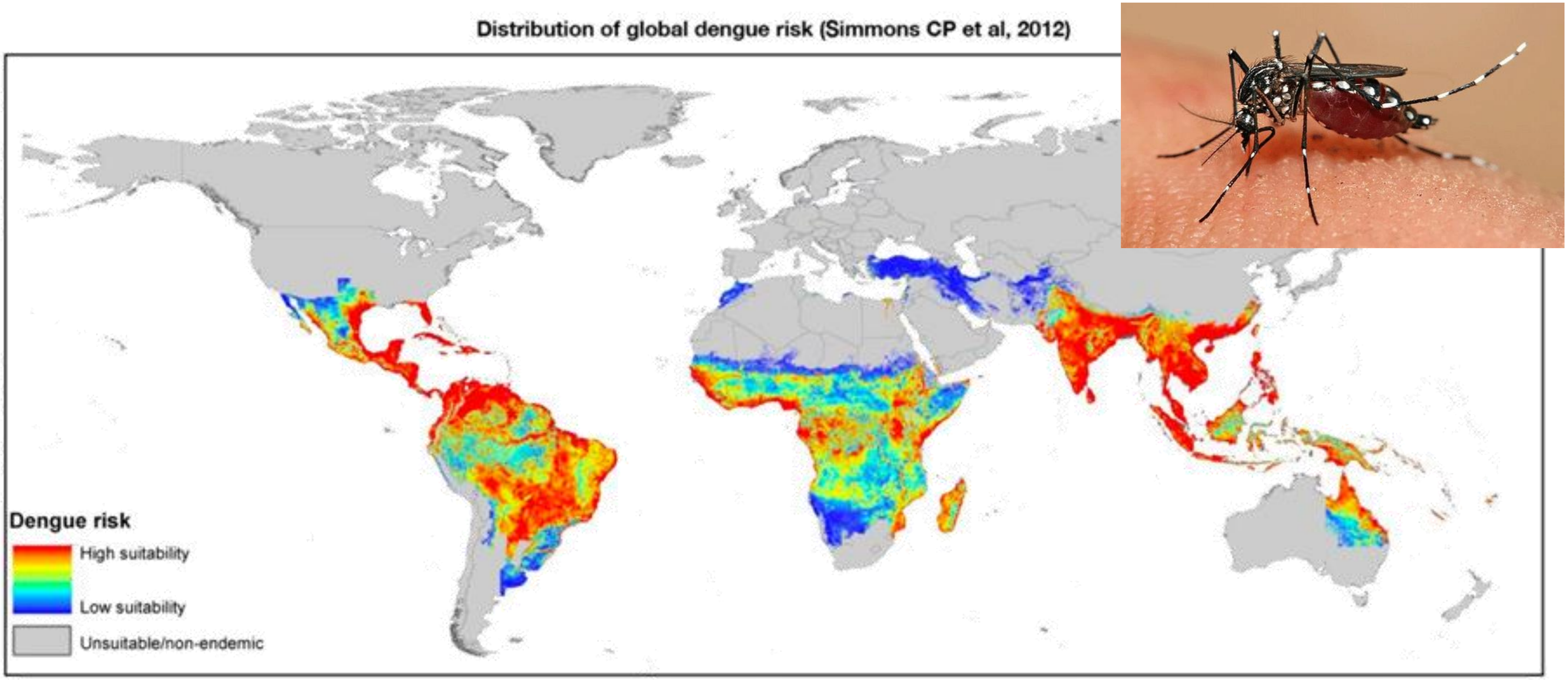
Published: 16 June 2021 https://doi.org/10.1098/rsif.2020.1006

Section

Abstract

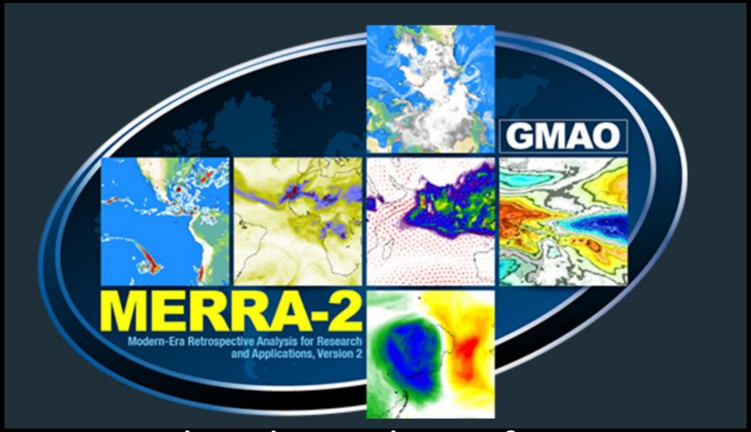
Transmission of dengue fever depends on a complex interplay of human, climate and mosquito dynamics, which often change in time and space. It is well known that its disease dynamics are highly influenced by multiple factors including population susceptibility to infection as well as by microclimates: 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

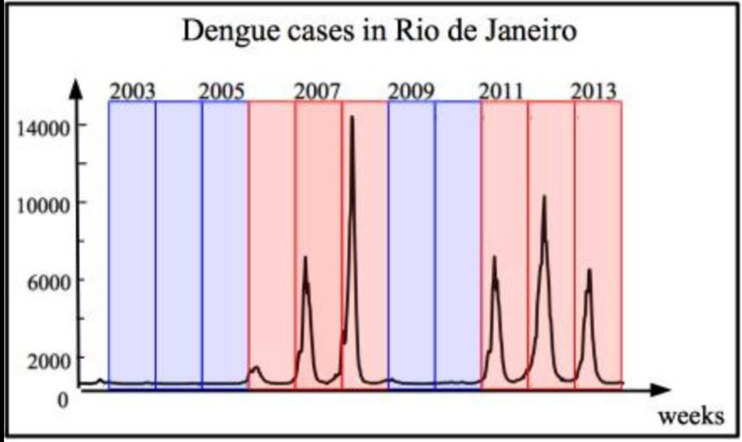


Distribution of global dengue risk (Simmons CP et al, 2012)

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

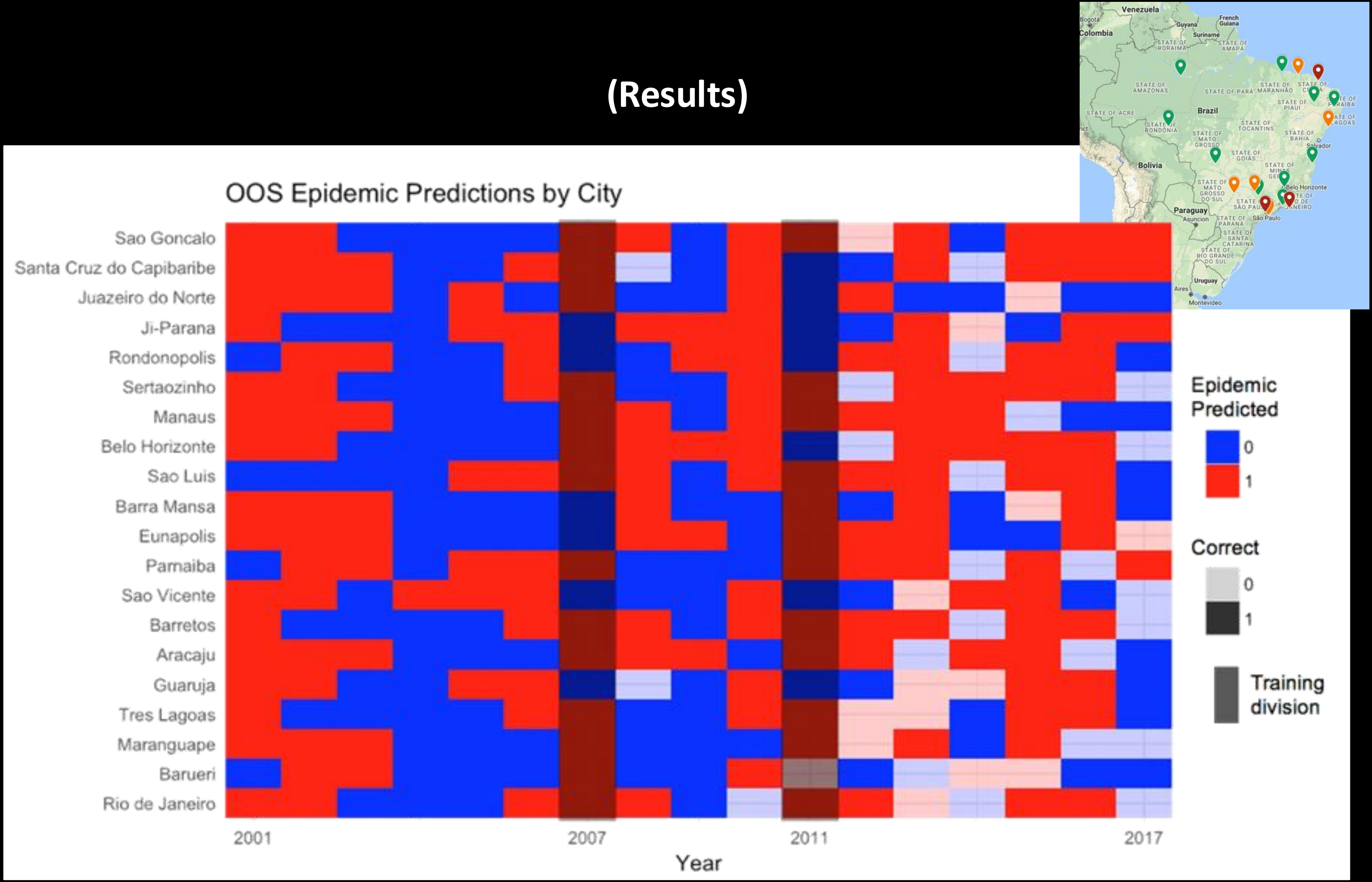
Our contribution:

- 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

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

(Results)



How do we incorporate DENV susceptibility cycles (3-4 years)?

- Data-driven Hidden Markov model

Transitioning to

Transitioning from

	0	1
00	**0.42**	**0.58**
11	**0.71**	**0.29**
000	**0.21**	**0.79**
111	**0.78**	**0.22**

Frequency of transition calculated from the data

How do we incorporate DENV susceptibility cycles (3-4 years)?

- Data-driven Hidden Markov model

Transitioning to

Transitioning from

	0	**1**
00	**0.42**	**0.58**
11	**0.71**	**0.29**
000	**0.21**	**0.79**
111	**0.78**	**0.22**

ENSEMBLE APPROACH

Decision Rule:

IF probability of transition > percent of model votes
THEN overturn the vote

How do we incorporate DENV susceptibility cycles (3-4 years)?

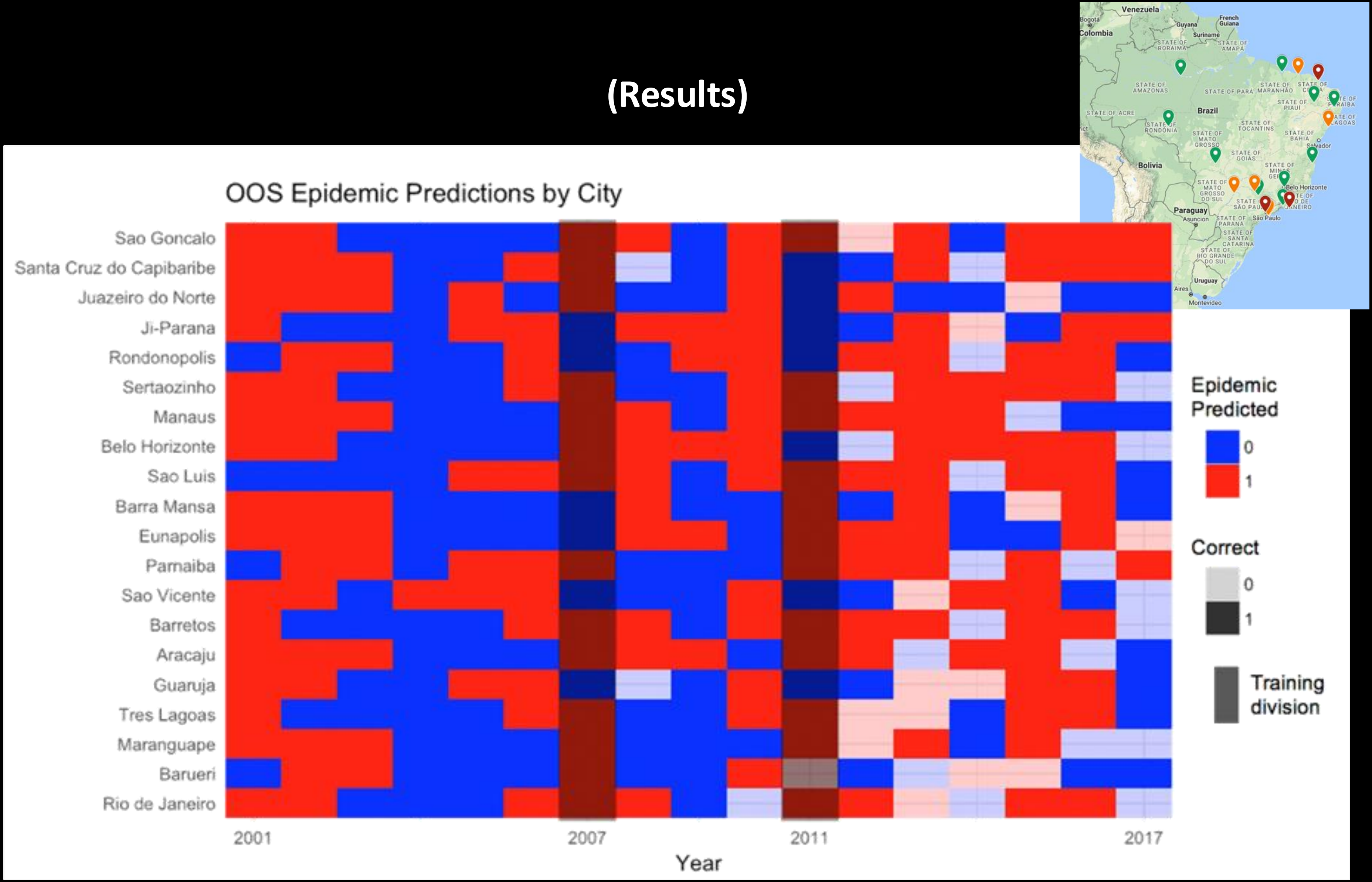
- Data-driven Hidden Markov model

OOS Evaluation Metric	Climate	Climate + DENV Cycle
Accuracy	71.7%	75%
Hit rate (Sensitivity)	81%	78%
Non-epidemic detection rate	58%	71%
No-information rate	60%	60%
P(Accuracy > No-Information Rate)	p = 0.005	p=0.0004

Pattern	N(overrides)	Correct	Incorrect
1110	5	4	1
110	3	2	1
0001	0	-	-
001	0	-	-

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

(Results)



scientific reports

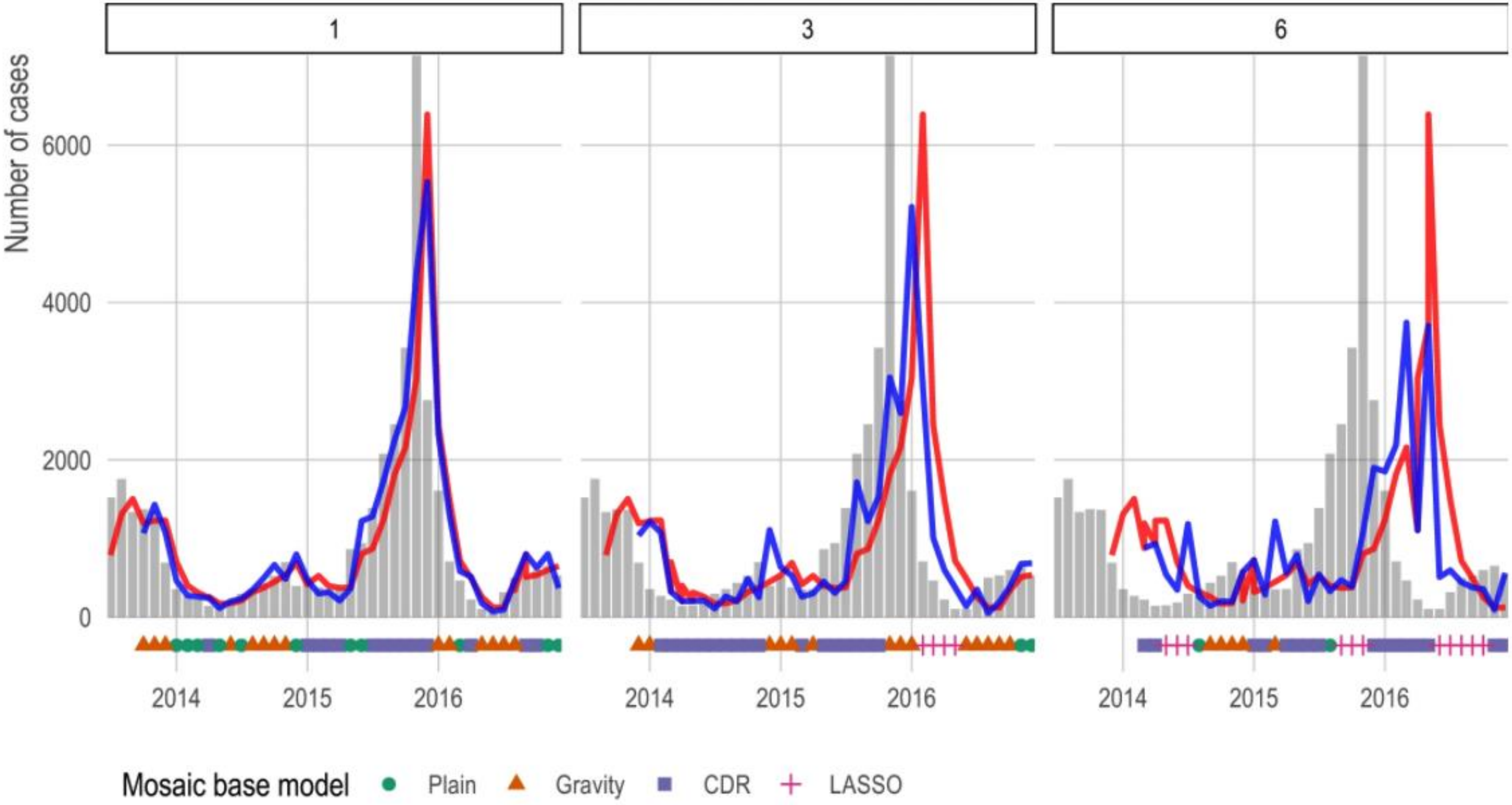
Article | Open Access | Published: 13 January 2021

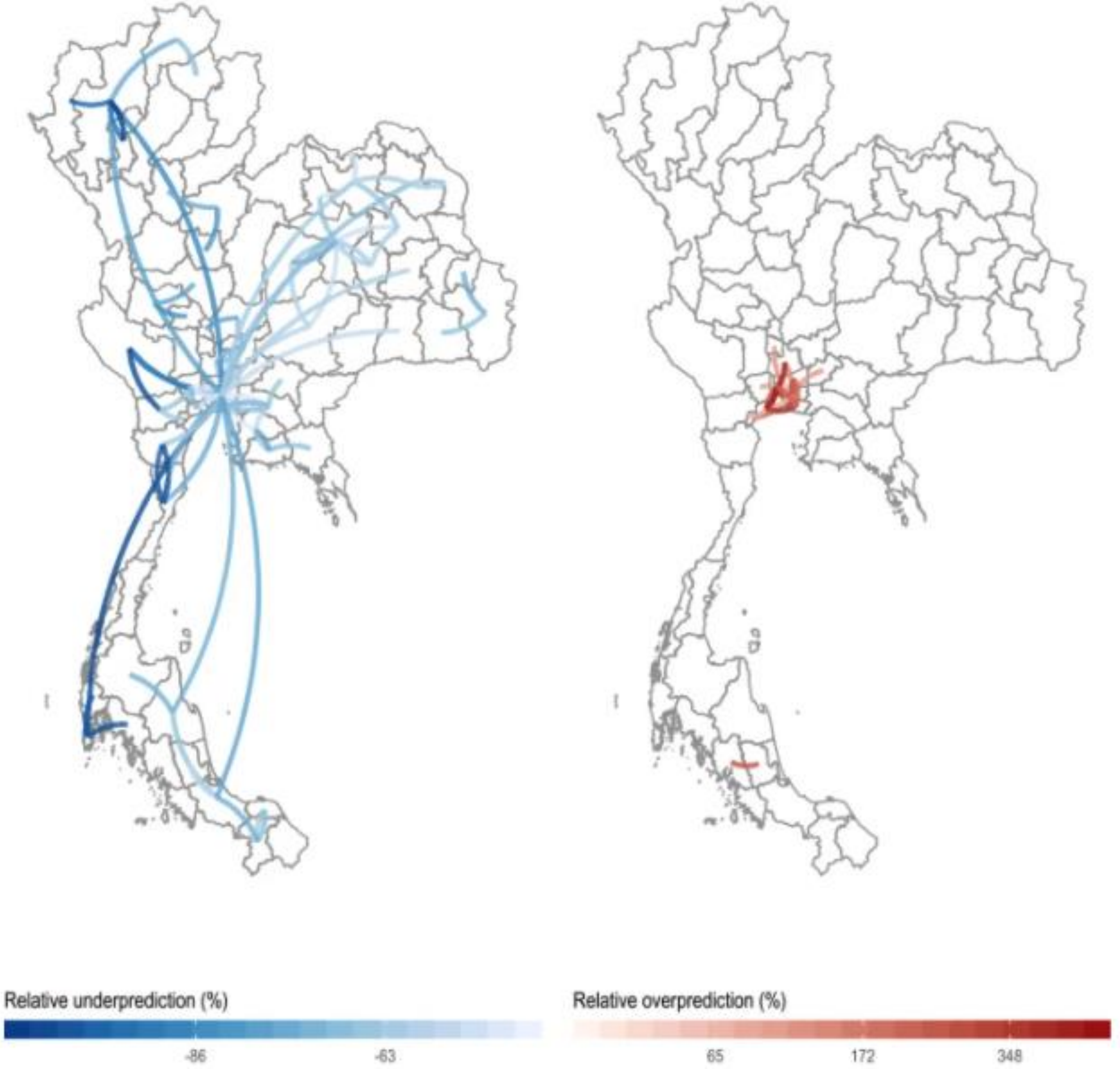
Incorporating human mobility data improves forecasts of Dengue fever in Thailand

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

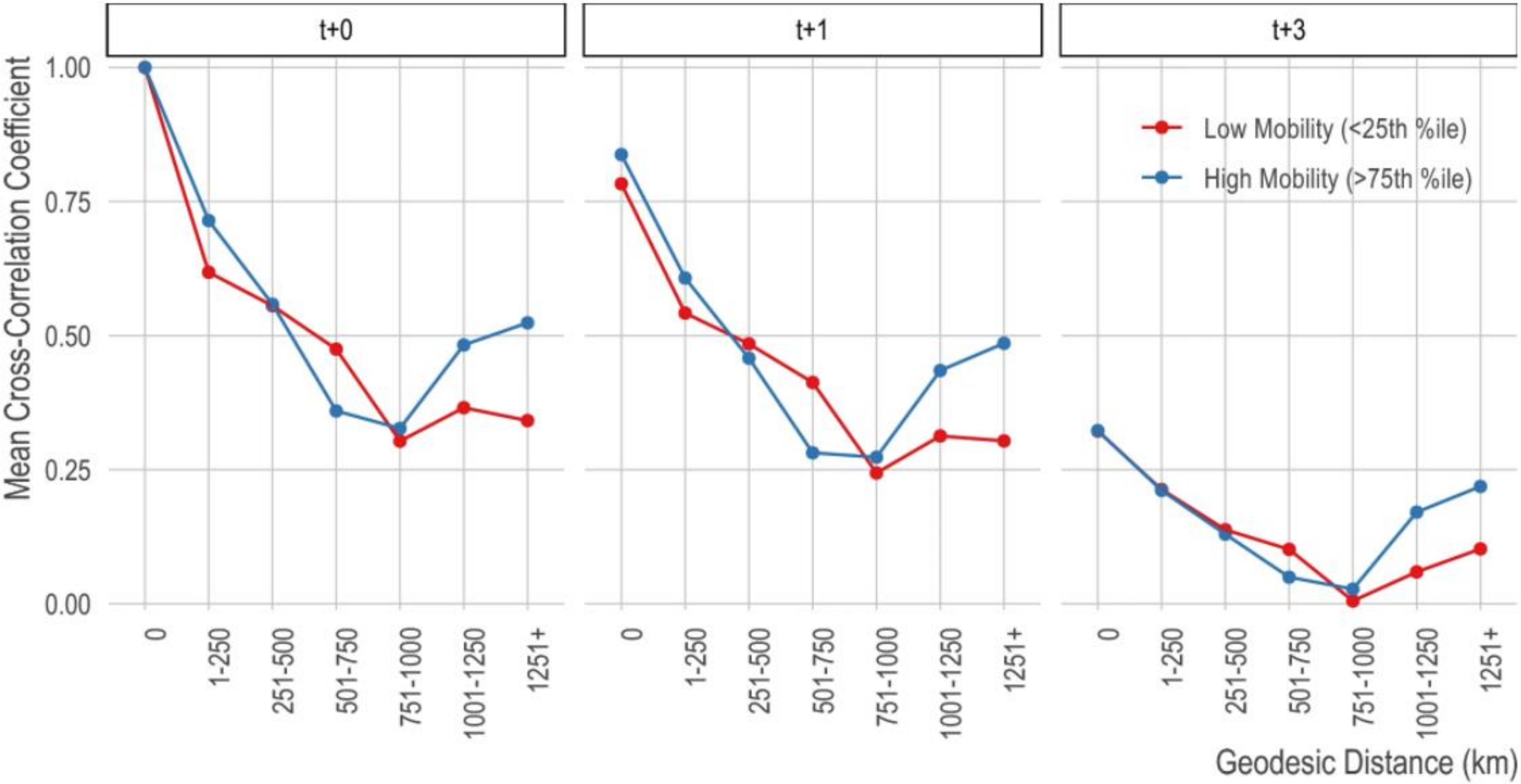
Scientific Reports **11**, Article number: 923 (2021) | Cite this article

1669 Accesses | **5** Altmetric | Metrics


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

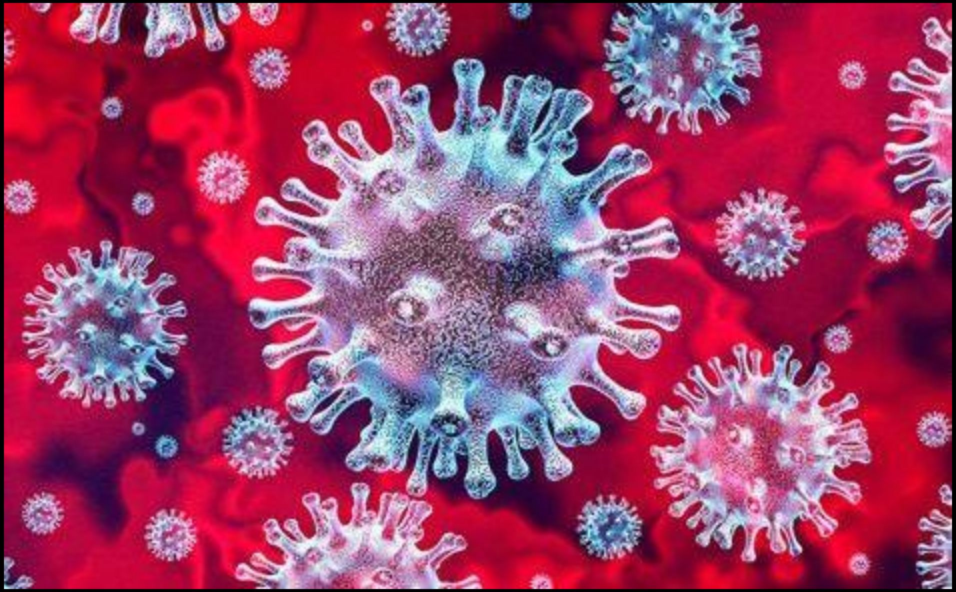


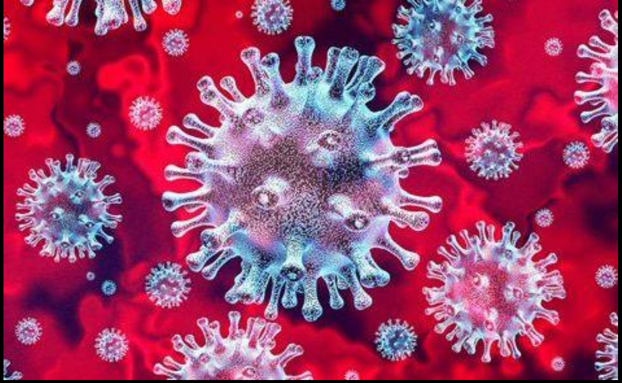
Under- and over-prediction of outlier travel. Relative under-prediction (left) and over-prediction (right) comparing observed mobility data (from CDRs) to estimated mobility data from the best fit gravity model. We defined relative prediction error as 100%*(PredictedTrips – ObservedTrips)/ObservedTrips.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

COVID-19





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

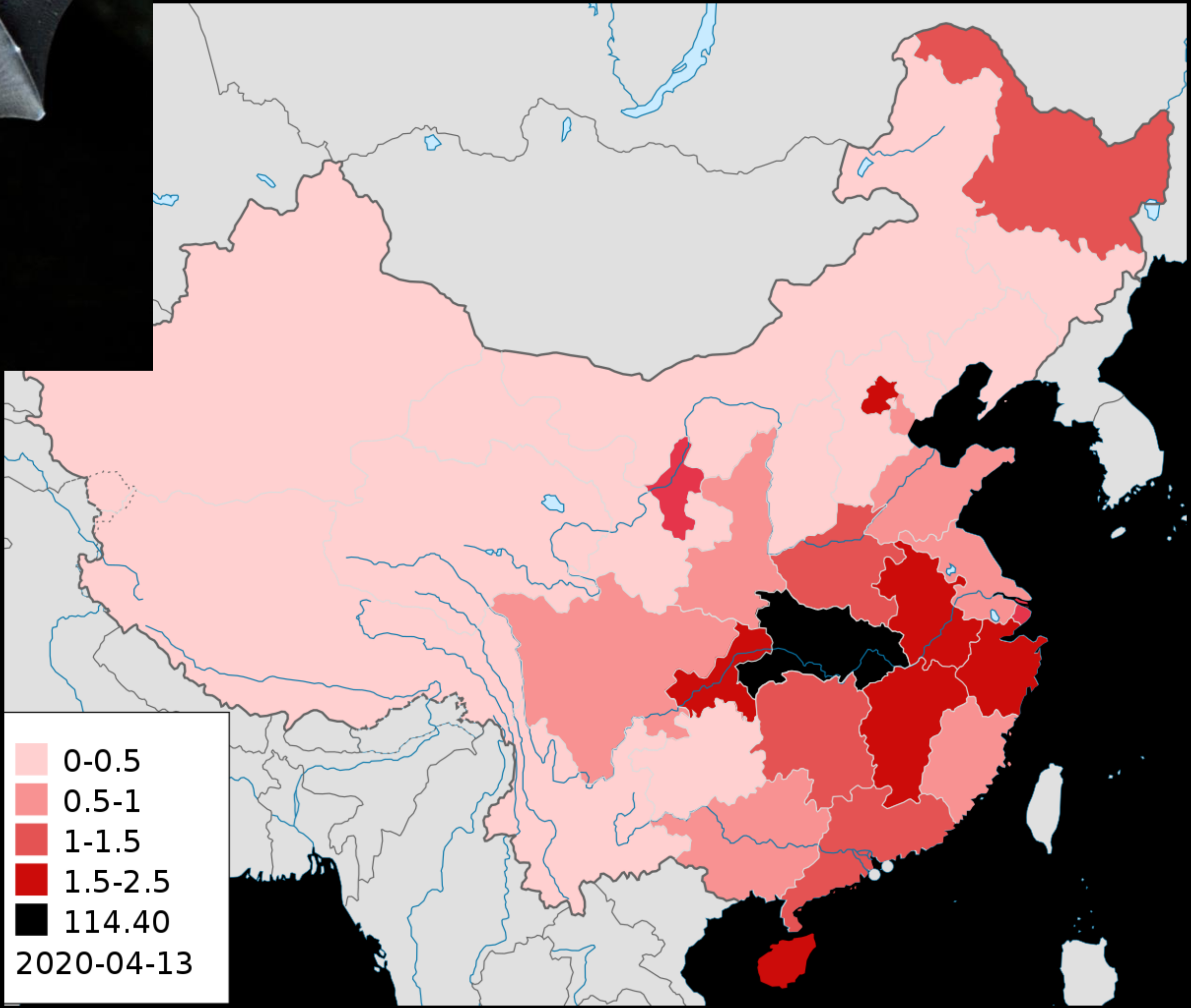
Caused by: severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)

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]

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

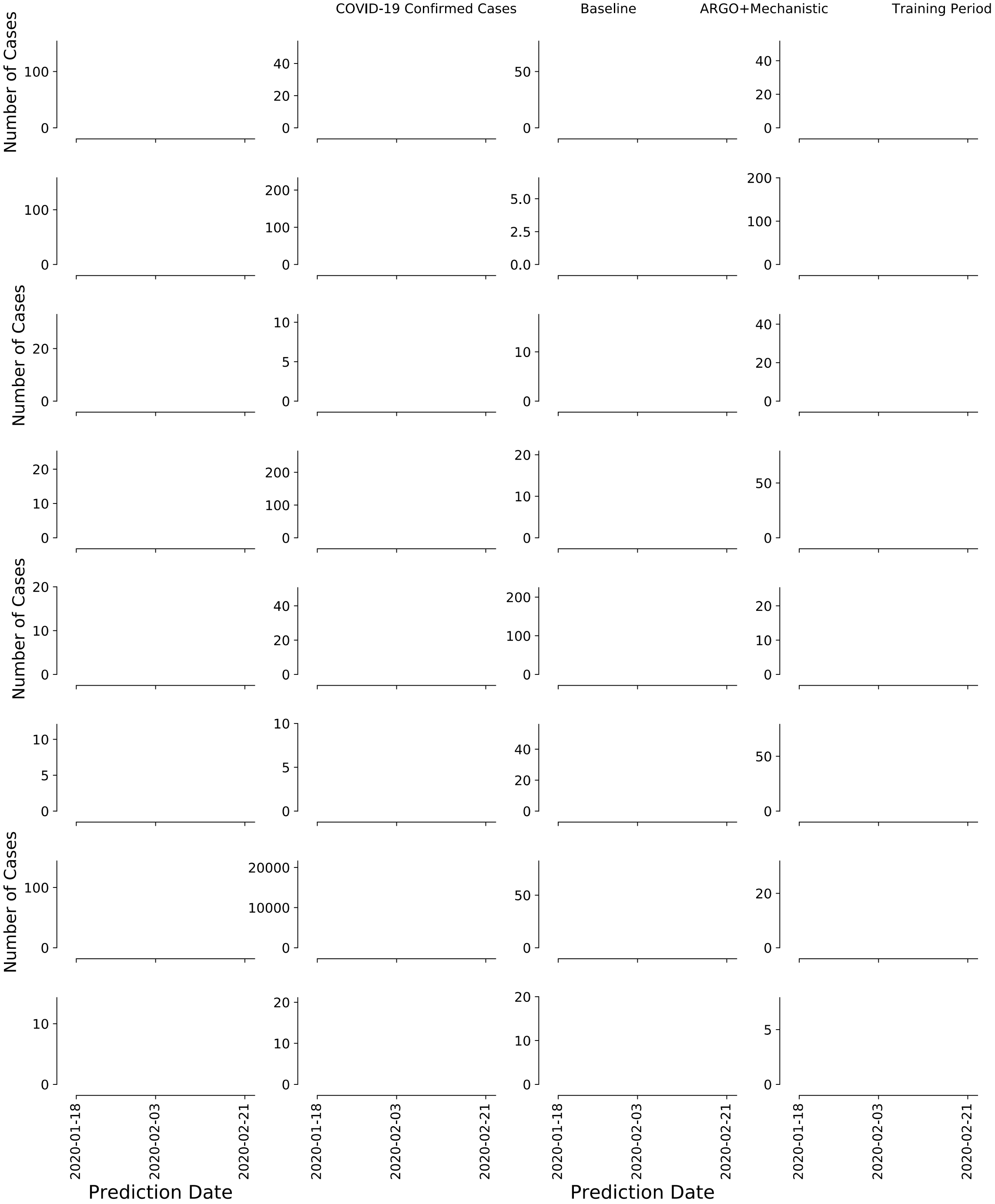
Related Article

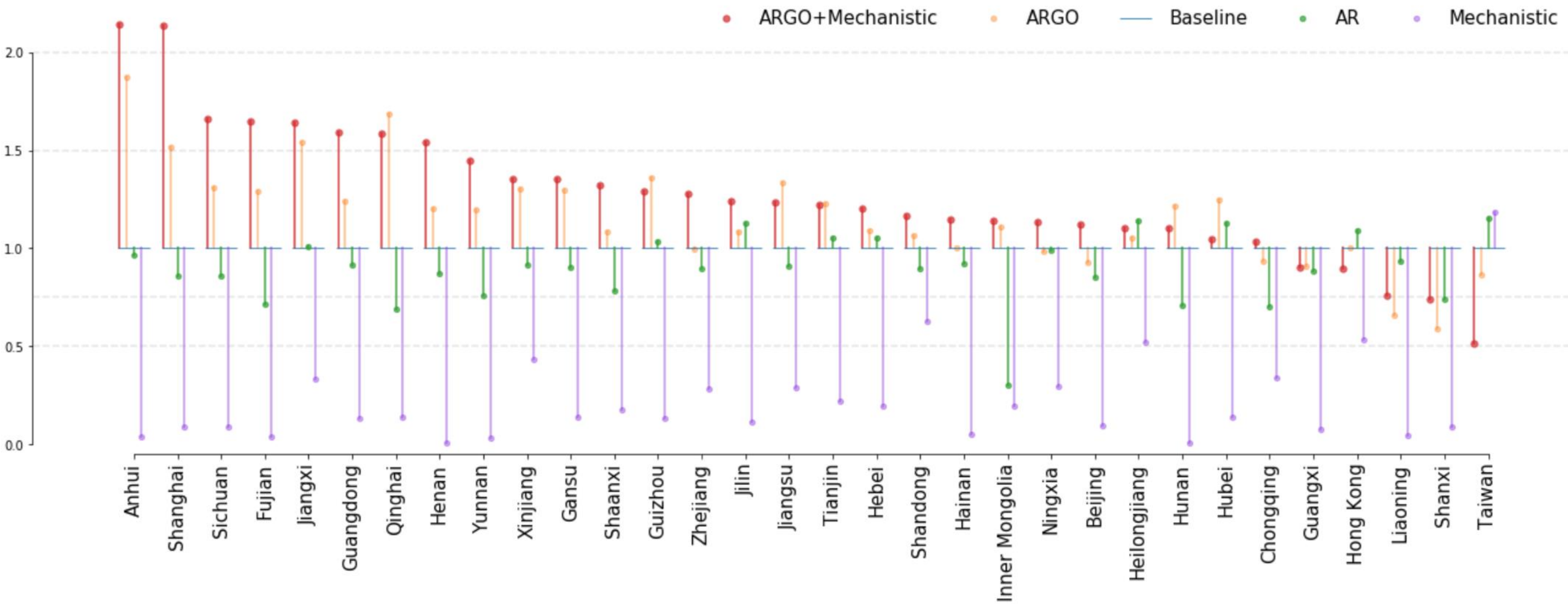
This is a corrected version. See correction statement in: https://www.jmir.org/2020/9/e23996/

Abstract

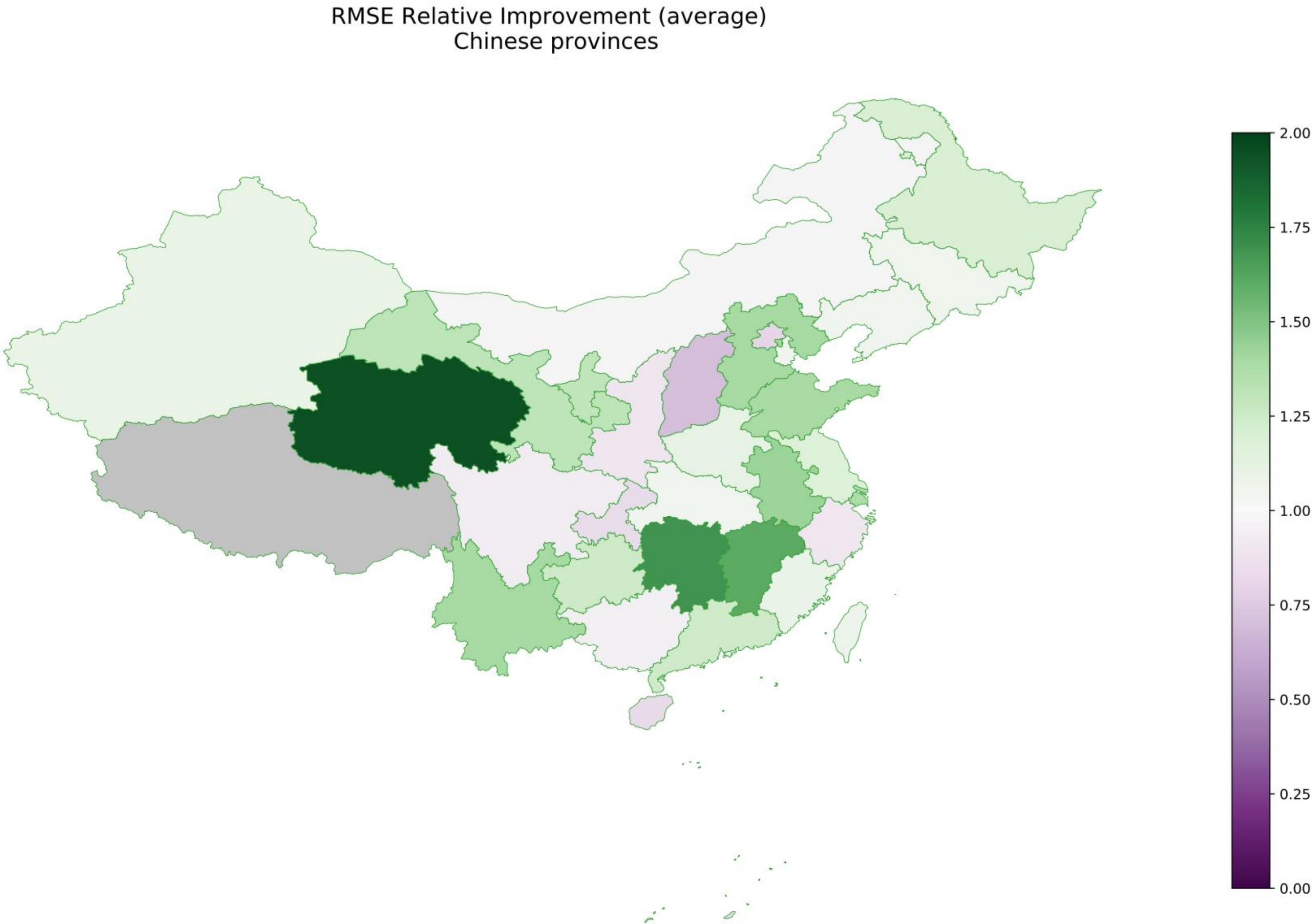
Background:
The inherent difficulty of identifying and monitoring emerging outbreaks caused by novel pathogens can lead to their rapid spread; and if left unchecked, they may become major public health threats to the planet. The ongoing coronavirus disease (COVID-19) outbreak, which has infected over 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.

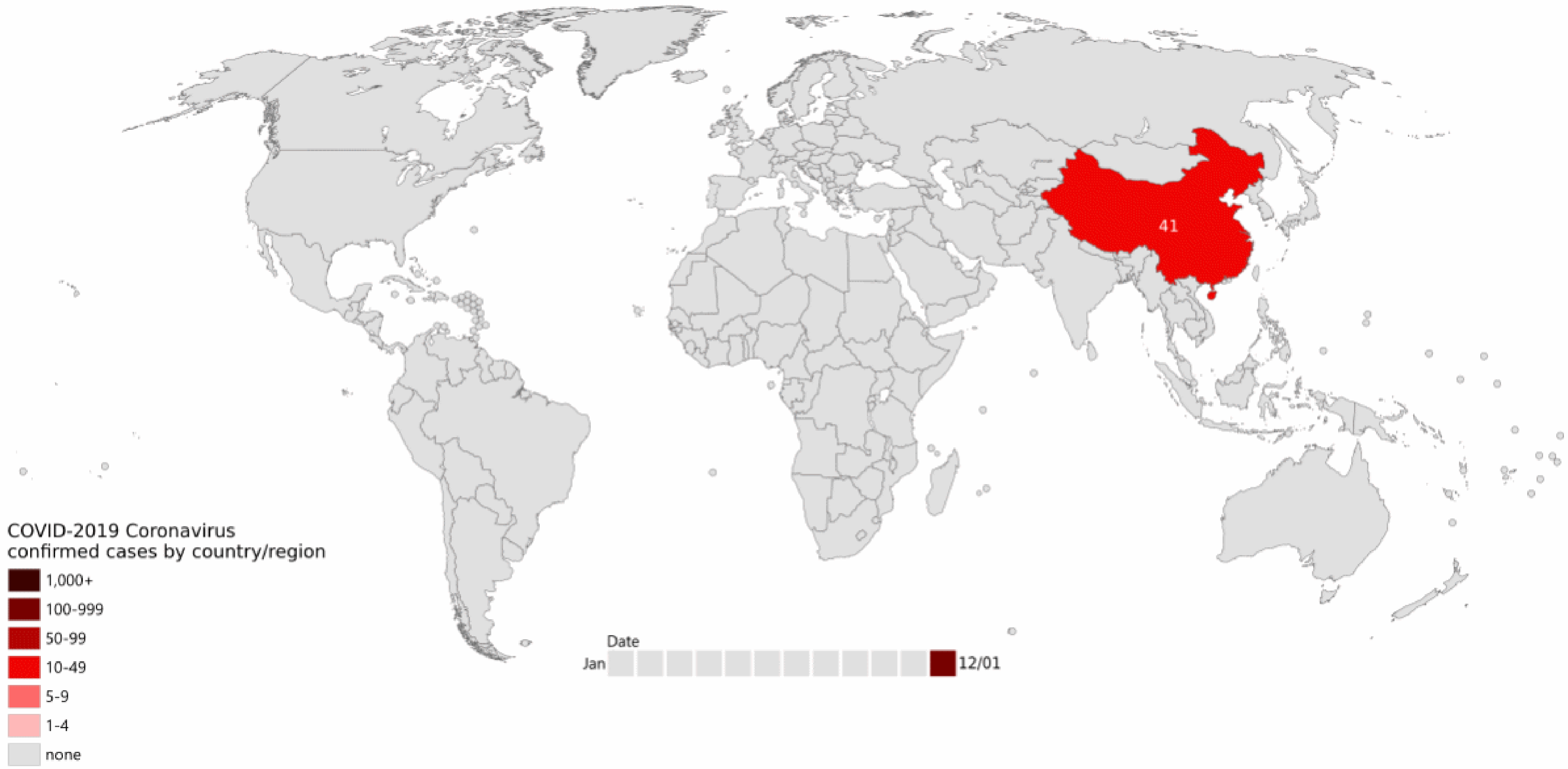

COVID-19 Confirmed Cases
Baseline
ARGO+Mechanistic
Training Period
Number of Cases
Prediction Date
2020-01-18
2020-02-03
2020-02-21



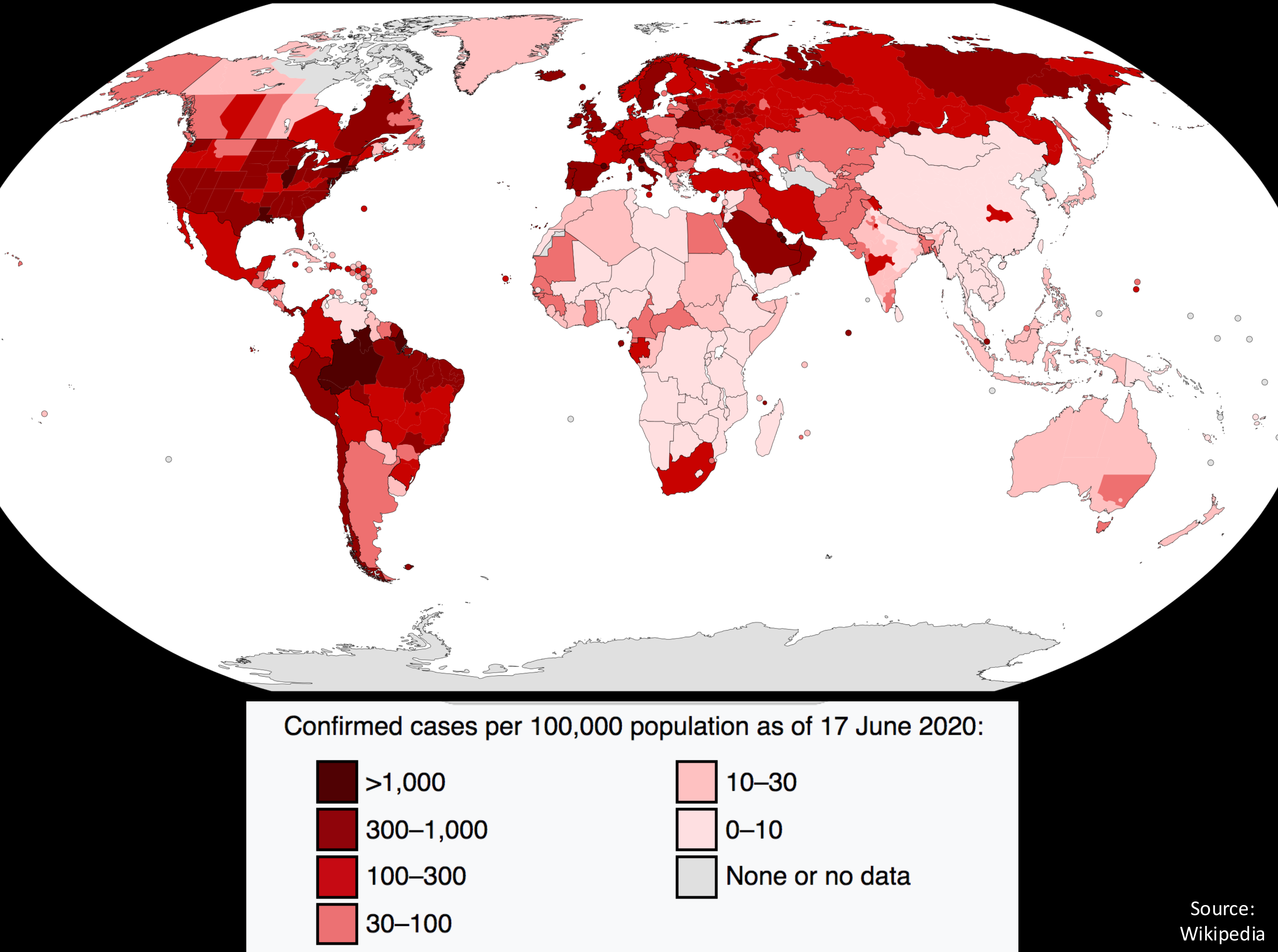
ARGO+Mechanistic
ARGO
Baseline
AR
Mechanistic
2.0
1.5
1.0
0.5
0.0
Anhui
Shanghai
Sichuan
Fujian
Jiangxi
Guangdong
Qinghai
Henan
Yunnan
Xinjiang
Gansu
Shaanxi
Guizhou
Zhejiang
Jilin
Jiangsu
Tianjin
Hebei
Shandong
Hainan
Inner Mongolia
Ningxia
Beijing
Heilongjiang
Hunan
Hubei
Chongqing
Guangxi
Hong Kong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



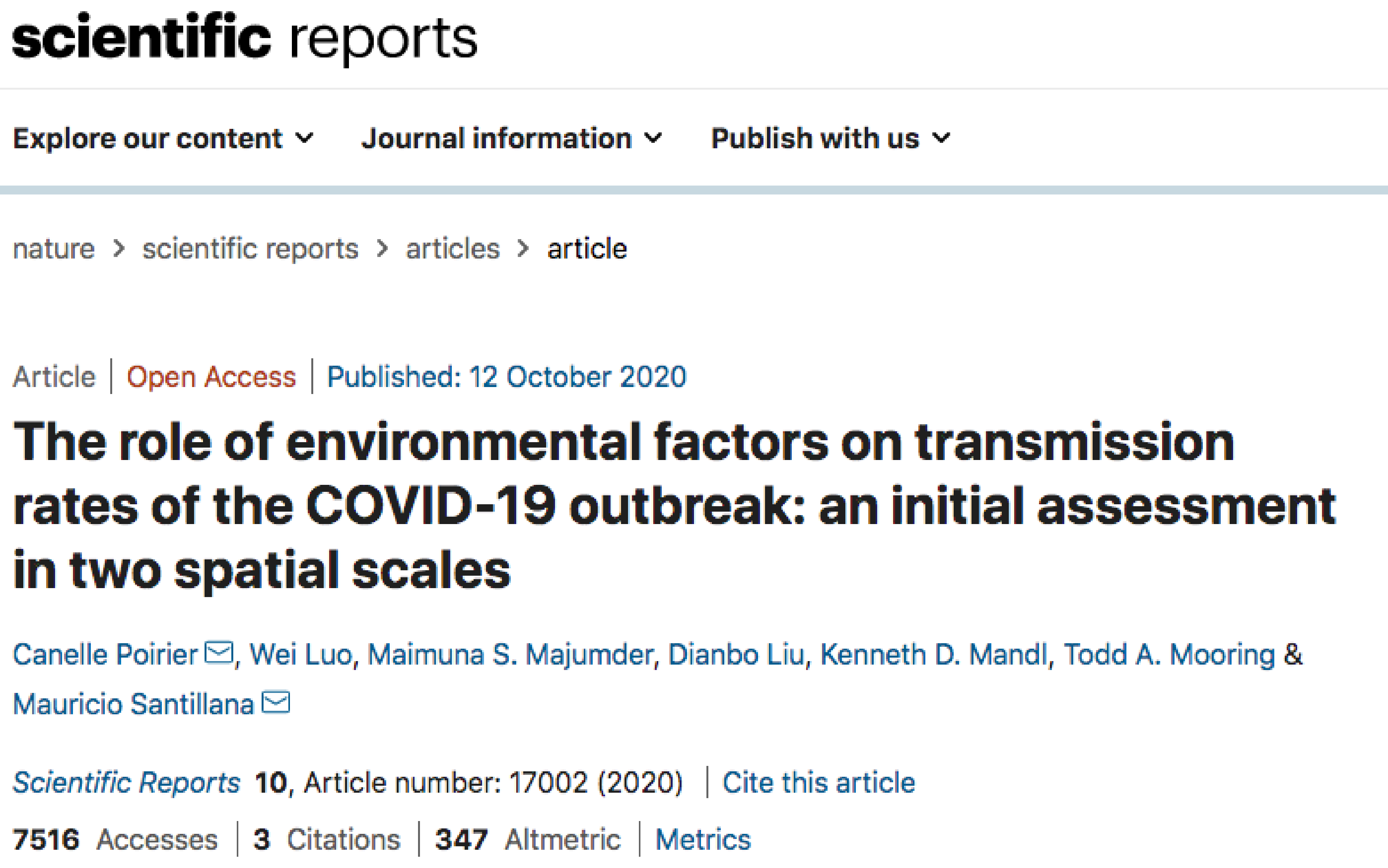
RMSE Relative Improvement (average)
Chinese provinces



Source:
Wikipedia



Source: Wikipedia



Article | Open Access | Published: 12 October 2020

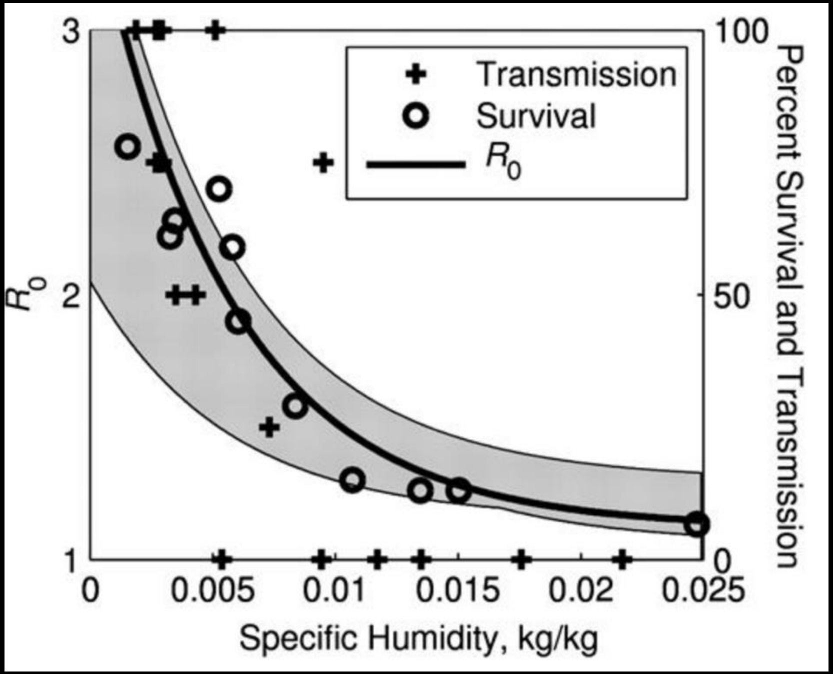
The role of environmental factors on transmission rates of the COVID-19 outbreak: an initial assessment in two spatial scales

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

7516 Accesses | **3** Citations | **347** Altmetric | Metrics

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."*

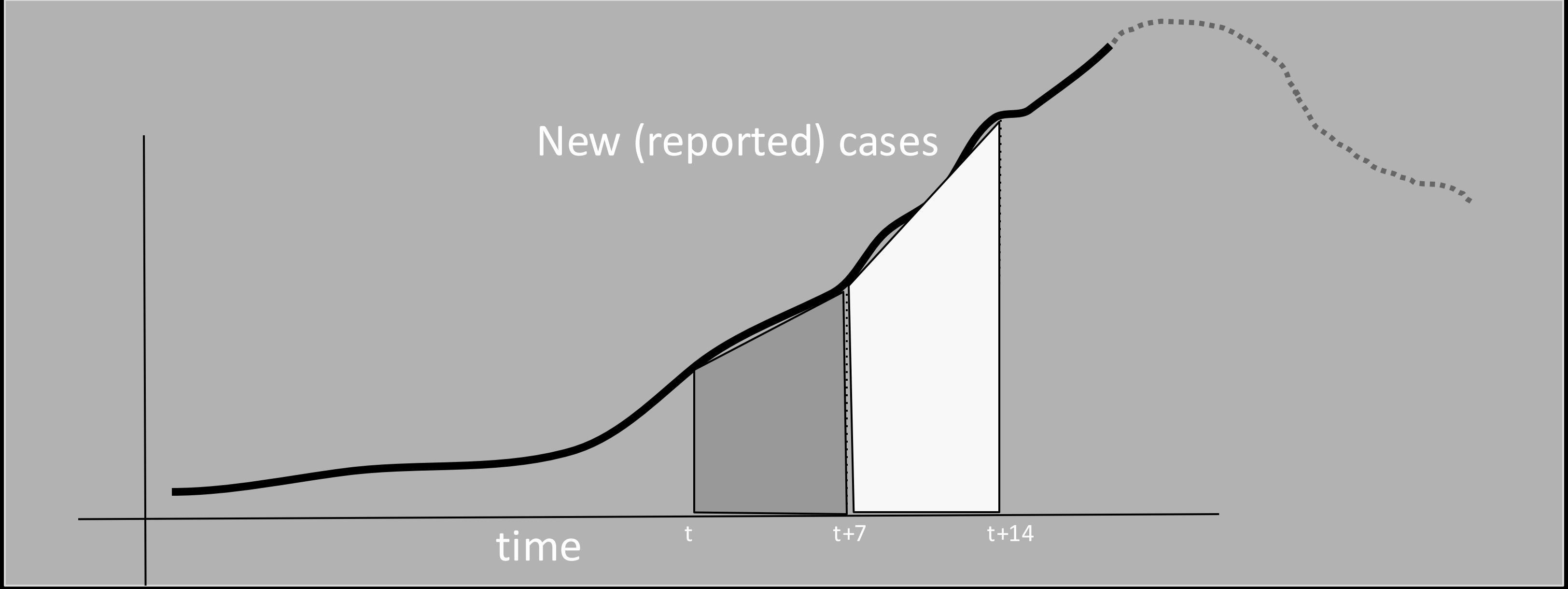
These observed relationships in influenza transmission have been assumed (without any evidence) for the ongoing COVID-19 outbreak. It has been stated that COVID-19 transmission will decrease as warmer temperatures (leading to higher absolute humidity conditions) are experienced in the upcoming spring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

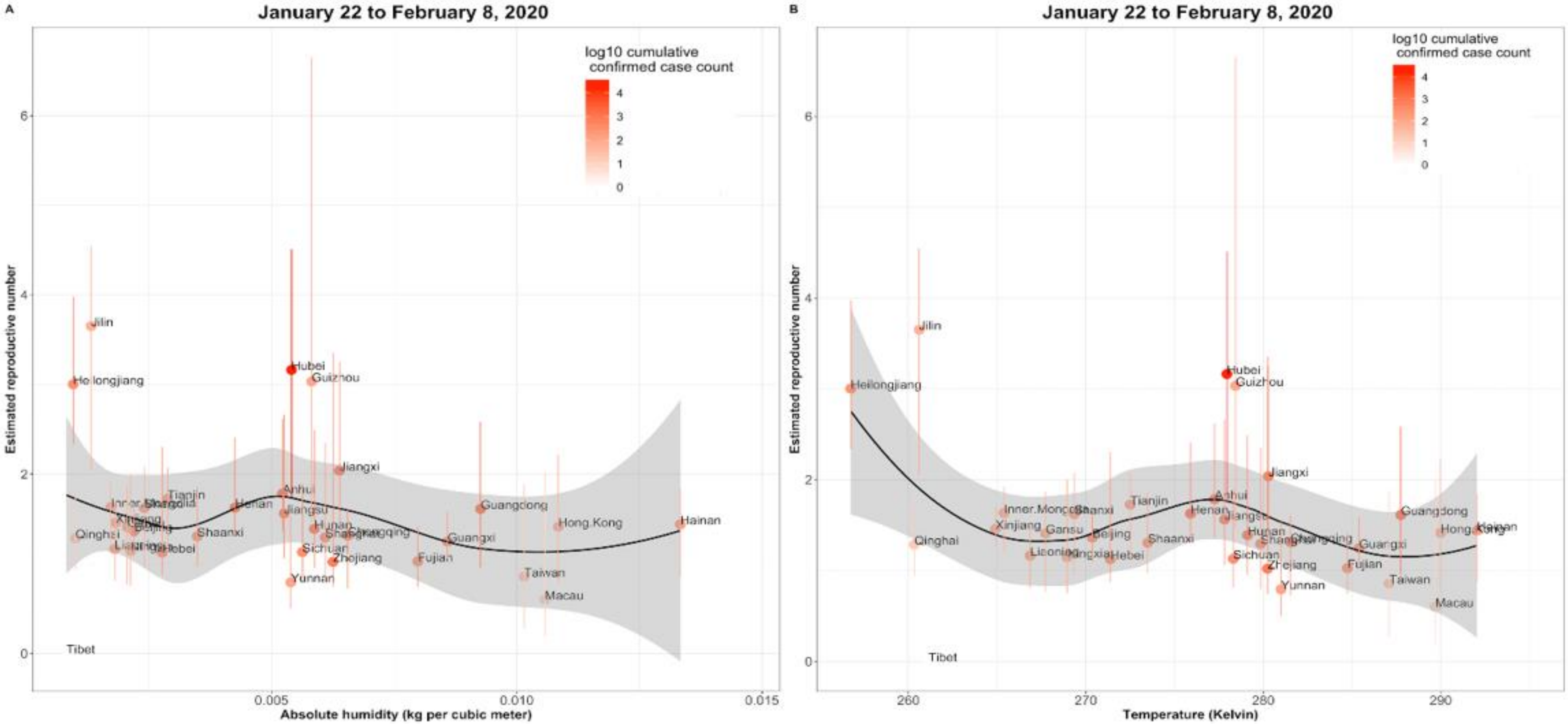
A novel coronavirus (SARS-CoV-2) was identified in Wuhan, Hubei Province, China, in December 2019 and has caused over 240,000 cases of COVID-19 worldwide as of March 19, 2020. Previous studies have supported an epidemiological hypothesis that cold and dry environments facilitate the survival and spread 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.

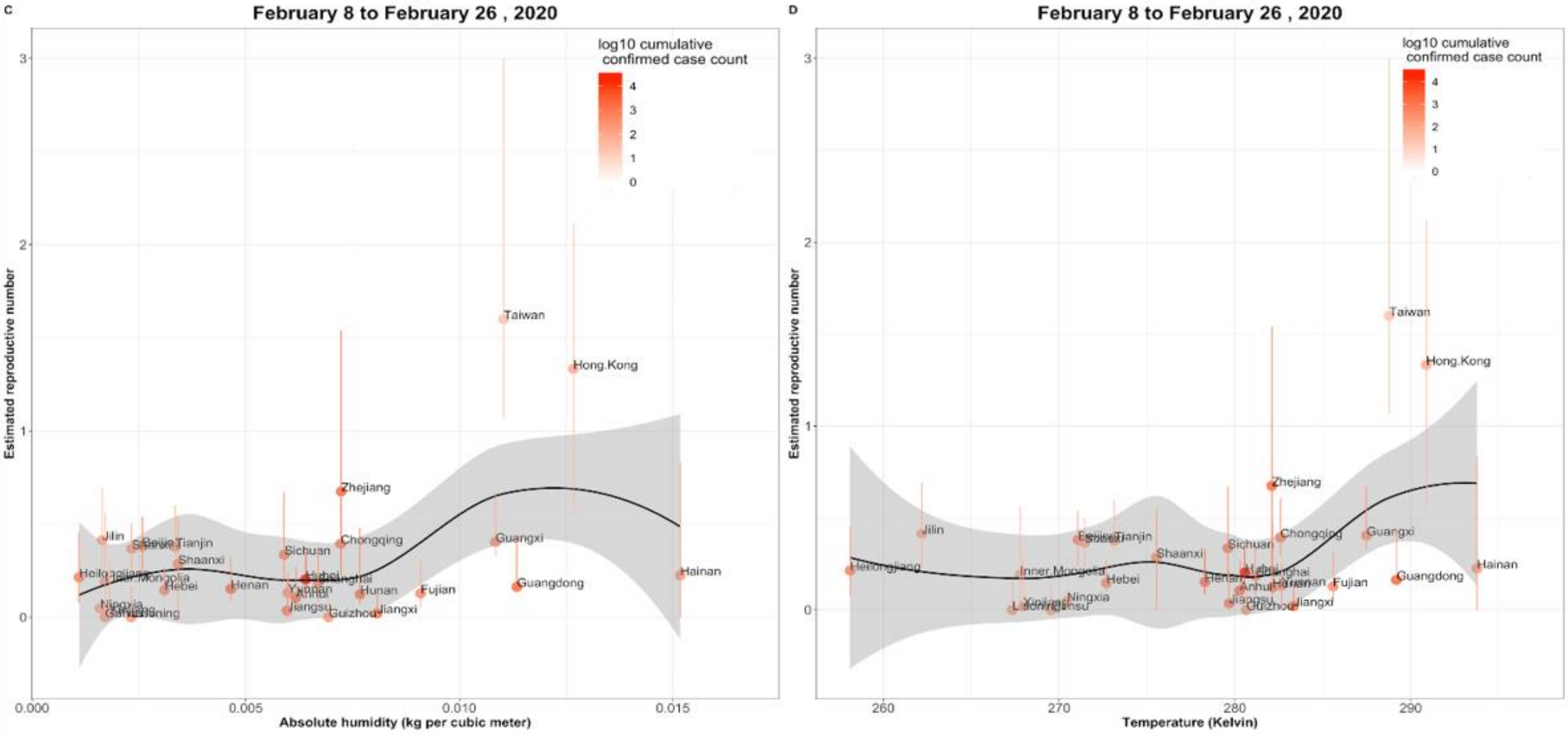
Calculating a proxy for R0

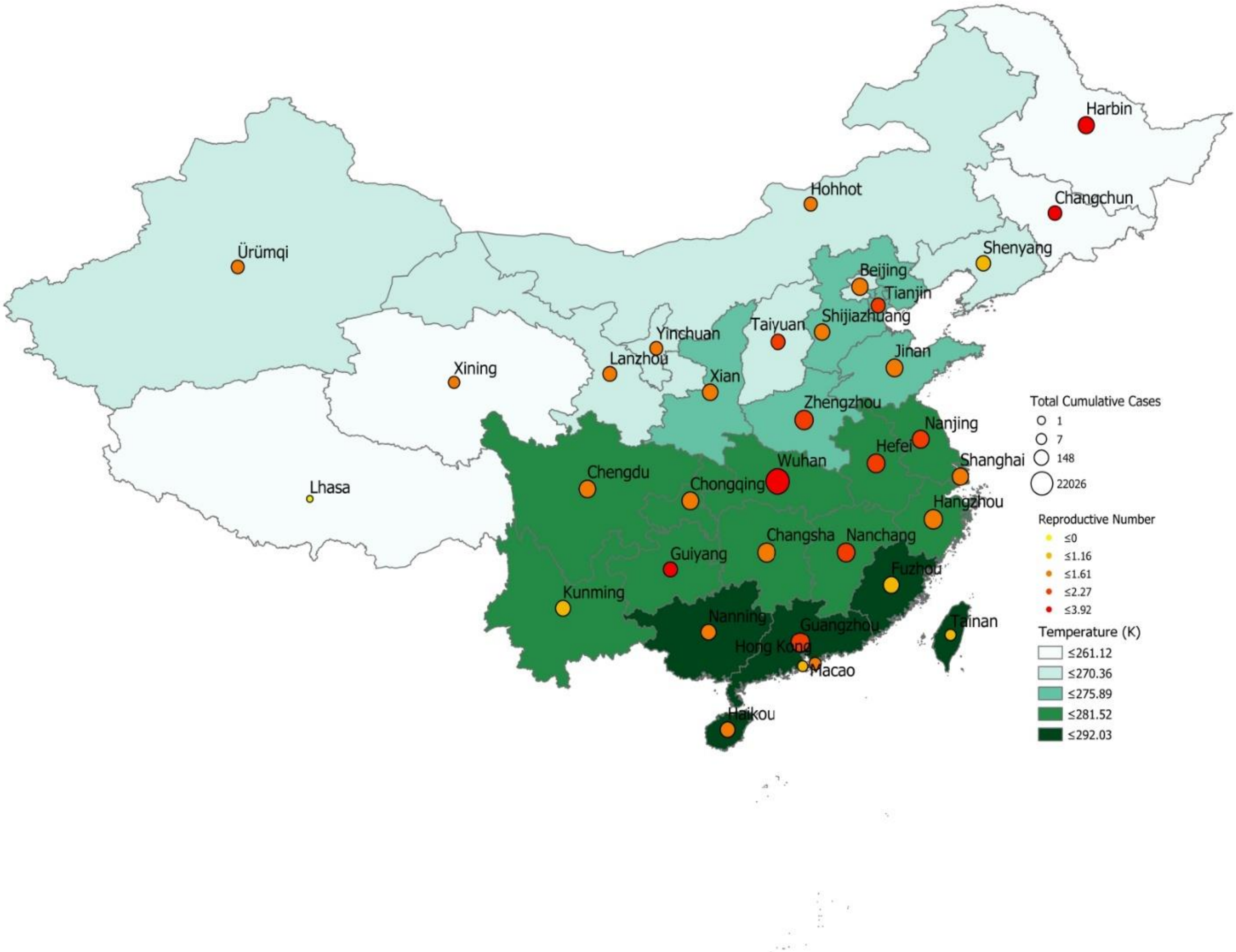


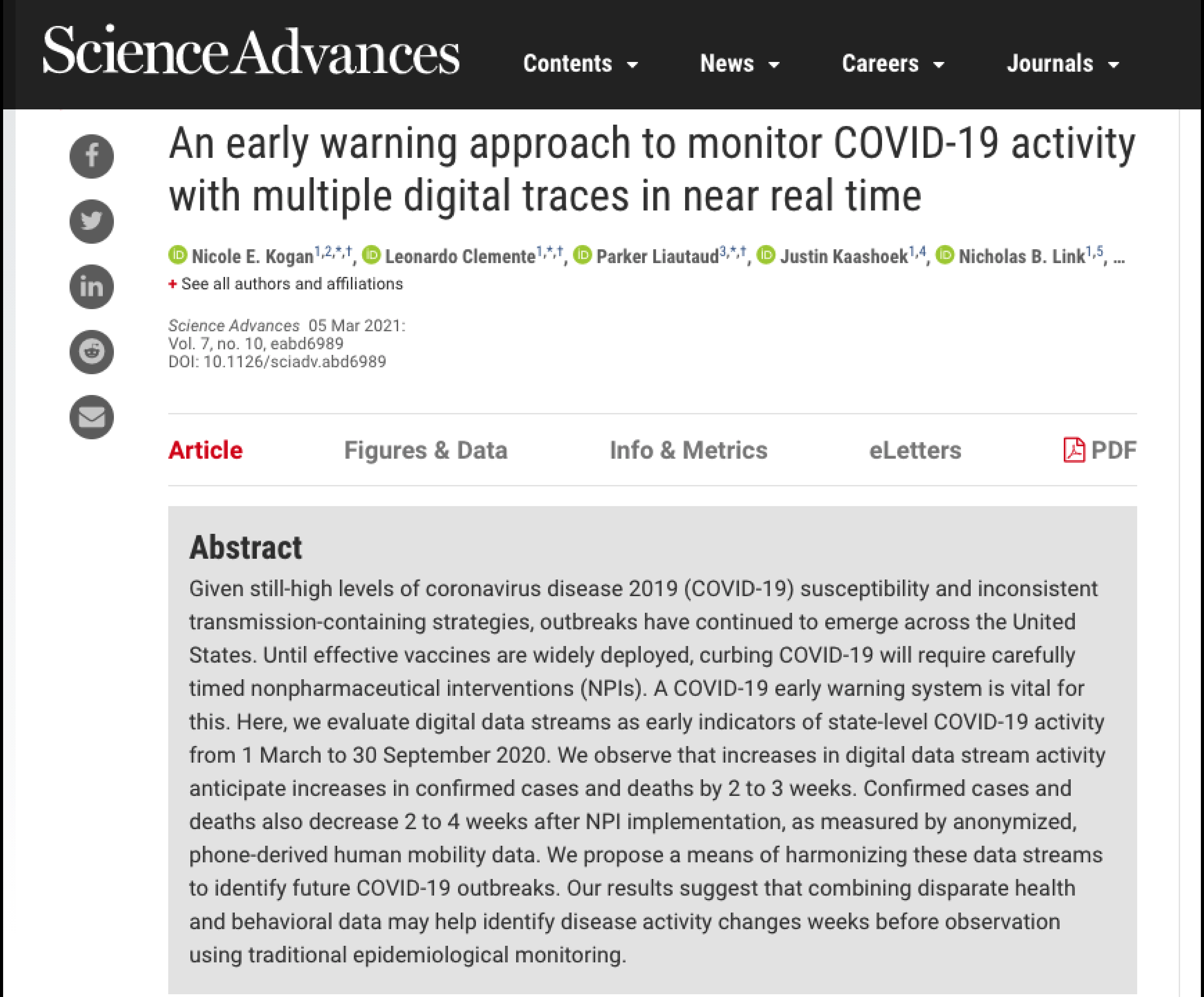
$$R_{proxy}(t,d) = \frac{C(t+2d) - C(t+d)}{C(t+d) - C(t)}$$

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
Inner Mongolia
Xinjiang
Beijing
Qinghai
Liaoning
Hebei
Shaanxi
Henan
Anhui
Jiangsu
Hunan
Shanghai
Chongqing
Sichuan
Zhejiang
Yunnan
Fujian
Guangxi
Guangdong
Hong Kong
Taiwan
Macau
Hainan
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
Inner Mongolia
Shanxi
Tianjin
Xinjiang
Gansu
Beijing
Shaanxi
Henan
Anhui
Jiangsu
Qinghai
Liaoning
Ningxia
Hebei
Hunan
Shanghai
Chongqing
Guangxi
Sichuan
Zhejiang
Yunnan
Fujian
Guangdong
Hong Kong
Hainan
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
Beijing
Tianjin
Shaanxi
Sichuan
Chongqing
Guangxi
Heilongjiang
Inner Mongolia
Hebei
Henan
Yunnan
Anhui
Hubei
Shanghai
Hunan
Fujian
Guangdong
Hainan
Ningxia
Gansu
Liaoning
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
Beijing
Shanxi
Tianjin
Shaanxi
Sichuan
Chongqing
Guangxi
Heilongjiang
Inner Mongolia
Hebei
Henan
Hubei
Shanghai
Anhui
Yunnan
Hunan
Fujian
Guangdong
Hainan
Xinjiang
Ningxia
Liaoning
Gansu
Jiangsu
Guizhou
Jiangxi
260
270
280
290
Temperature (Kelvin)


Harbin
Changchun
Hohhot
Shenyang
Ürümqi
Beijing
Tianjin
Shijiazhuang
Taiyuan
Yinchuan
Lanzhou
Xining
Xian
Jinan
Zhengzhou
Nanjing
Hefei
Wuhan
Shanghai
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



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], ...

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

U

kinsa

cuebiq

GLEAM
GLOBAL EPIDEMIC AND MOBILITY MODEL

COVID-19 gold standards

The COVID Tracking Project

JOHNS HOPKINS UNIVERSITY & MEDICINE | CORONAVIRUS RESOURCE CENTER

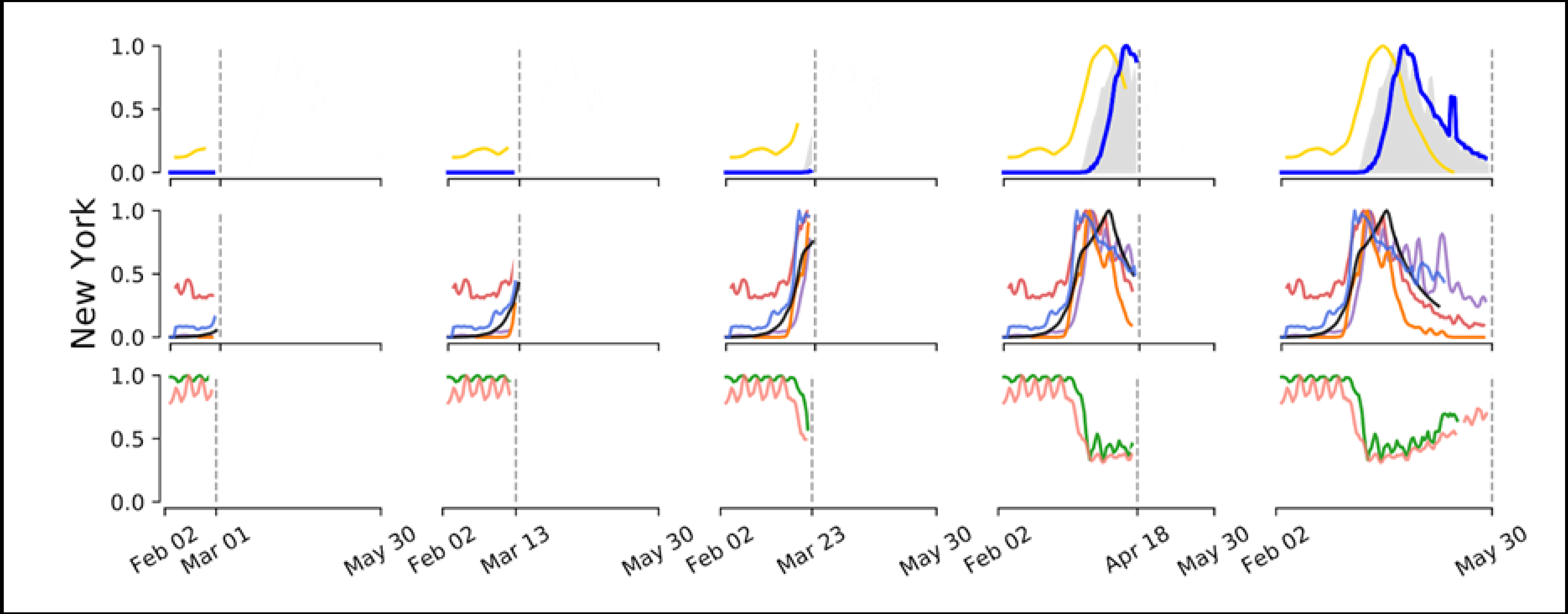
CDC *

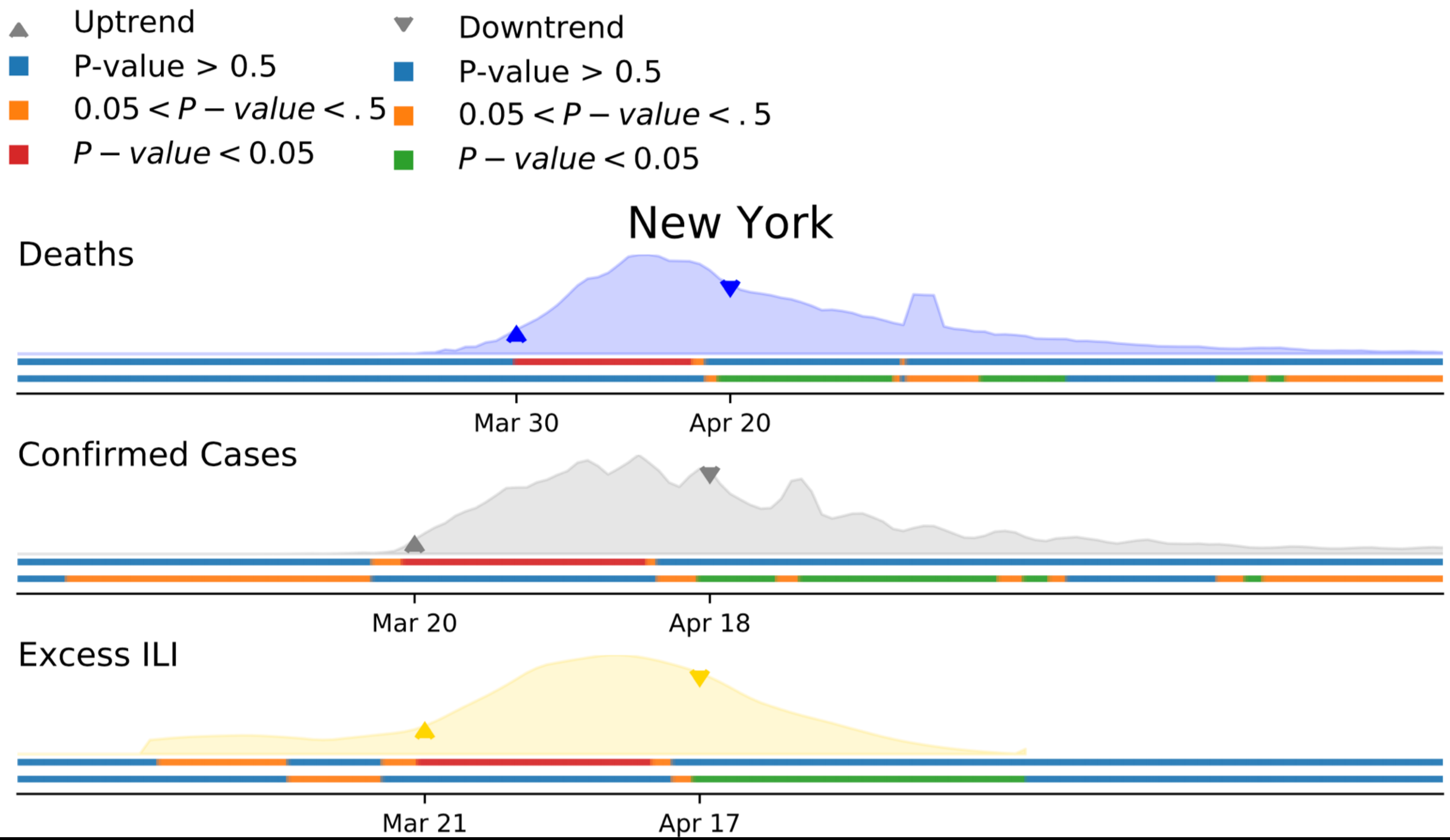
* ILINet

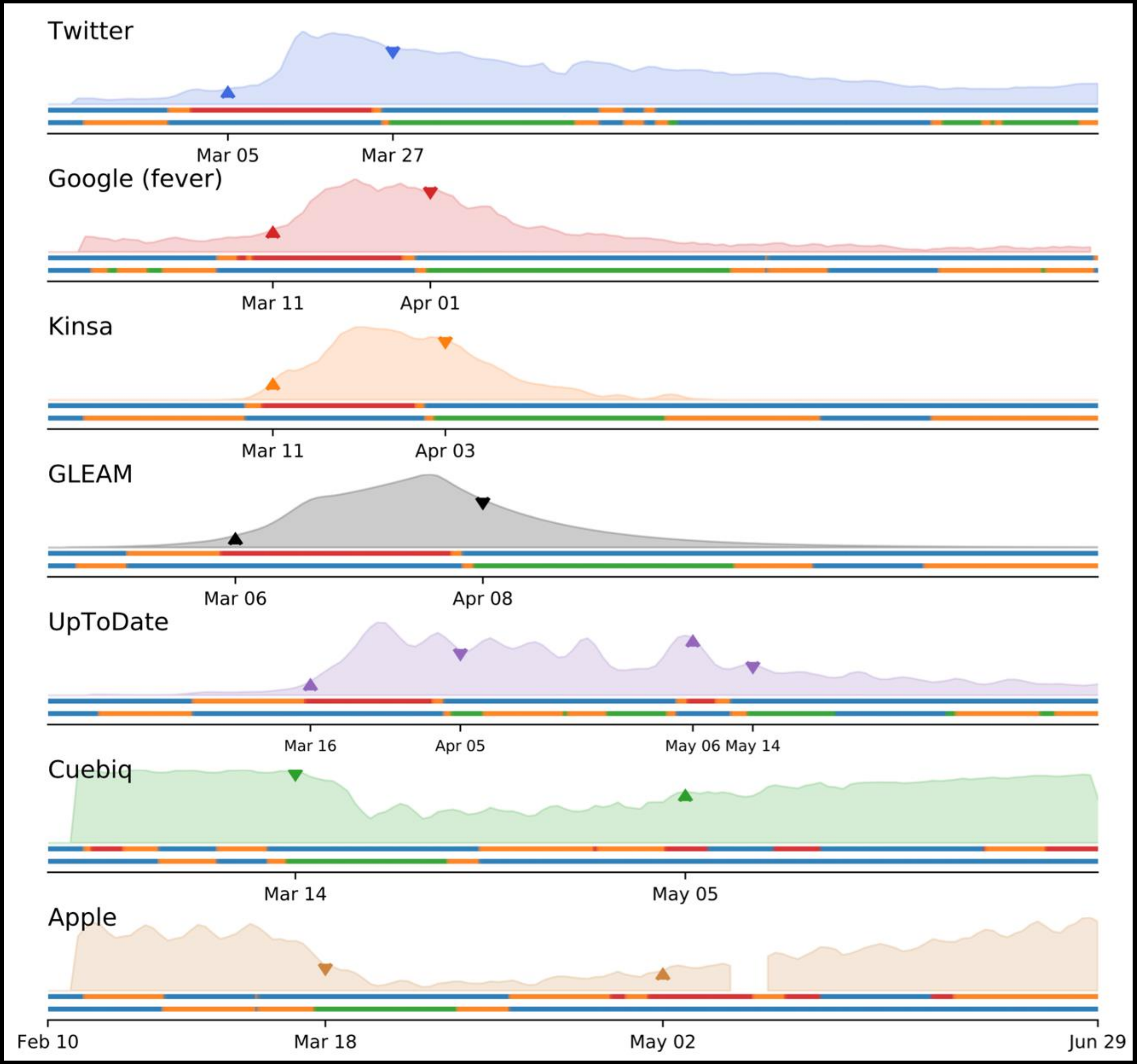
Fig. I - Time series for COVID-19 proxies and gold standards

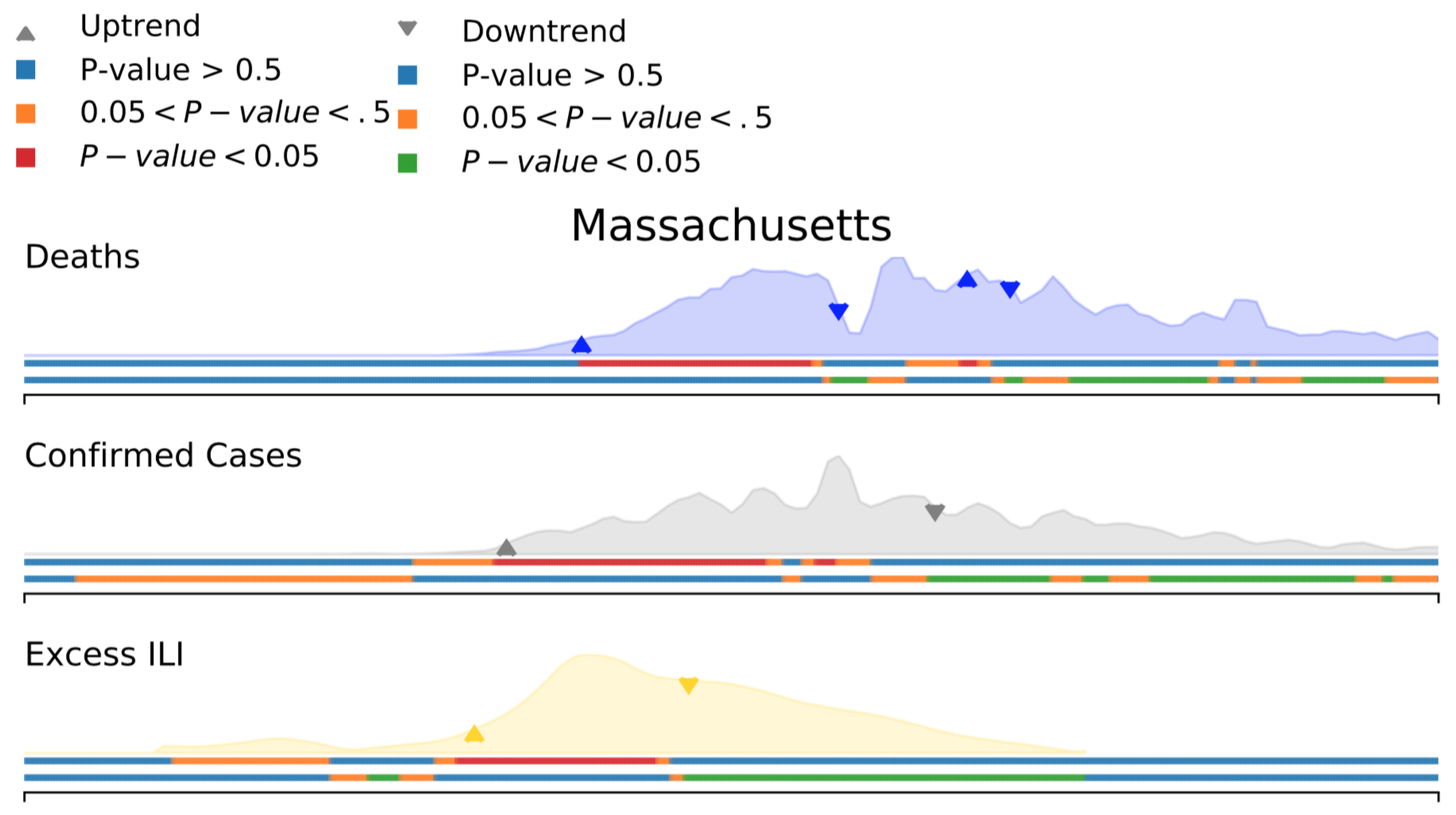
Legend (*delays represent lags in data availability*):

- Excess ILI (Reported Weekly, delayed up to 12 days)
- Confirmed Deaths, (1 day delay)
- Current Date
- Confirmed Cases (1 day delay)
- Google Search Activity (2 day delay)
- UpToDate Search Activity (3 day delay)
- Smart Thermometer Data (1 day delay)
- GLEAM model (No delay)
- Twitter data (2 day delay)
- Cuebiq Mobility (1 day delay)
- Apple Mobility (1 day delay)




▲ 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

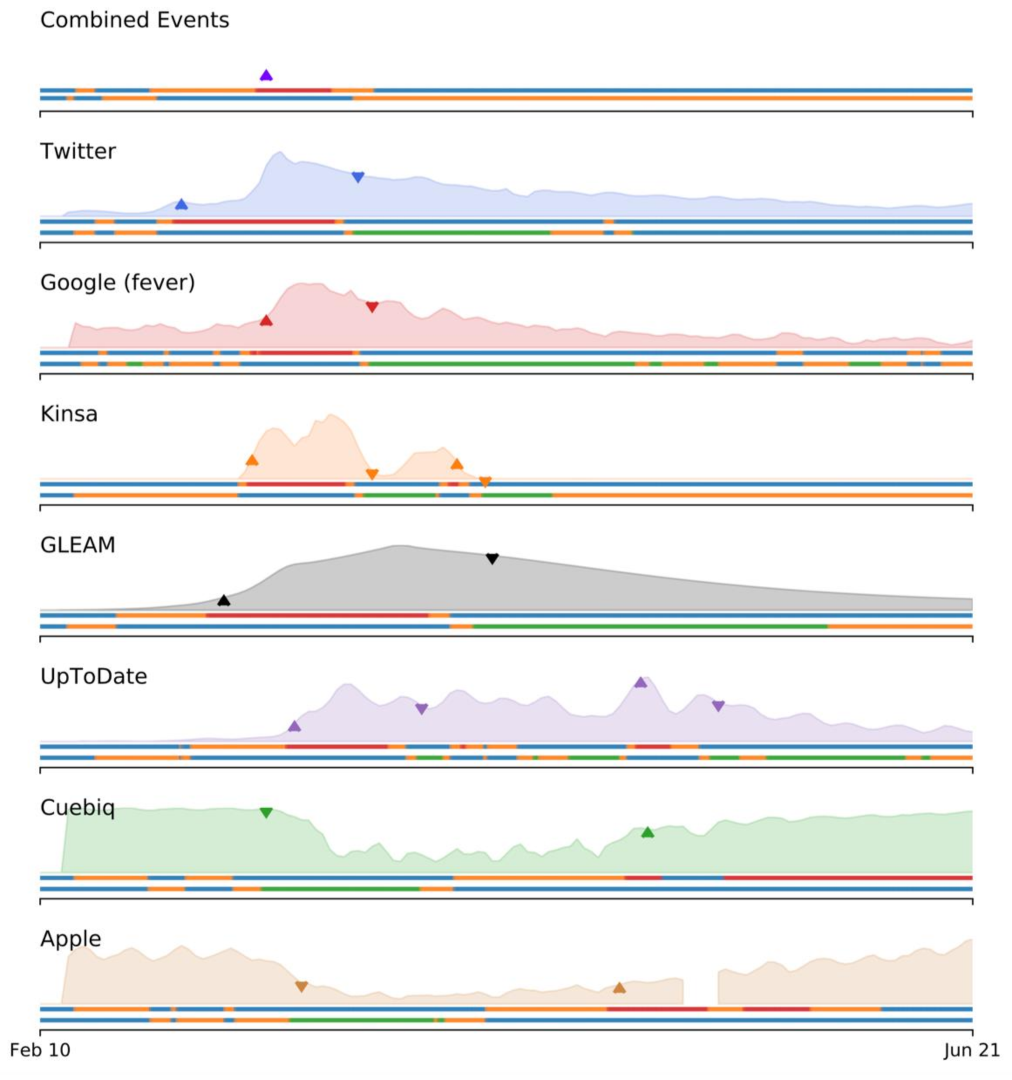

Combined Events
Twitter
Google (fever)
Kinsa
GLEAM
UpToDate
Cuebiq
Apple
Feb 10
Jun 21

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

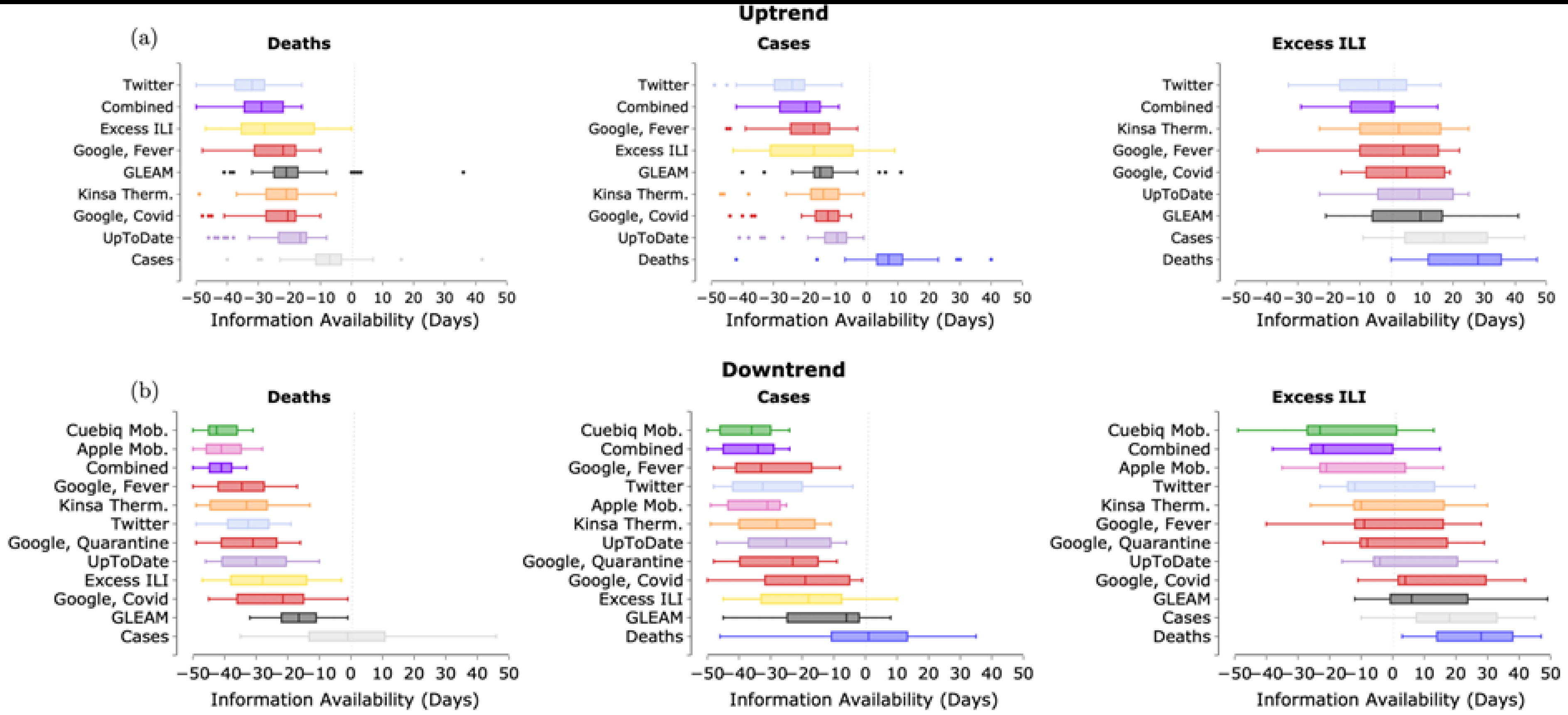
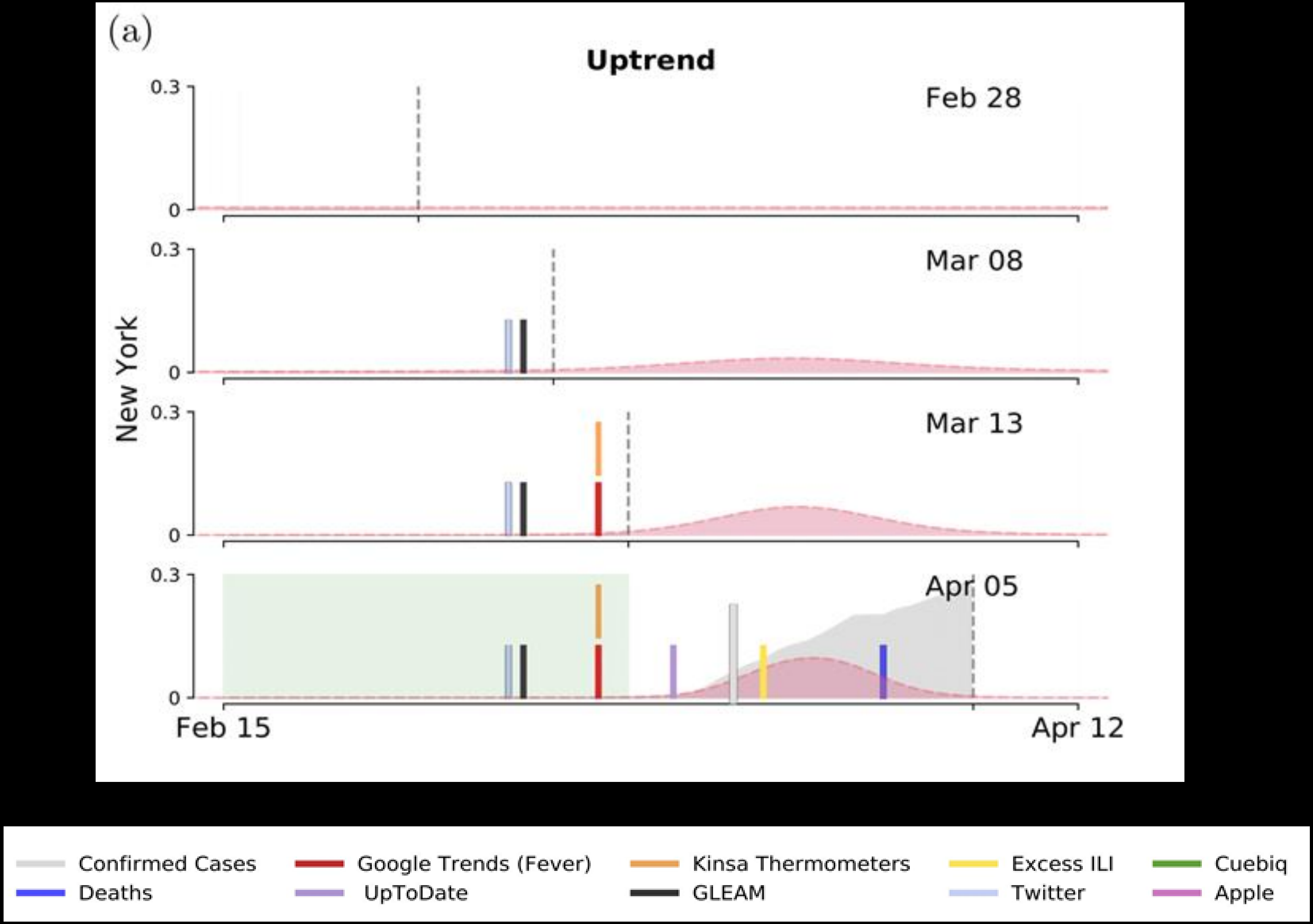
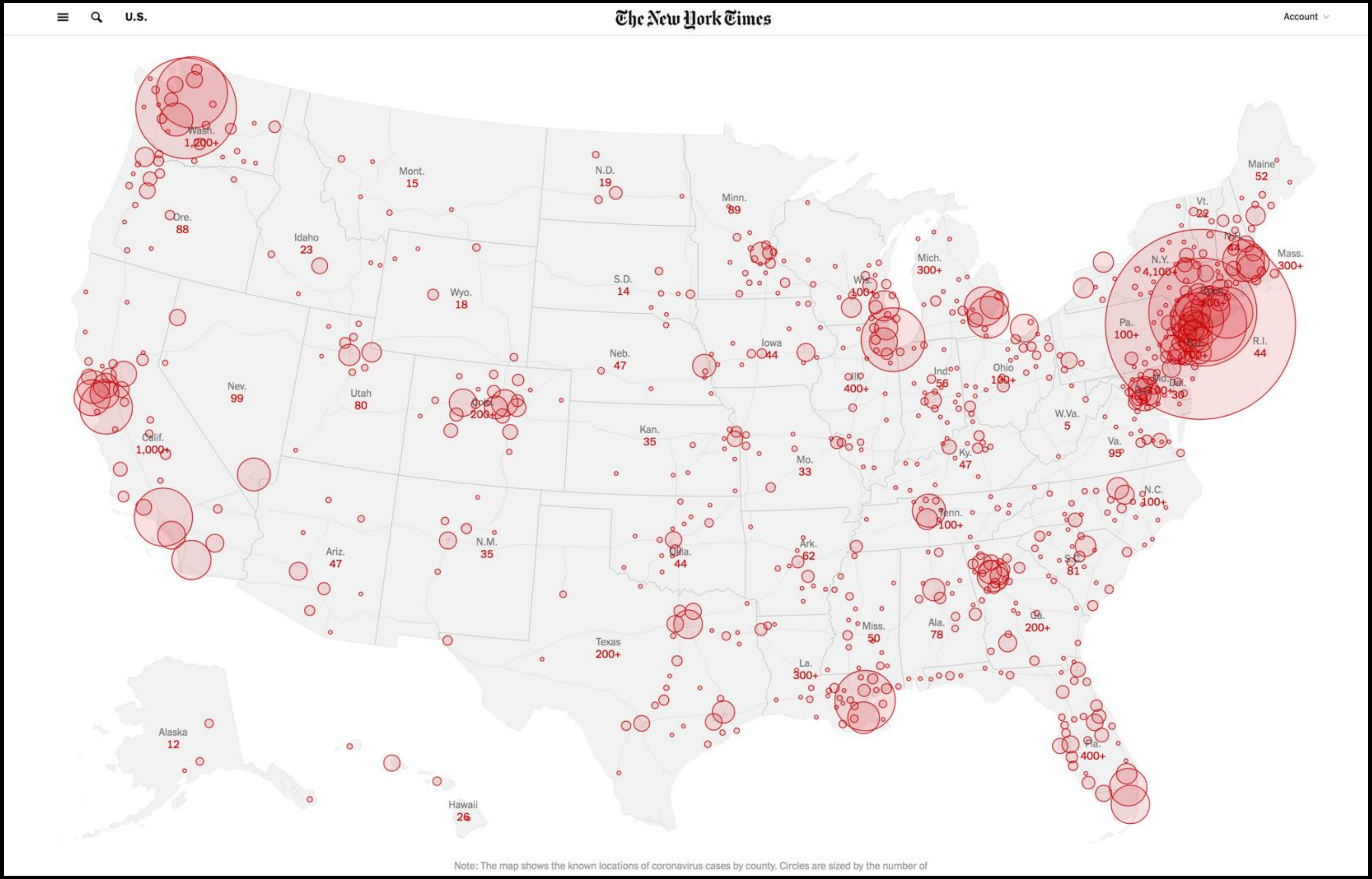


Fig. IV - Evolving posterior probability distribution for time-to-event estimation in New York



Legend:

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



Note: The map shows the known locations of coronavirus cases by county. Circles are sized by the number of

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

RESEARCH ARTICLE

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

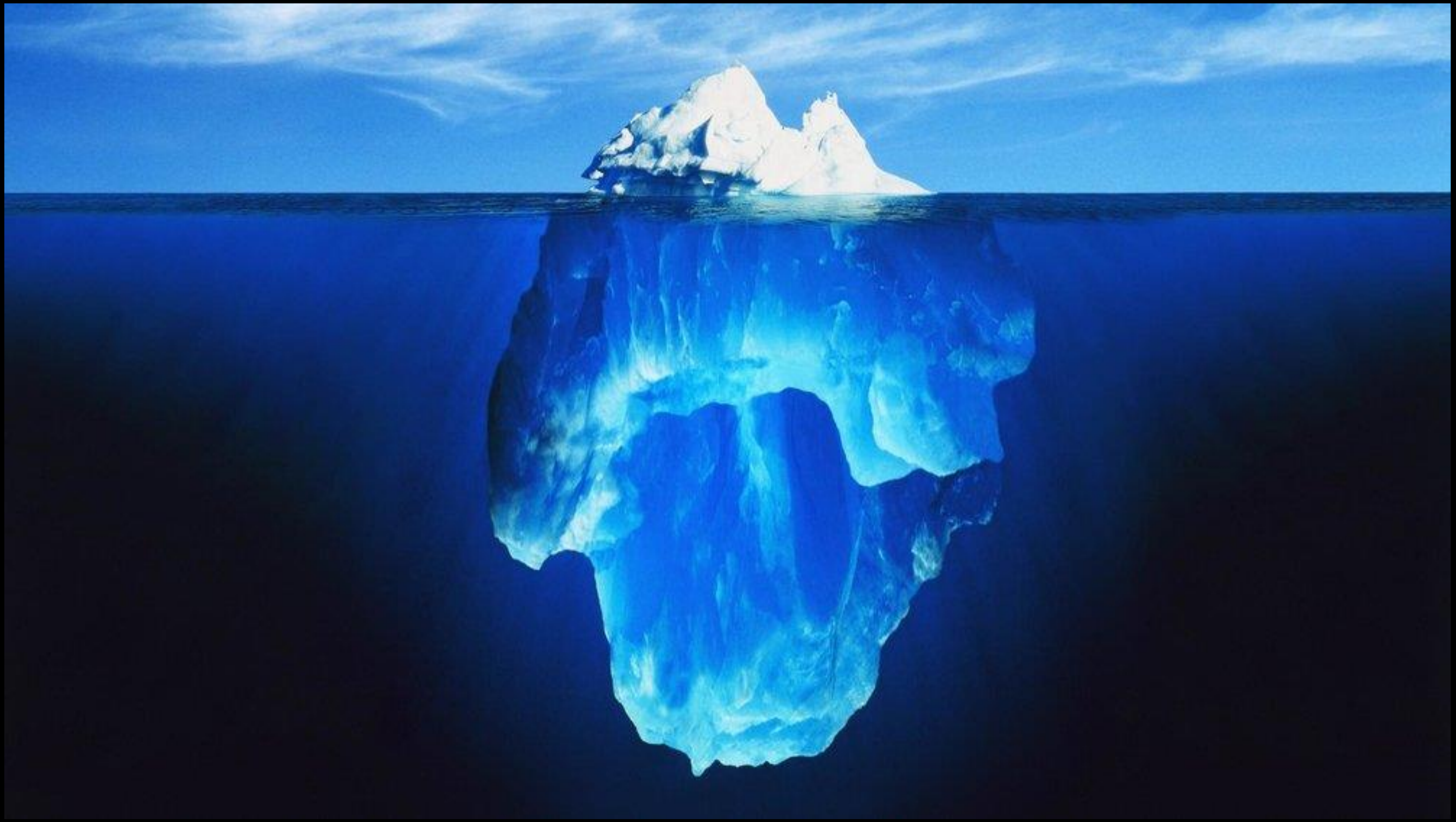
Ruiyun Li[1,*], Sen Pei[2,*,†], Bin Chen[3,*], Yimeng Song[4], Tao Zhang[5], Wan Yang[6], Jeffrey Shaman[2,†]
+ See all authors and affiliations

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

Abstract

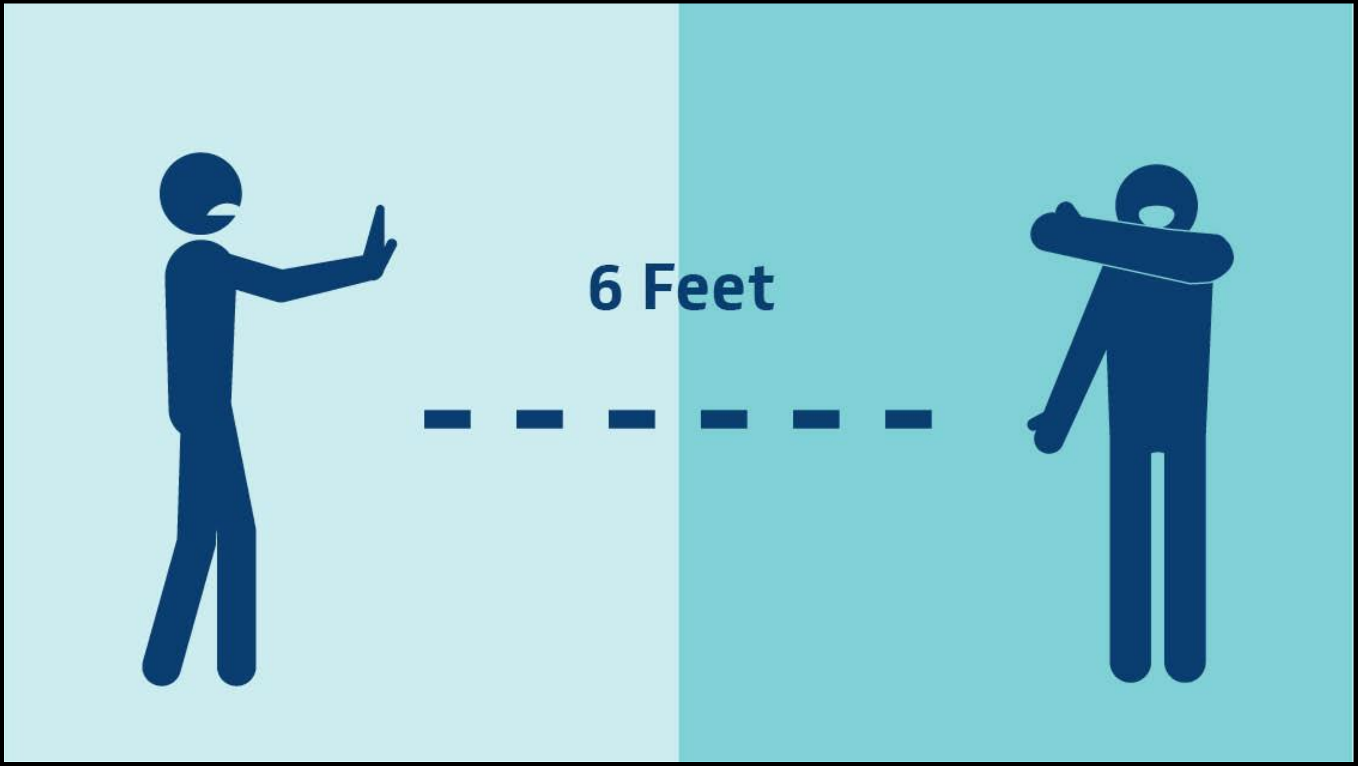
Estimation of the prevalence and contagiousness of undocumented novel coronavirus (SARS-CoV2) infections is critical for understanding the overall prevalence and pandemic potential of this disease. Here we use observations of reported infection within China, in conjunction with mobility data, a networked 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.

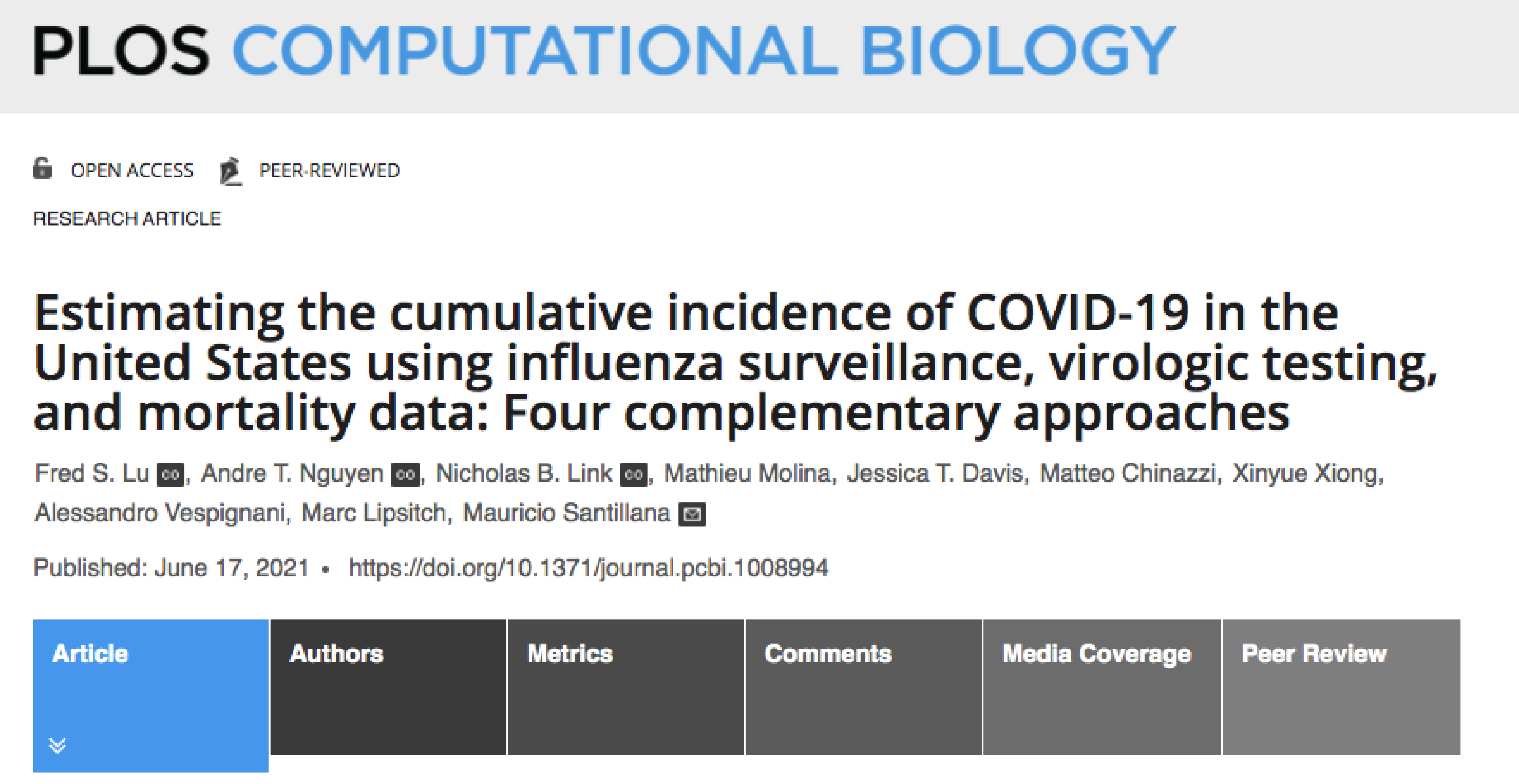
Why should we estimate **prevalence**?



tok resources.org

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



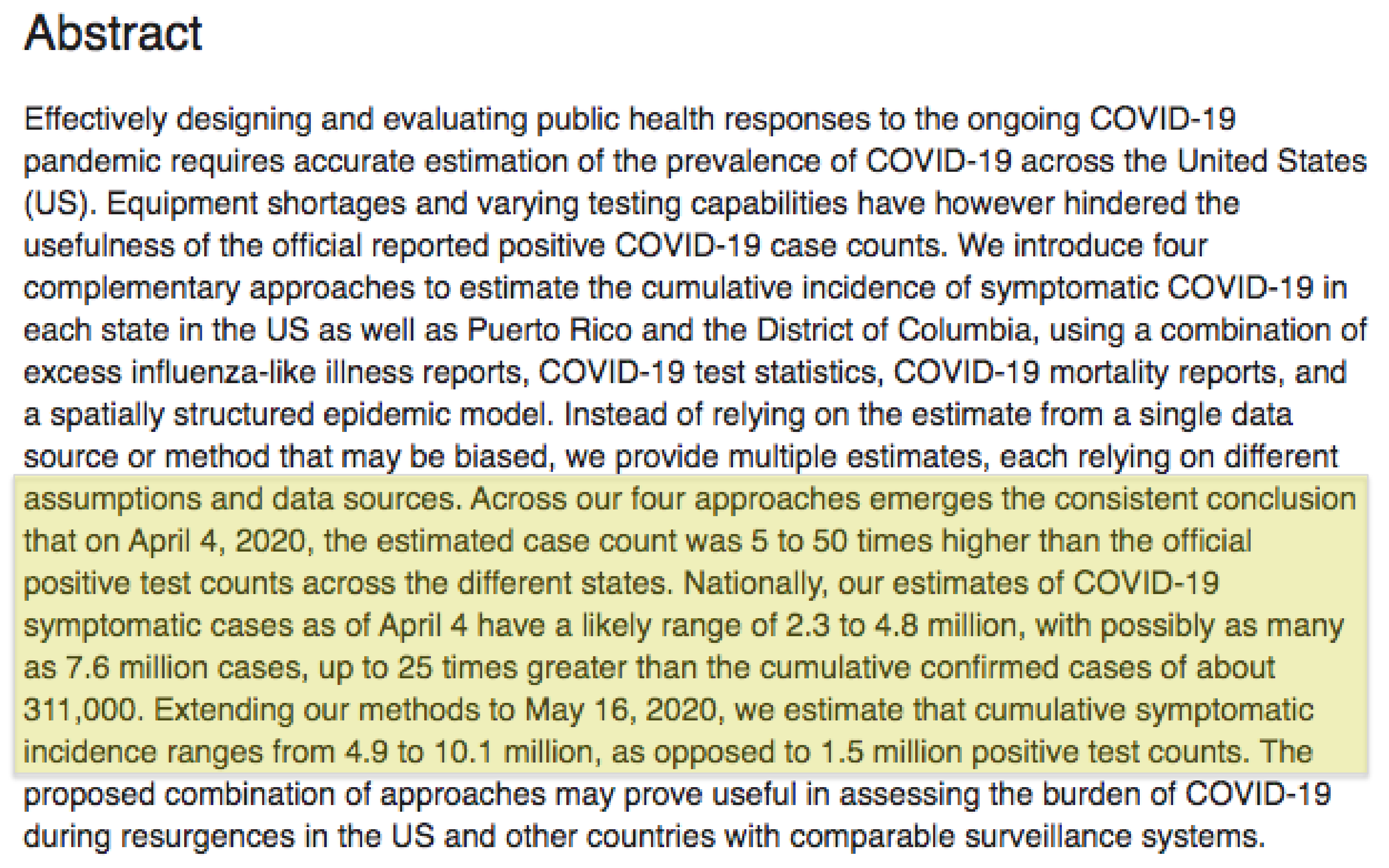

PLOS COMPUTATIONAL BIOLOGY

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

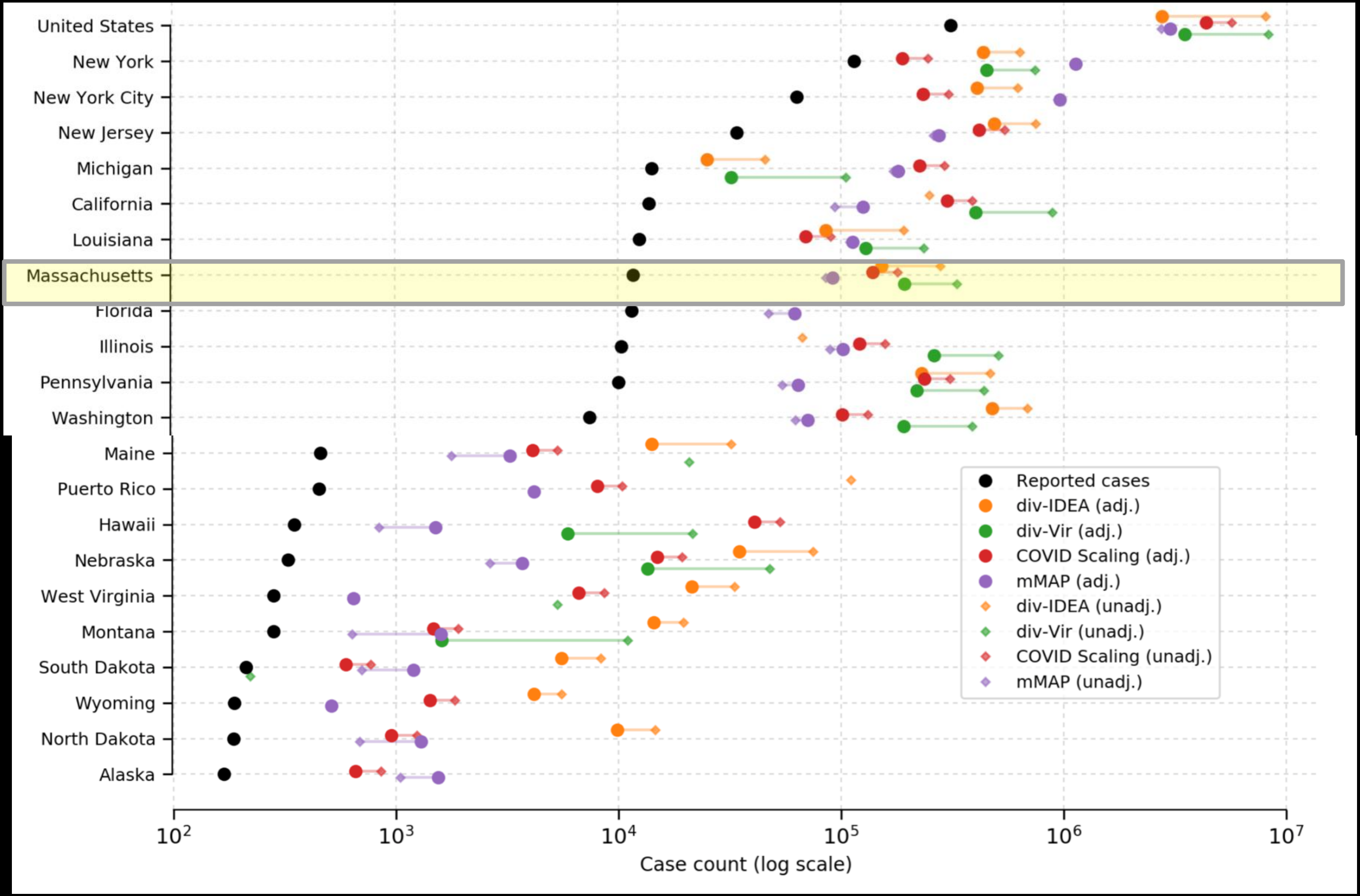


Abstract

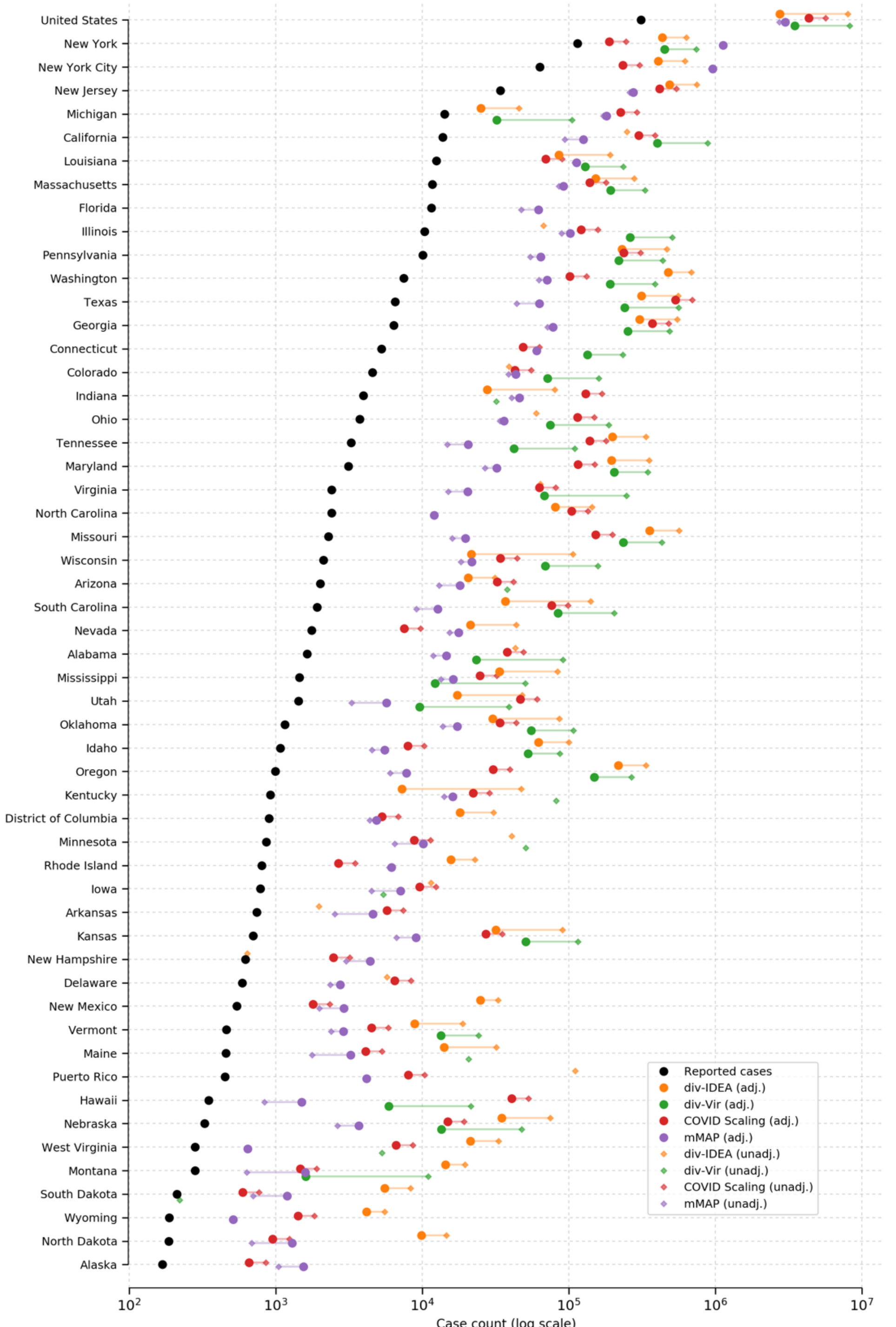
Effectively designing and evaluating public health responses to the ongoing COVID-19 pandemic requires accurate estimation of the prevalence of COVID-19 across the United States (US). Equipment shortages and varying testing capabilities have however hindered the usefulness of the official reported positive 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

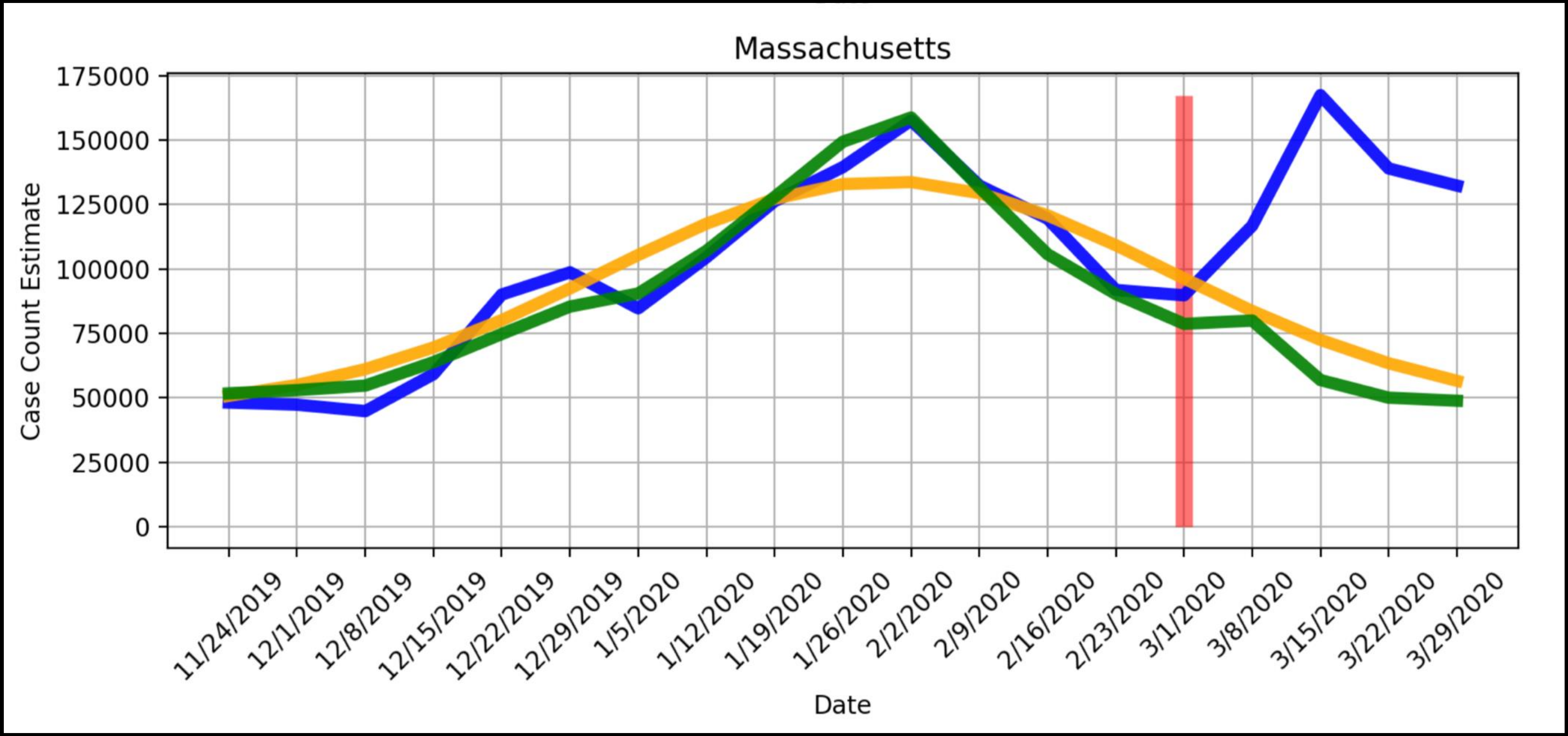


As of April 4th, 2020



As of April 4th, 2020

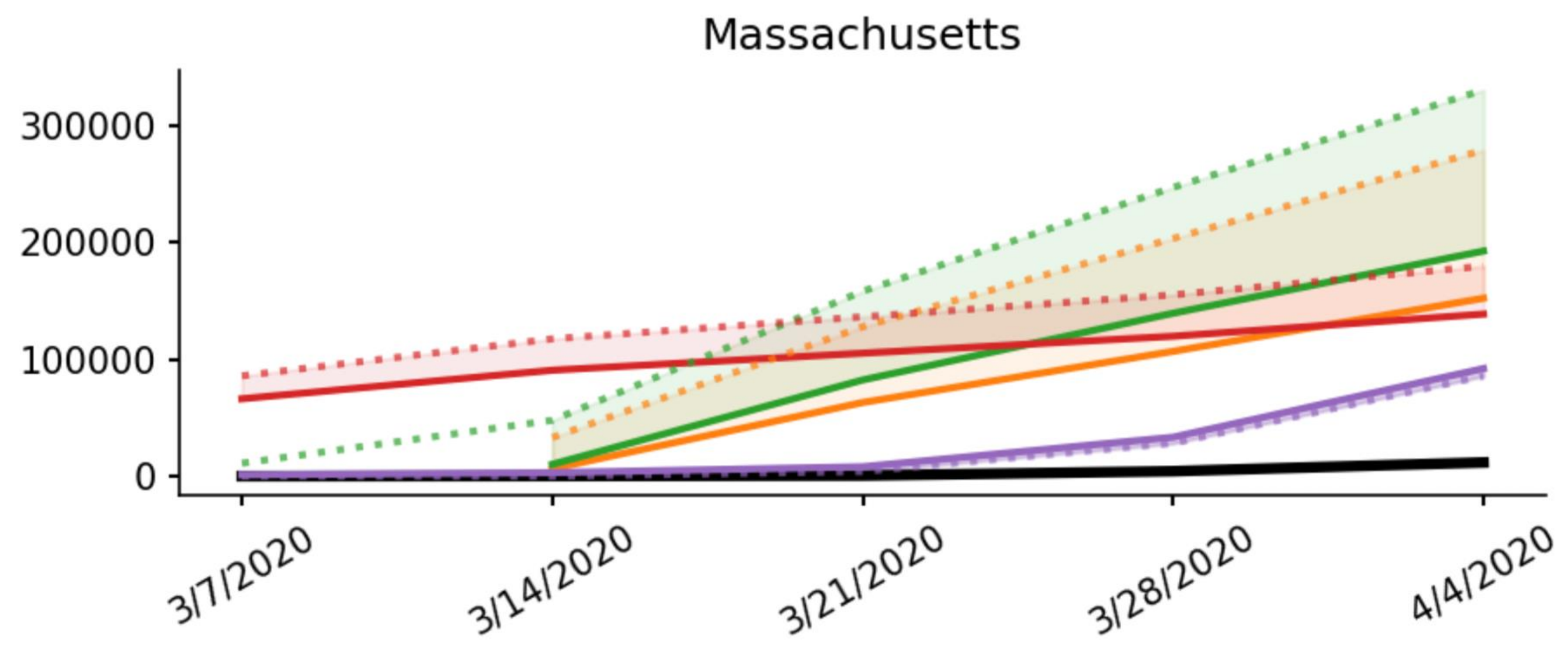
In MA we estimate about **120,000** COVID-19 infected
Compare to



As of April 4th, 2020

As of April 16th, 2020

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



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.

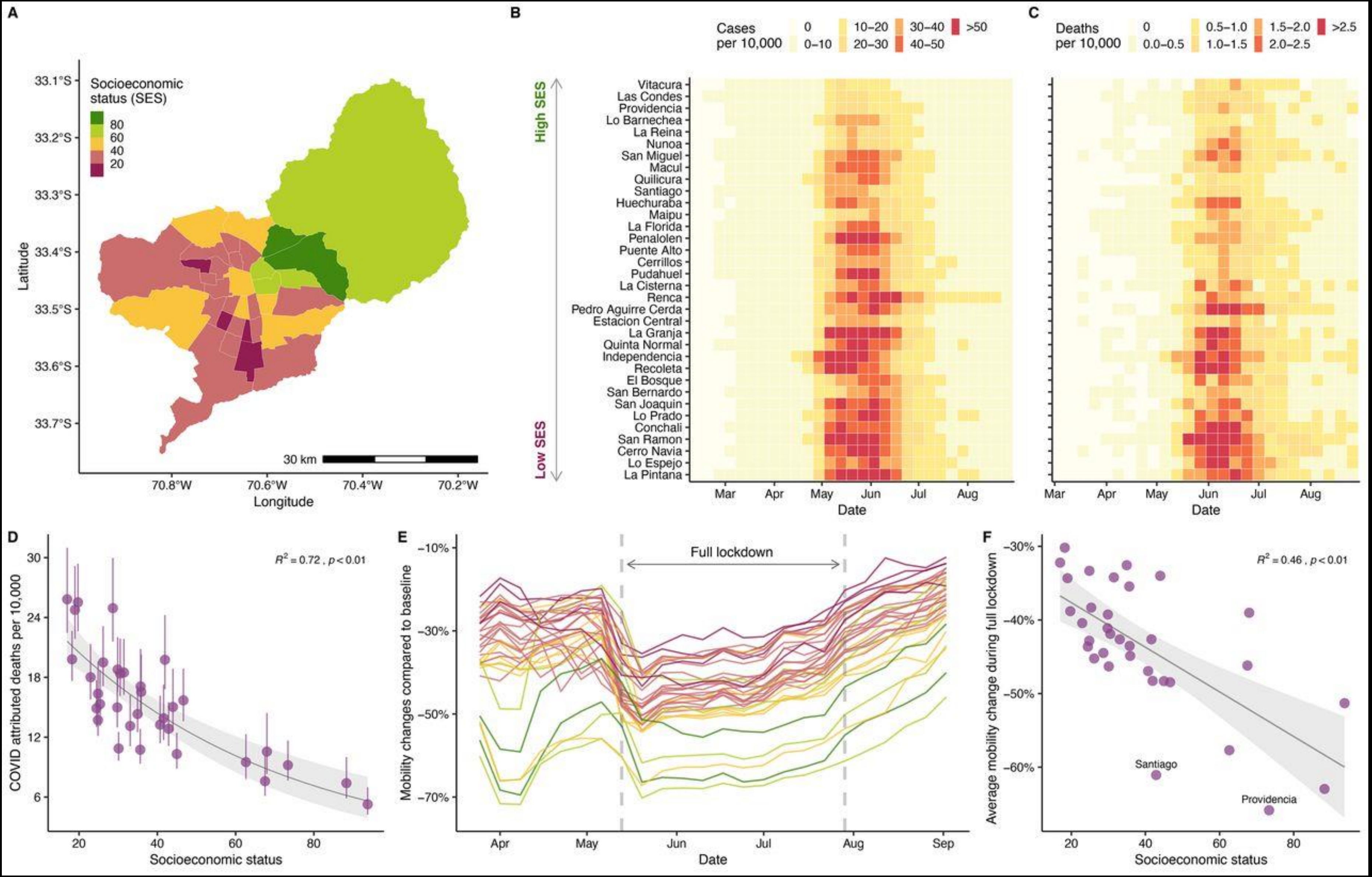
Pamela Martinez
University of Illinois
Urbana-Champaign

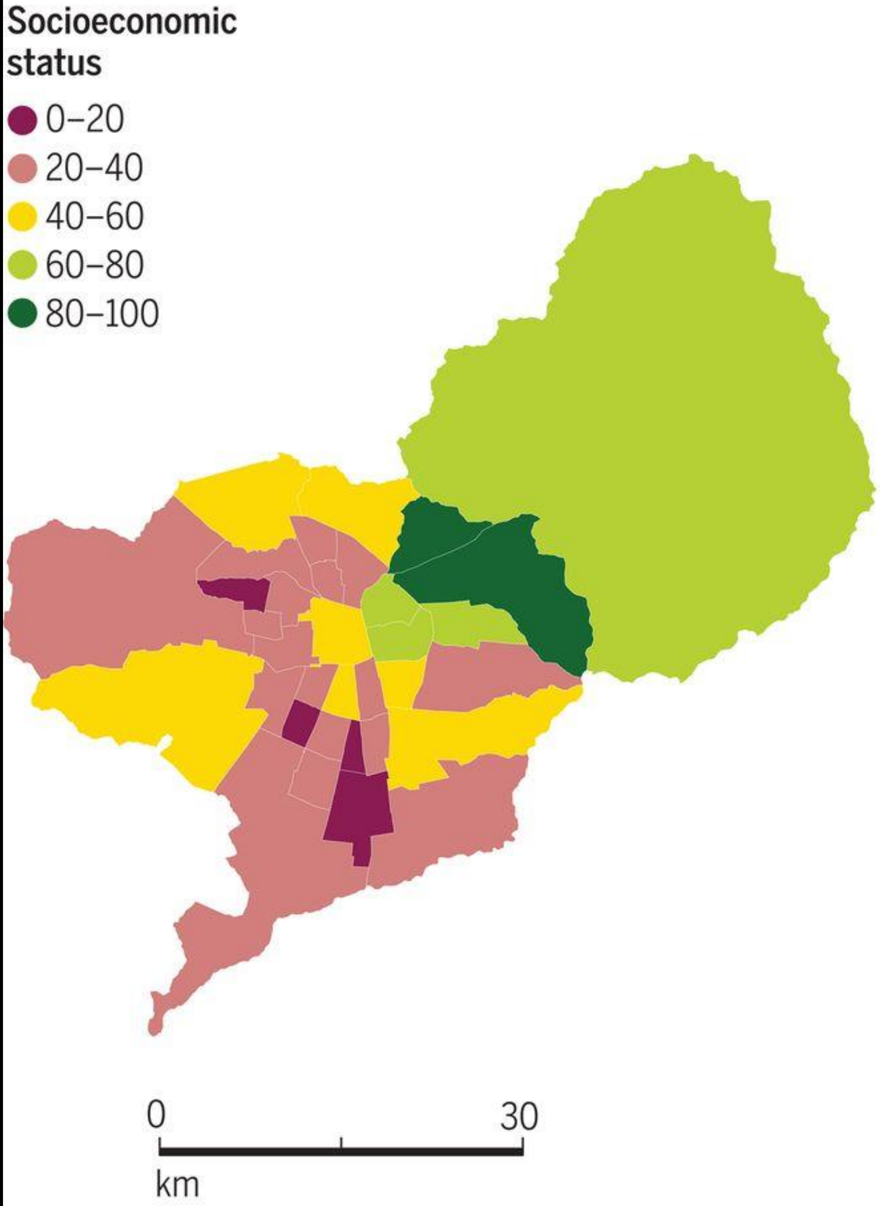
Gonzalo Mena
University of Oxford

Ayesha Mahmud
University of
California, Berkeley

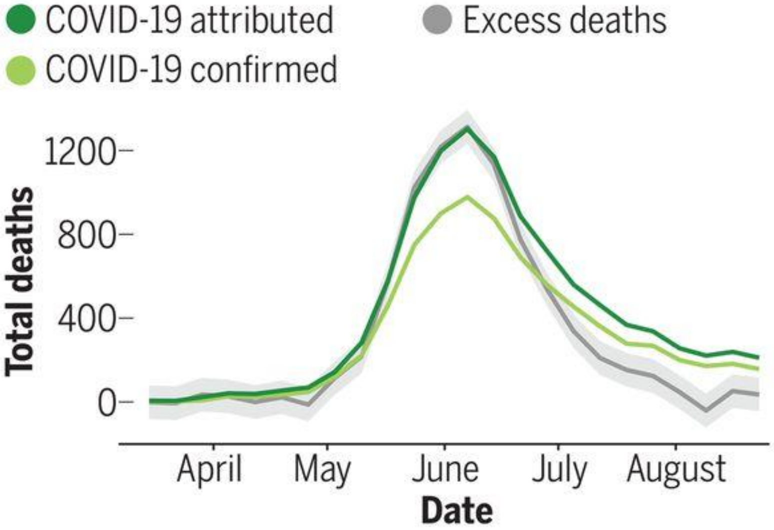
Pablo Marquet
Pontificia Universidad
Catolica de Chile

Caroline Buckee
Harvard University

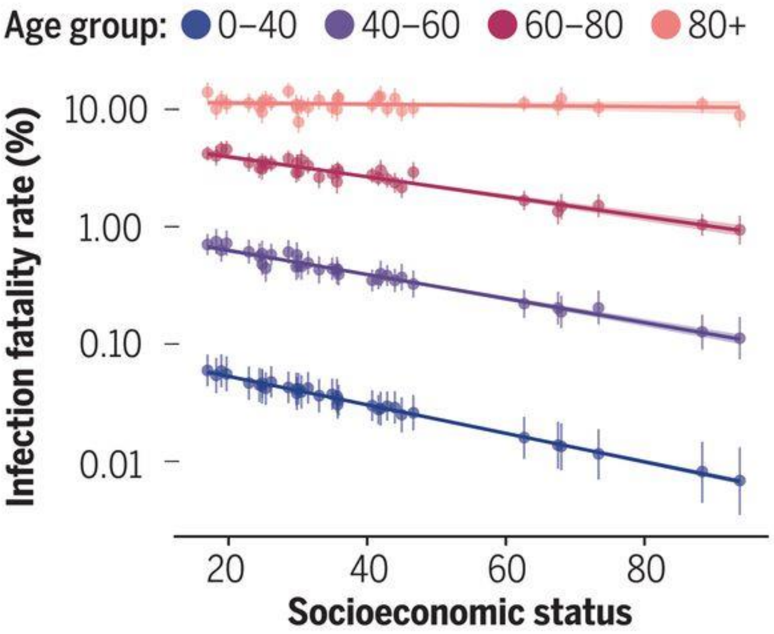

A
Socioeconomic status (SES)
80
60
40
20
33.1°S
33.2°S
33.3°S
33.4°S
33.5°S
33.6°S
33.7°S
Latitude
30 km
70.8°W
70.6°W
70.4°W
70.2°W
Longitude
B
Cases per 10,000
0
0–10
10–20
20–30
30–40
40–50
>50
High SES
Low SES
Vitacura
Las Condes
Providencia
Lo Barnechea
La Reina
Nunoa
San Miguel
Macul
Quilicura
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
$R^2 = 0.72$, $p < 0.01$
30
24
18
12
6
COVID attributed deaths per 10,000
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



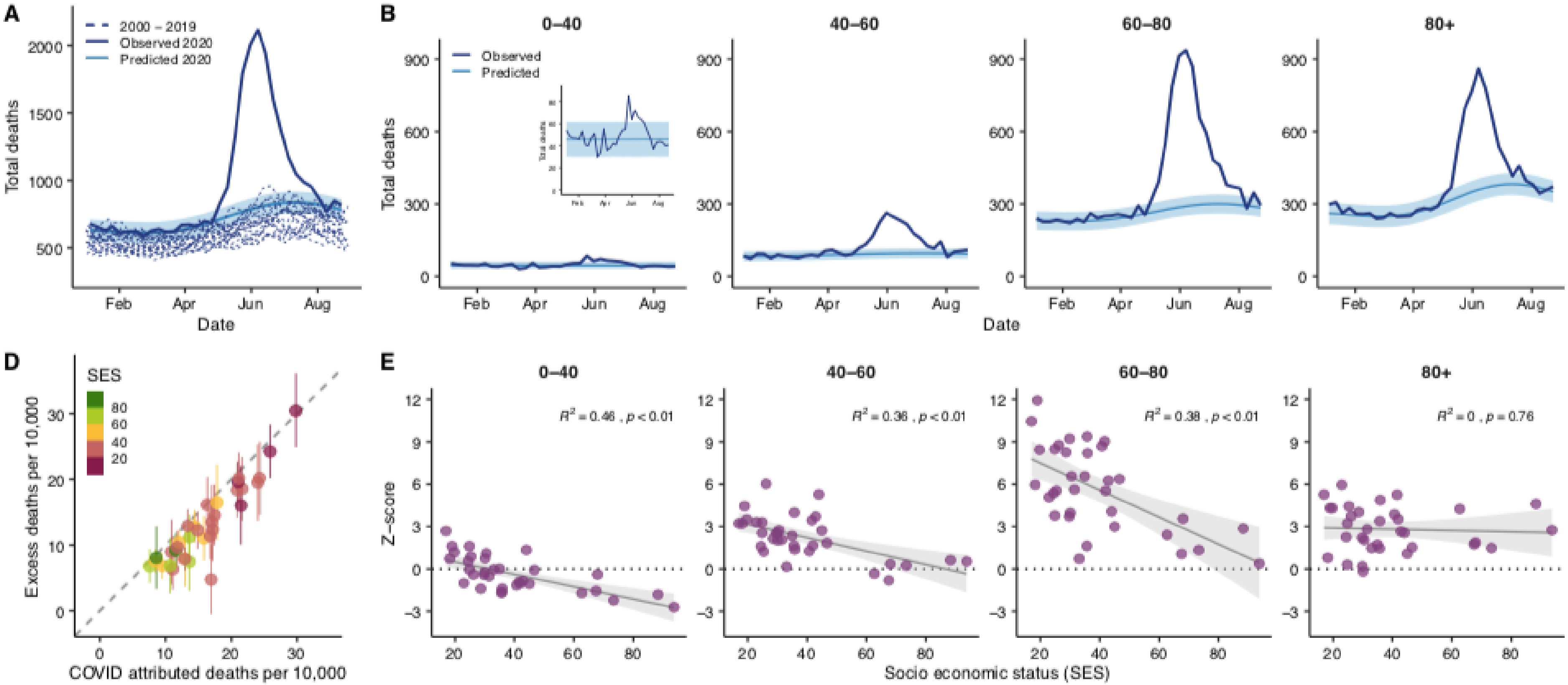
Municipalities of the Greater Santiago area of Chile

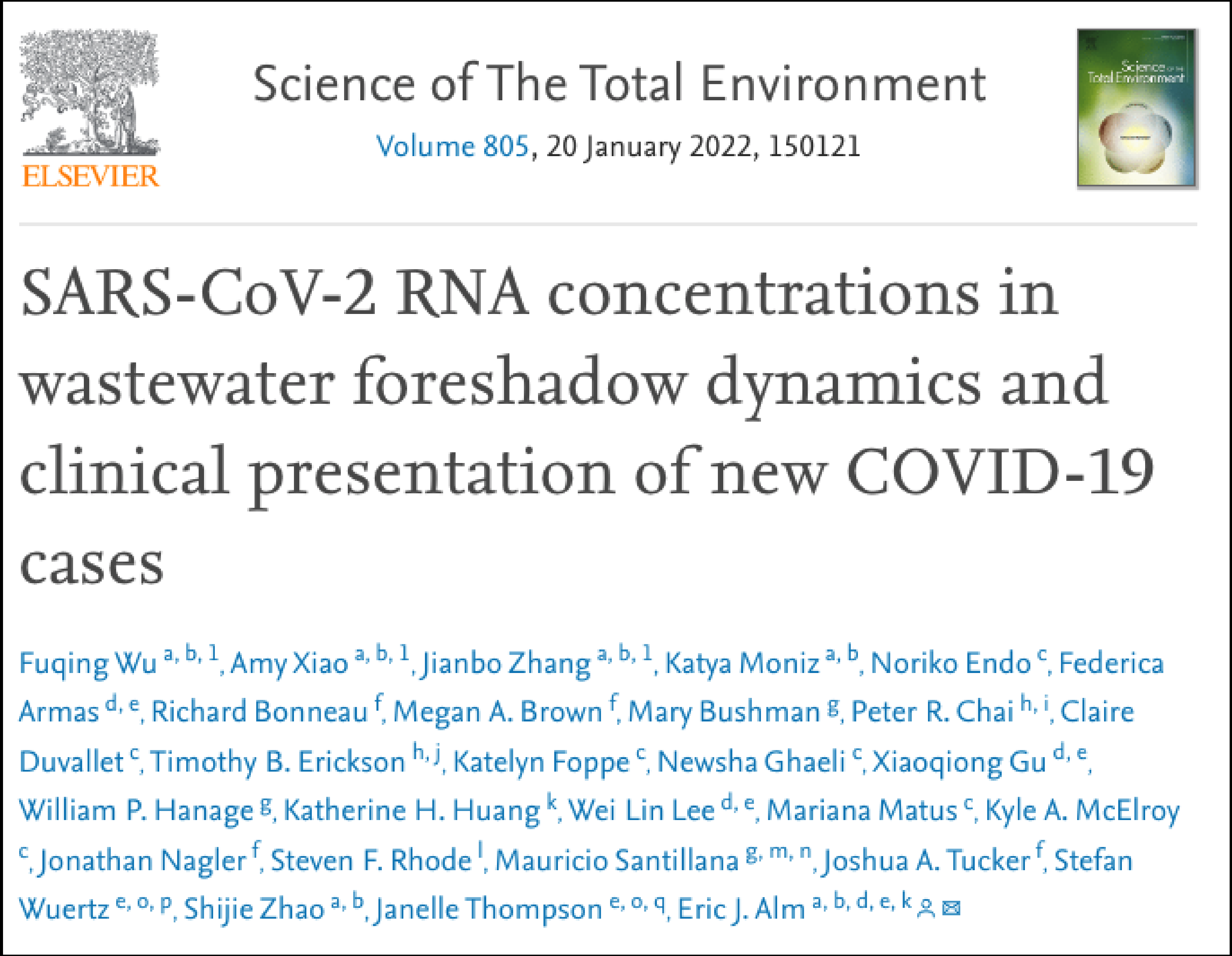


Comparison of COVID-19 deaths with excess deaths for the Greater Santiago area



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
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)



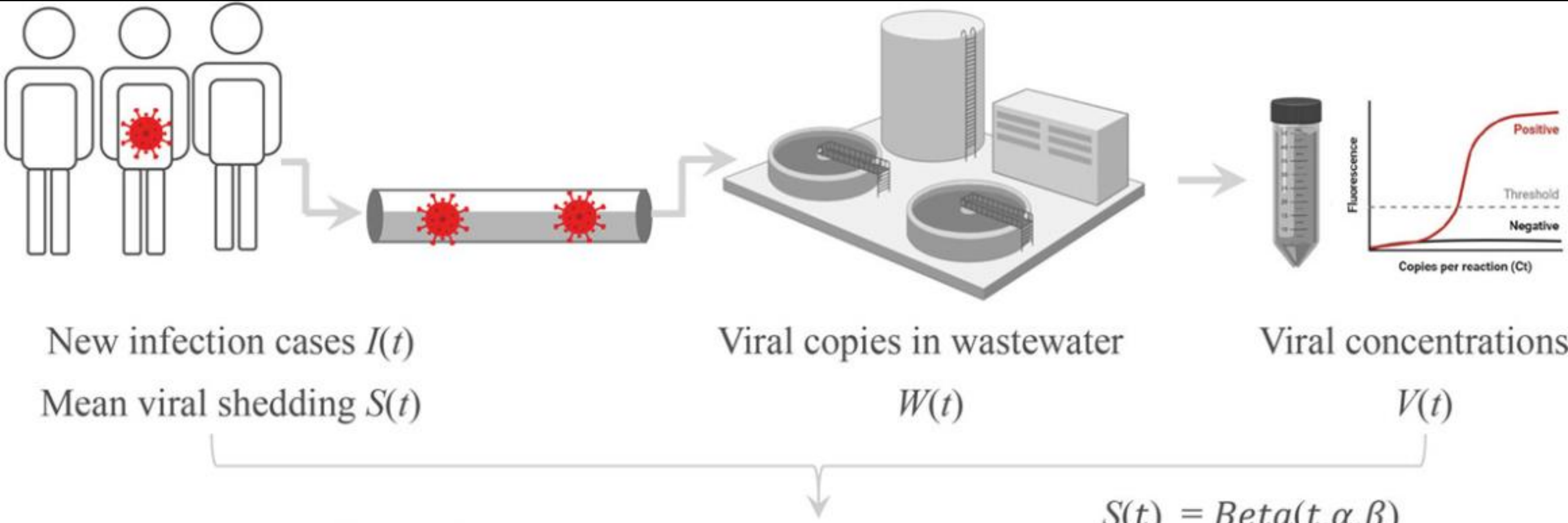
ELSEVIER

Science of The Total Environment

Volume 805, 20 January 2022, 150121

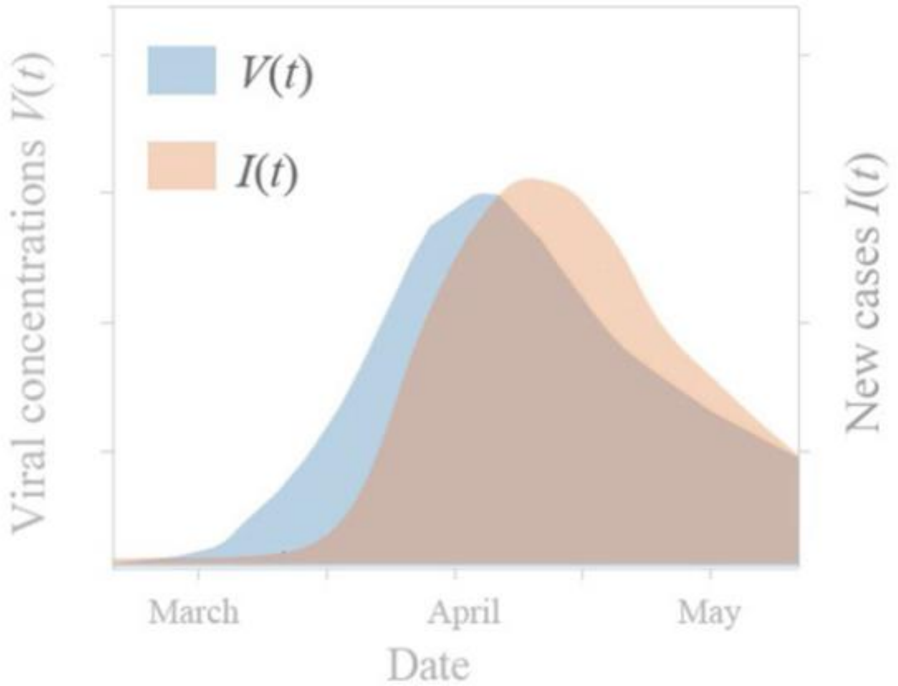
SARS-CoV-2 RNA concentrations in wastewater foreshadow dynamics and clinical presentation of new COVID-19 cases

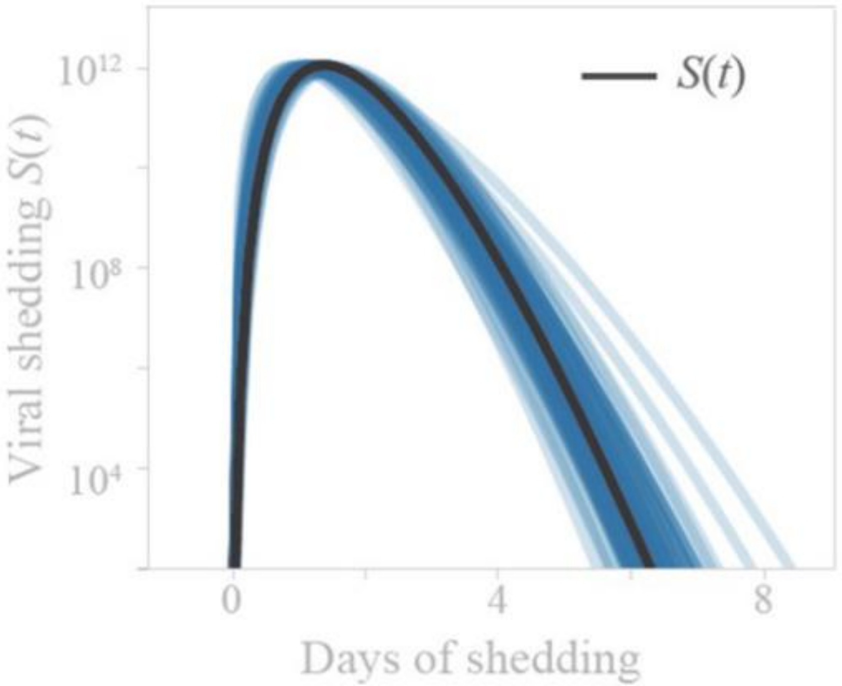
Fuqing Wu [a, b, 1], Amy Xiao [a, b, 1], Jianbo Zhang [a, b, 1], Katya Moniz [a, b], Noriko Endo [c], Federica Armas [d, e], Richard Bonneau [f], Megan A. Brown [f], Mary Bushman [g], Peter R. Chai [h, i], Claire Duvallet [c], Timothy B. Erickson [h, j], Katelyn Foppe [c], Newsha Ghaeli [c], Xiaoqiong 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] ✉



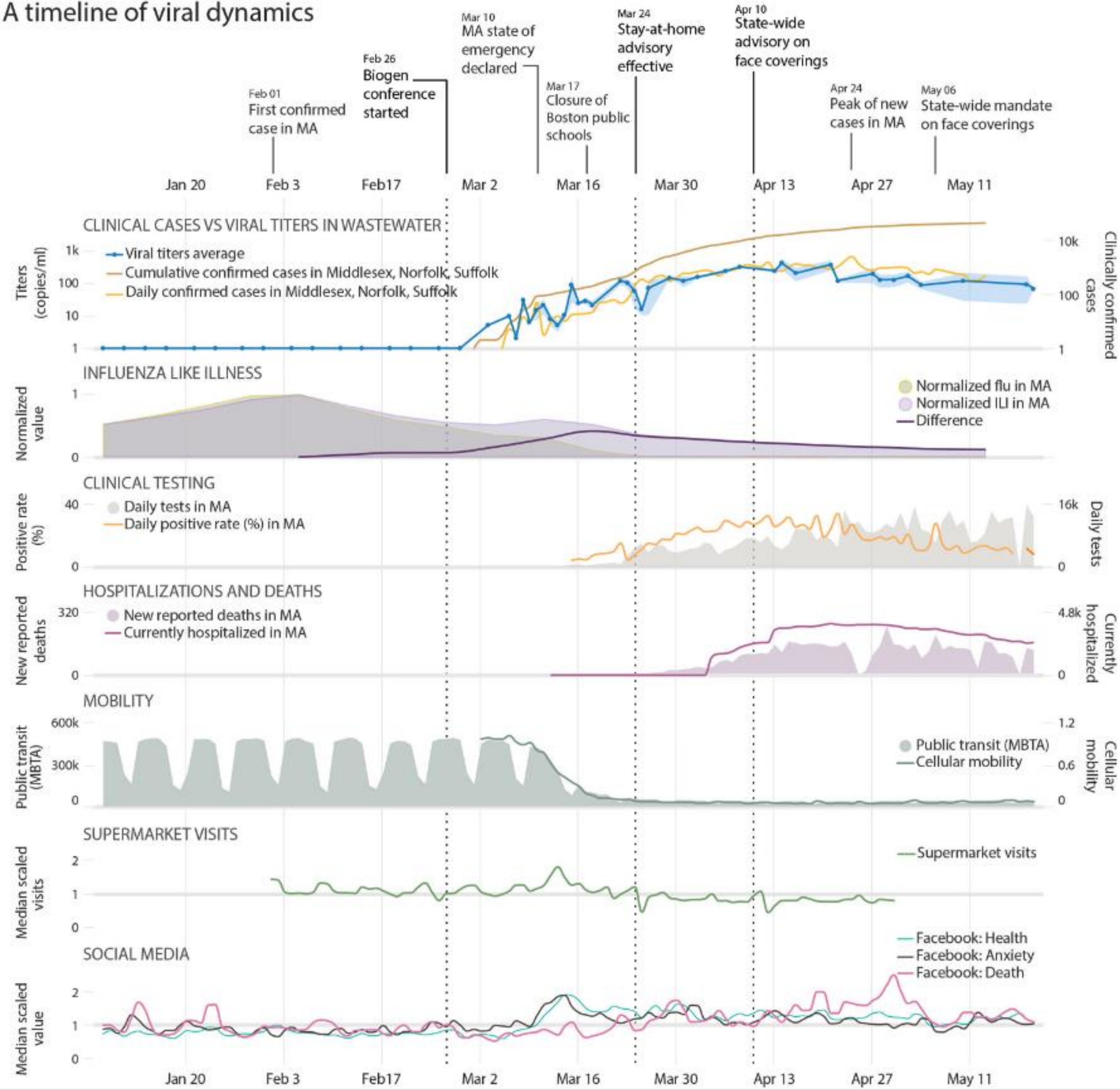
$$V(t) \propto I(t)$$

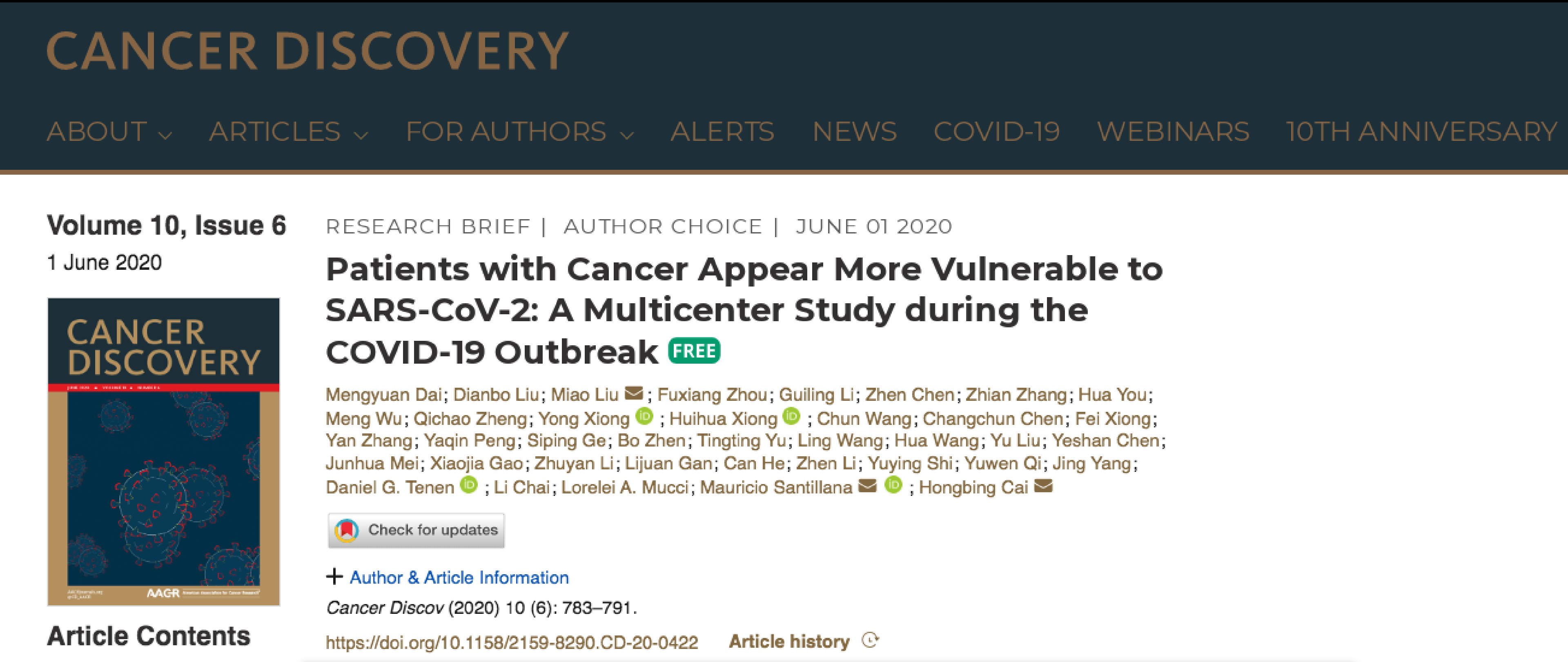
$$S(t) = Beta(t,\alpha,\beta)$$
$$W(t) = S(t) * I(t)$$





A timeline of viral dynamics





Volume 10, Issue 6
1 June 2020

Article Contents

RESEARCH BRIEF | AUTHOR CHOICE | JUNE 01 2020

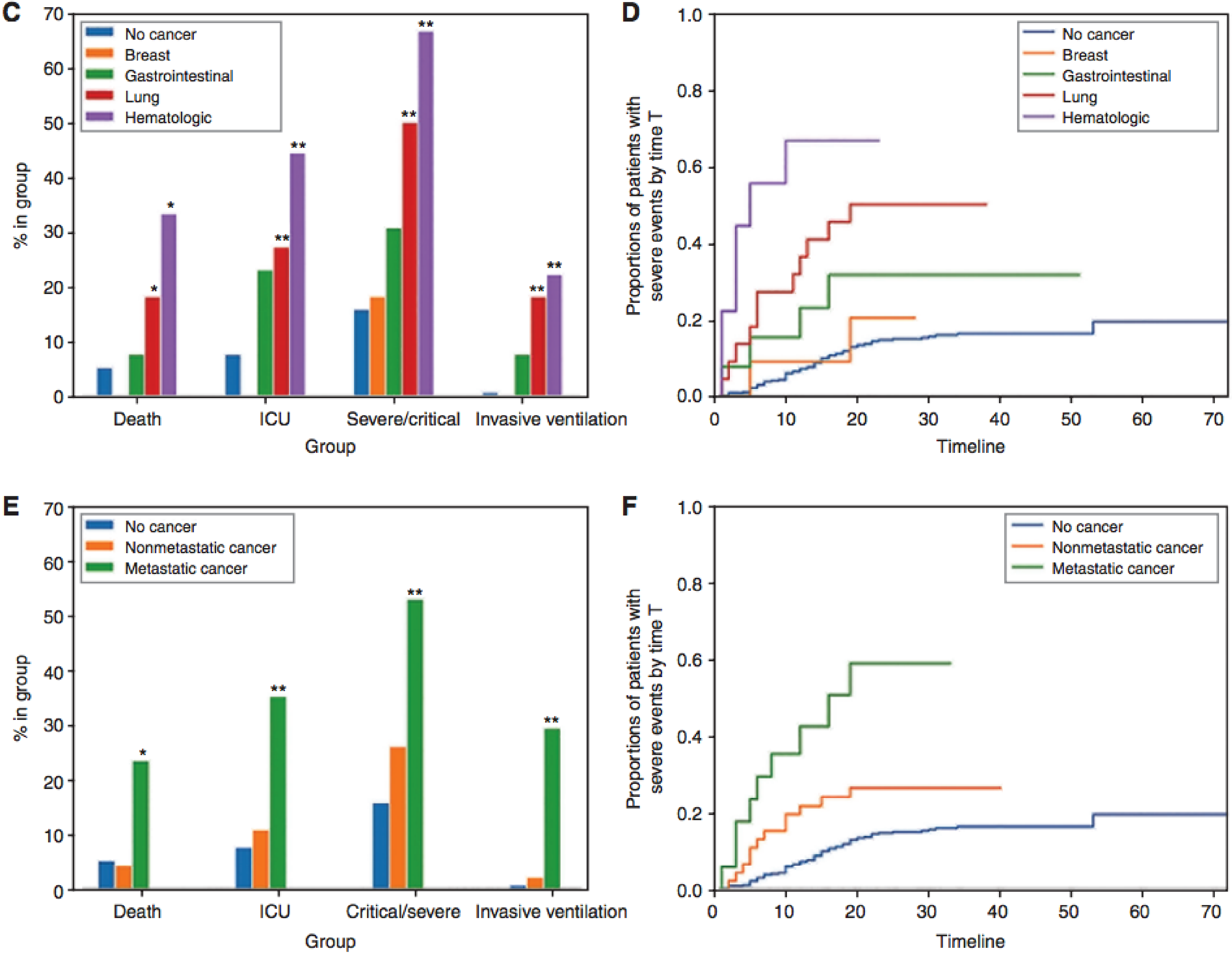
Patients with Cancer Appear More Vulnerable to SARS-CoV-2: A Multicenter Study during the COVID-19 Outbreak FREE

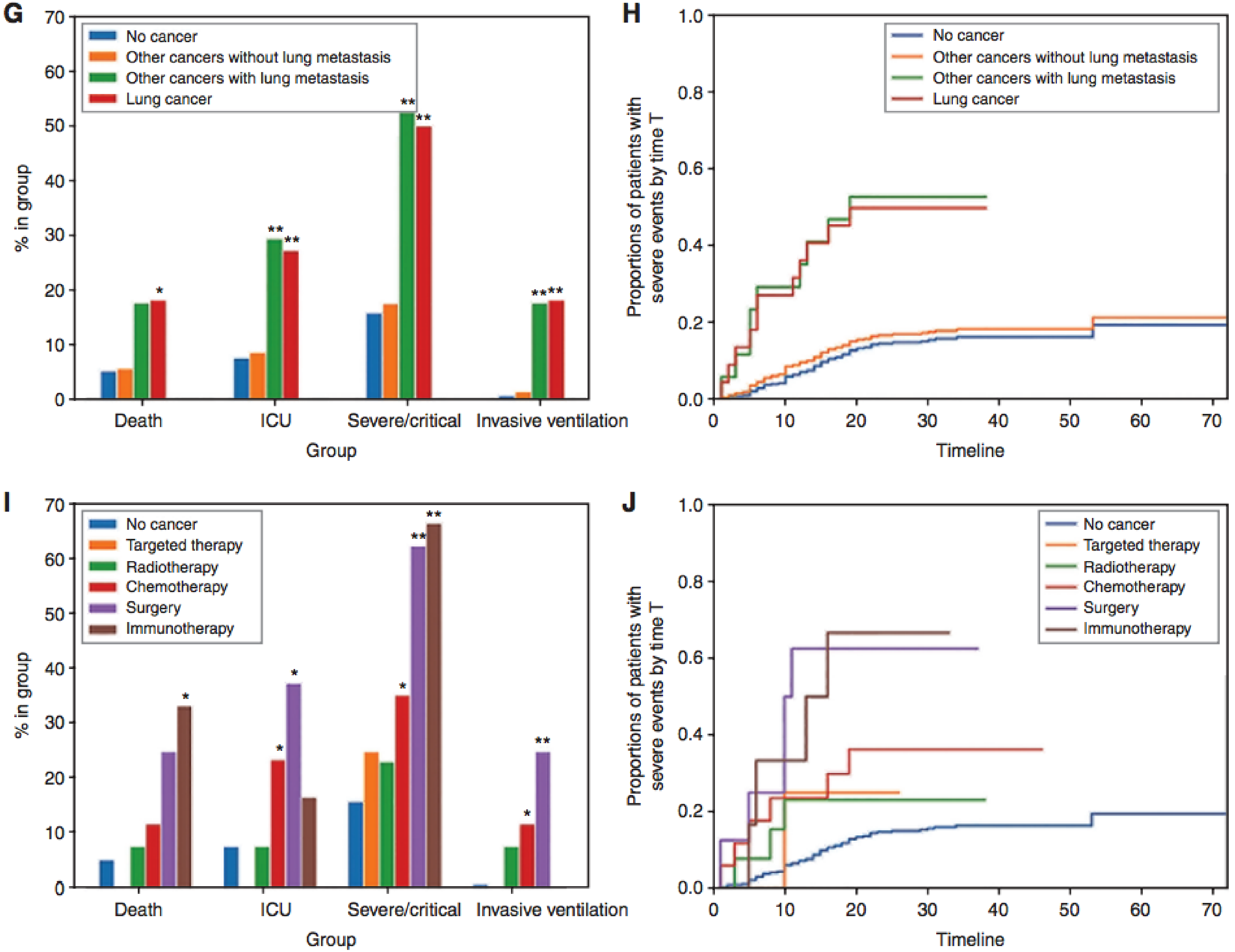
Mengyuan Dai; Dianbo Liu; Miao Liu; Fuxiang Zhou; Guiling Li; Zhen Chen; Zhian Zhang; Hua You; Meng Wu; Qichao Zheng; Yong Xiong; Huihua Xiong; Chun Wang; Changchun Chen; Fei Xiong; Yan Zhang; Yaqin Peng; Siping Ge; Bo Zhen; Tingting Yu; Ling Wang; Hua Wang; Yu Liu; Yeshan Chen; Junhua Mei; Xiaojia Gao; 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

+ Author & Article Information

Cancer Discov (2020) 10 (6): 783–791.

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


C
70
60
50
40
30
20
10
0
% in group
No cancer
Breast
Gastrointestinal
Lung
Hematologic
*
**
Death
ICU
Severe/critical
Invasive ventilation
Group
D
1.0
0.8
0.6
0.4
0.2
0.0
Proportions of patients with severe events by time T
No cancer
Breast
Gastrointestinal
Lung
Hematologic
0
10
20
30
40
50
60
70
Timeline
E
% in group
No cancer
Nonmetastatic cancer
Metastatic cancer
*
**
Death
ICU
Critical/severe
Invasive ventilation
Group
F
Proportions of patients with severe events by time T
No cancer
Nonmetastatic cancer
Metastatic cancer
Timeline


G
% in group
0
10
20
30
40
50
60
70
No cancer
Other cancers without lung metastasis
Other cancers with lung metastasis
Lung cancer
*
**
**
**
**
**
**
Death
ICU
Severe/critical
Invasive ventilation
Group
H
Proportions of patients with severe events by time T
0.0
0.2
0.4
0.6
0.8
1.0
No cancer
Other cancers without lung metastasis
Other cancers with lung metastasis
Lung cancer
0
10
20
30
40
50
60
70
Timeline
I
% in group
0
10
20
30
40
50
60
70
No cancer
Targeted therapy
Radiotherapy
Chemotherapy
Surgery
Immunotherapy
*
*
*
*
**
**
*
**
Death
ICU
Severe/critical
Invasive ventilation
Group
J
Proportions of patients with severe events by time T
0.0
0.2
0.4
0.6
0.8
1.0
No cancer
Targeted therapy
Radiotherapy
Chemotherapy
Surgery
Immunotherapy
0
10
20
30
40
50
60
70
Timeline

The Evolving Roles of US Political Partisanship and Social Vulnerability in the COVID-19 Pandemic from February 2020 - February 2021

35 Pages • Posted: 8 Oct 2021

Justin Kaashoek
Boston Children's Hospital; Harvard Paulson School of Engineering and Applied Sciences

Christian Testa
Harvard T.H. Chan School of Public Health - Department of Social and Behavioral Sciences

Jarvis Chen
Harvard University - Department of Social and Behavioral Sciences

Lucas Stolerman
Boston Children's Hospital - Computational Health Informatics Program

Nancy Krieger
Harvard University - Department of Social and Behavioral Sciences

William P. Hanage
Harvard T.H. Chan School of Public Health - Department of Epidemiology

Mauricio Santillana
Boston Children's Hospital; Harvard University - Harvard Medical School; Harvard University - T.H. Chan School of Public Health

Date Written: September 30, 2021

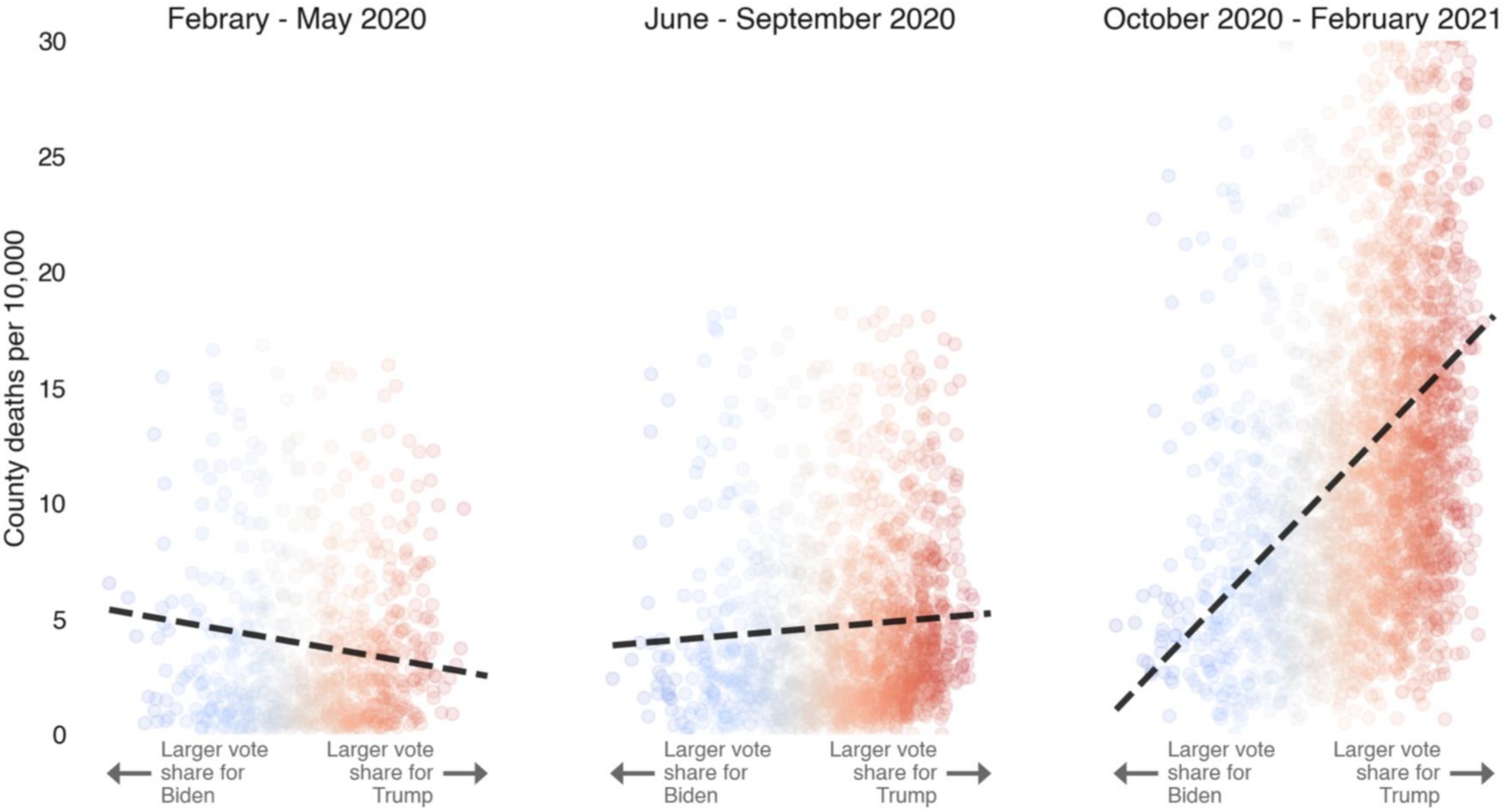
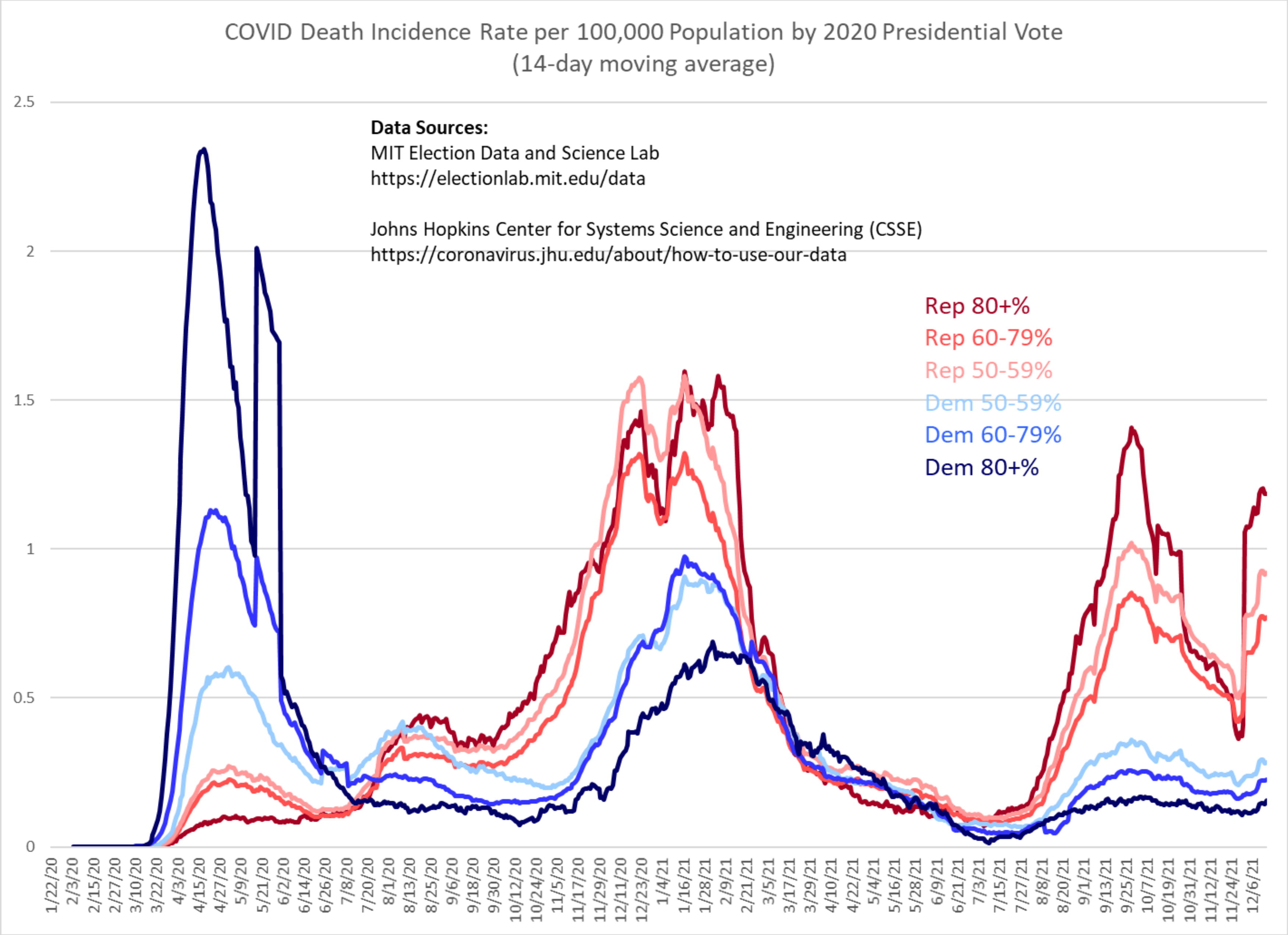


Figure 2: Evolution of COVID-19 deaths vs Political leaning. COVID-19 attributed deaths (per 10,000) at the county level as a function of vote share in favor of J. Biden (Democratic) vs D.J. Trump (Republican), the 2020 presidential candidates, during the three time periods of interest. Inspiration for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



COVID Death Incidence Rate per 100,000 Population by 2020 Presidential Vote
(14-day moving average)

Data Sources:
MIT Election Data and Science Lab
https://electionlab.mit.edu/data

Johns Hopkins Center for Systems Science and Engineering (CSSE)
https://coronavirus.jhu.edu/about/how-to-use-our-data

Not from our team with updates. Credit @maolesen

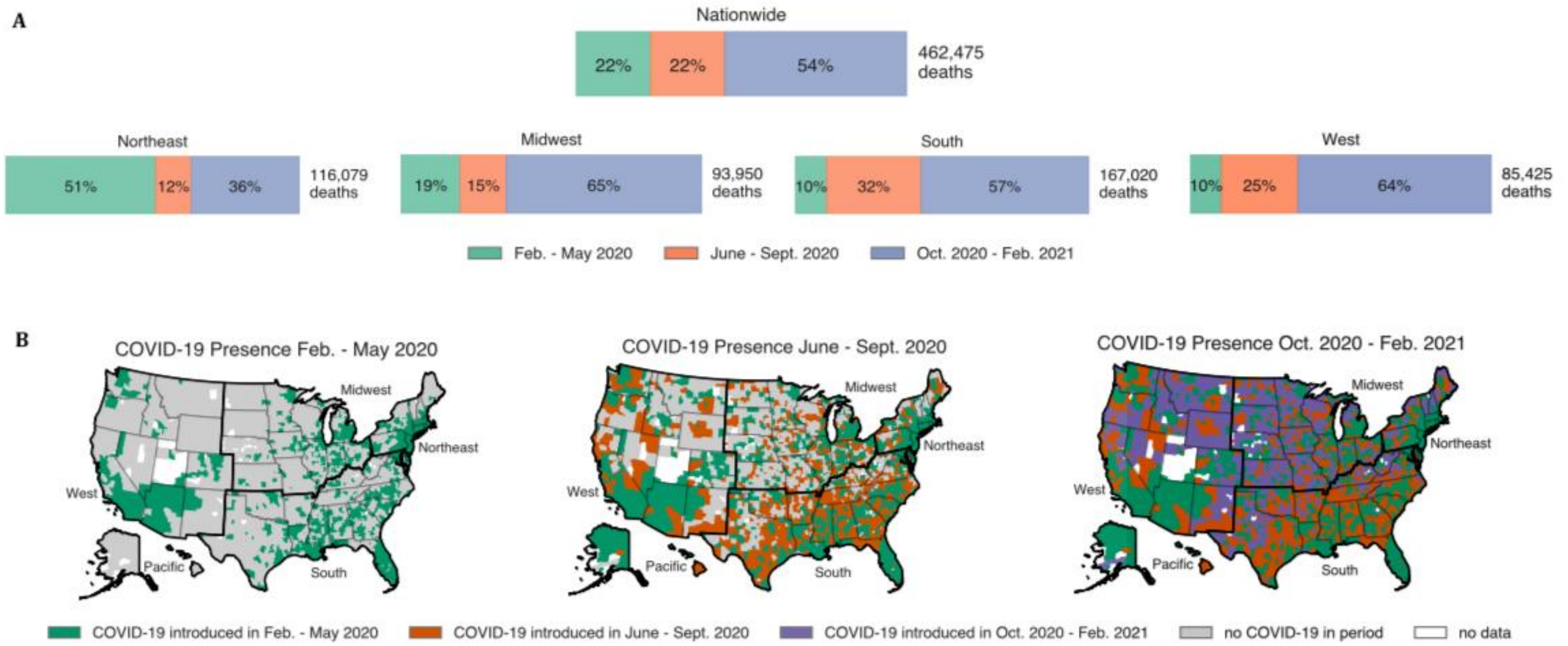
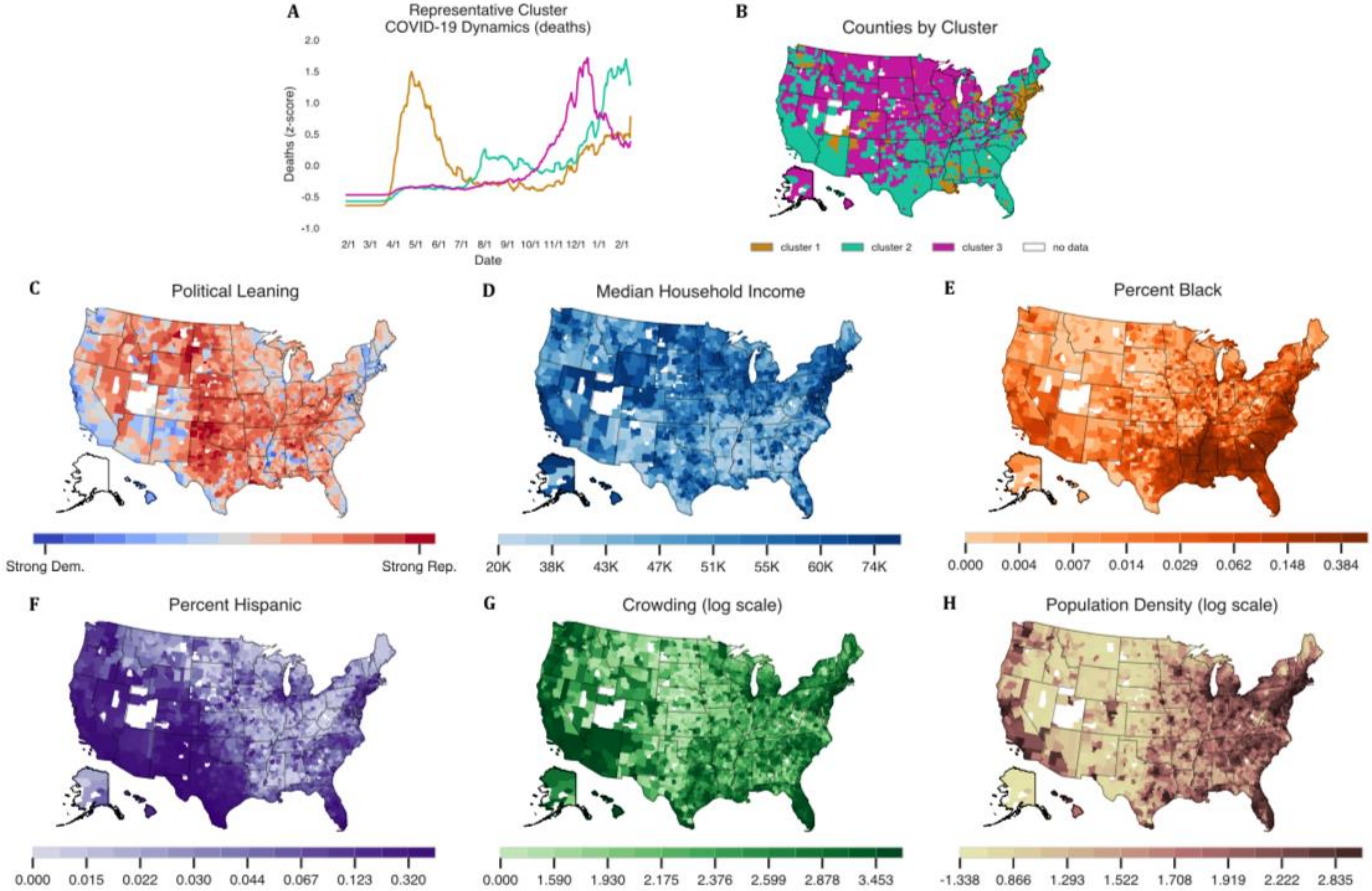
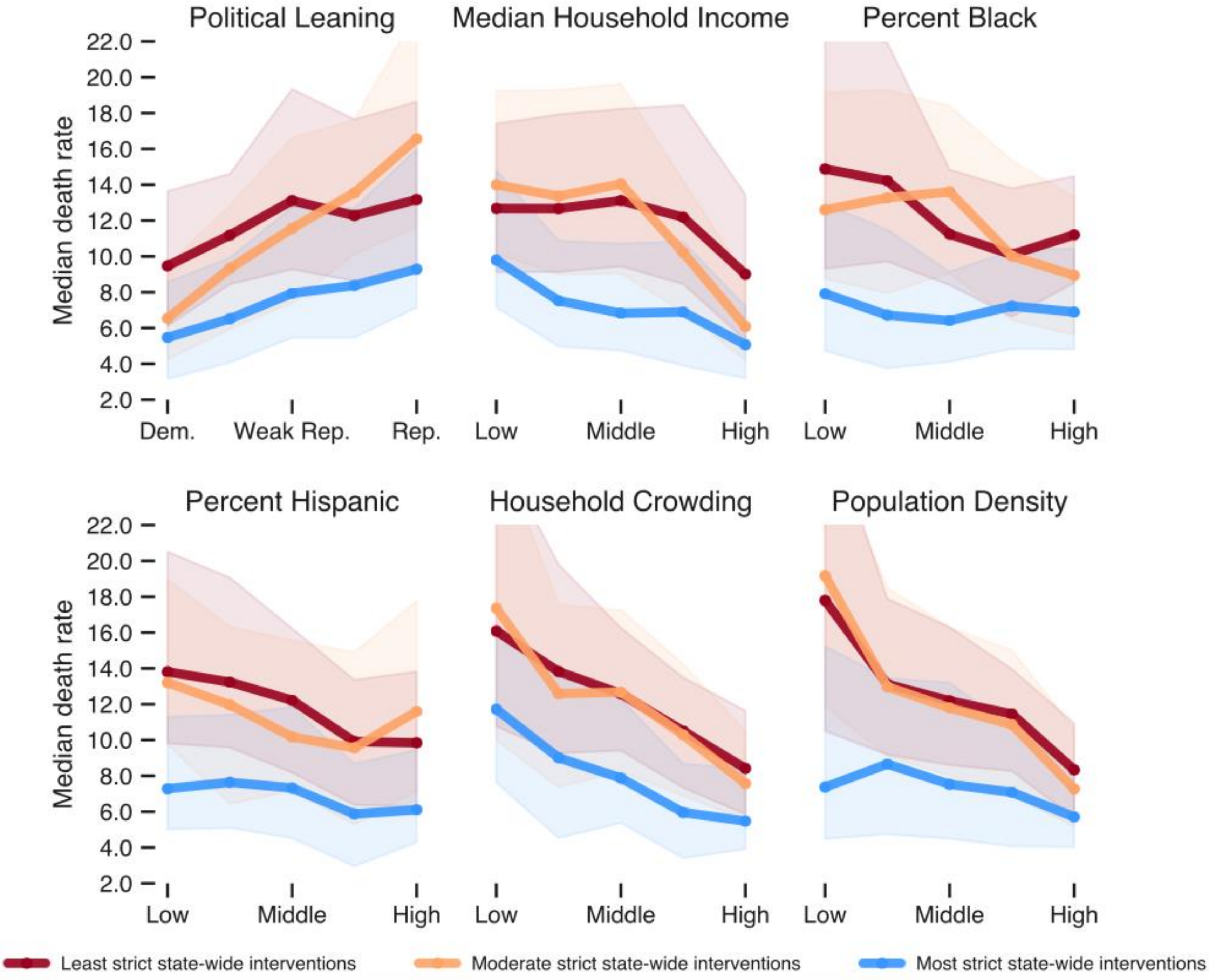
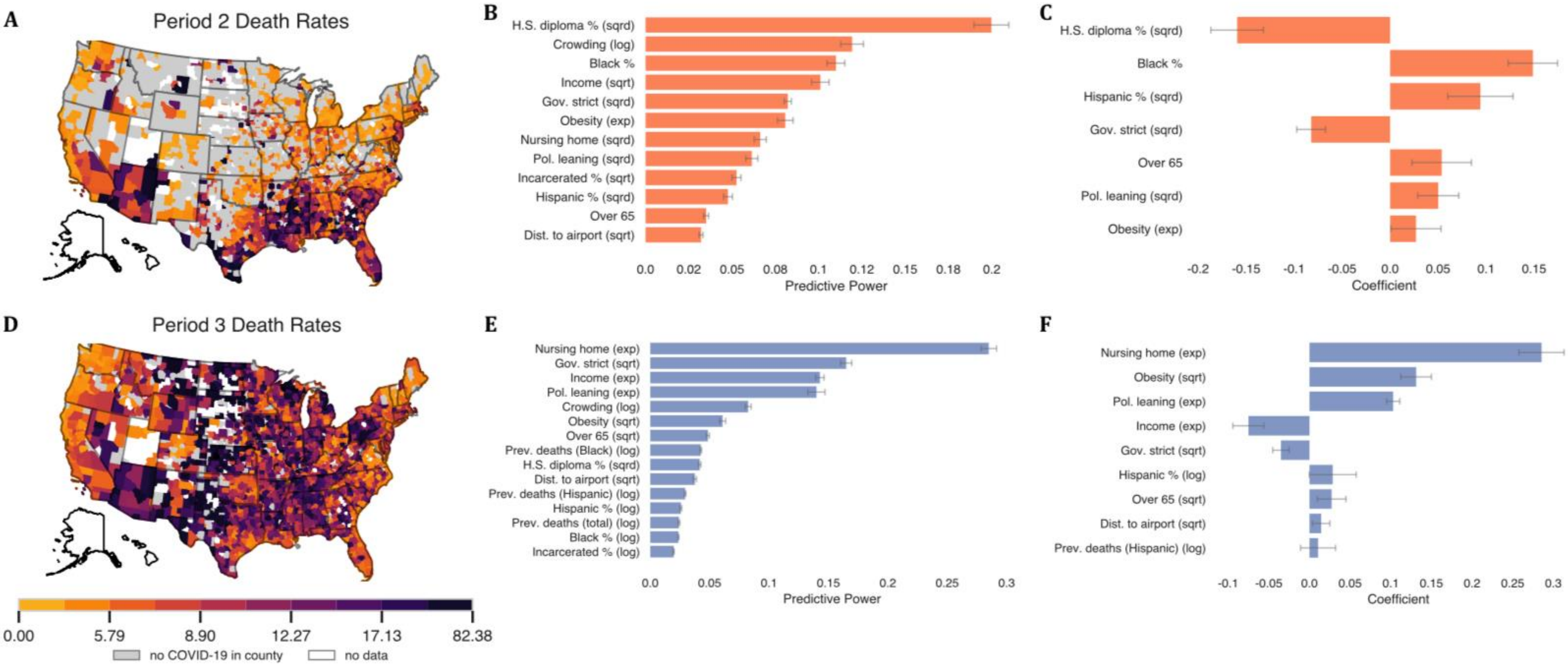
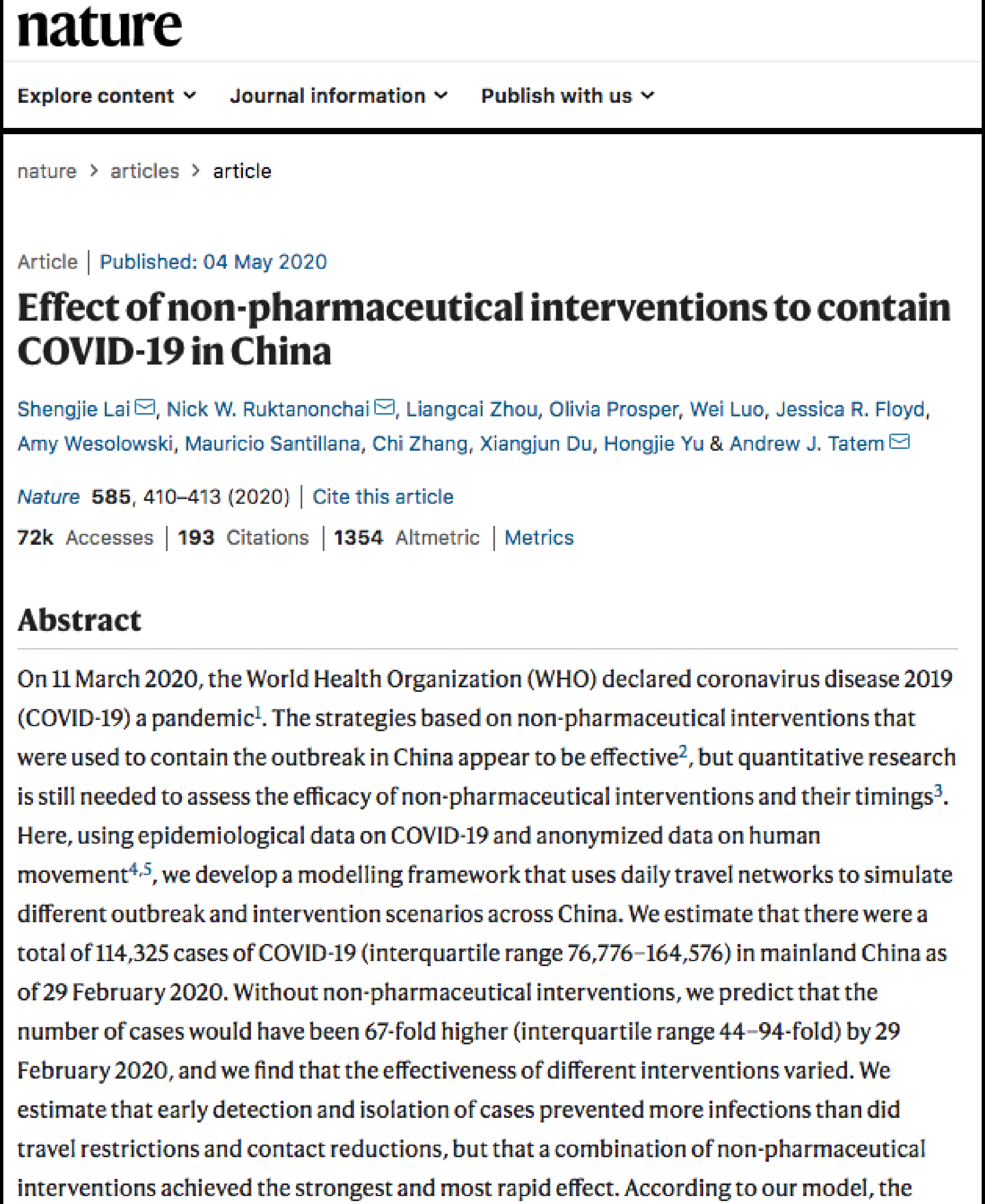


Figure 3: A breakdown of COVID-19 presence across the time periods of interest. (**A**) The percentage of deaths in the nation by time period, both nationwide and by Census region. Other than the Northeast, which was hit hard in the first period, the nation was hit hardest in period 3, as pointed out 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
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
Low
Middle
High
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
Low
Middle
High
Low
Middle
High
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
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
0.00
5.79
8.90
12.27
17.13
82.38
no COVID-19 in county
no data



Article | Published: 04 May 2020

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

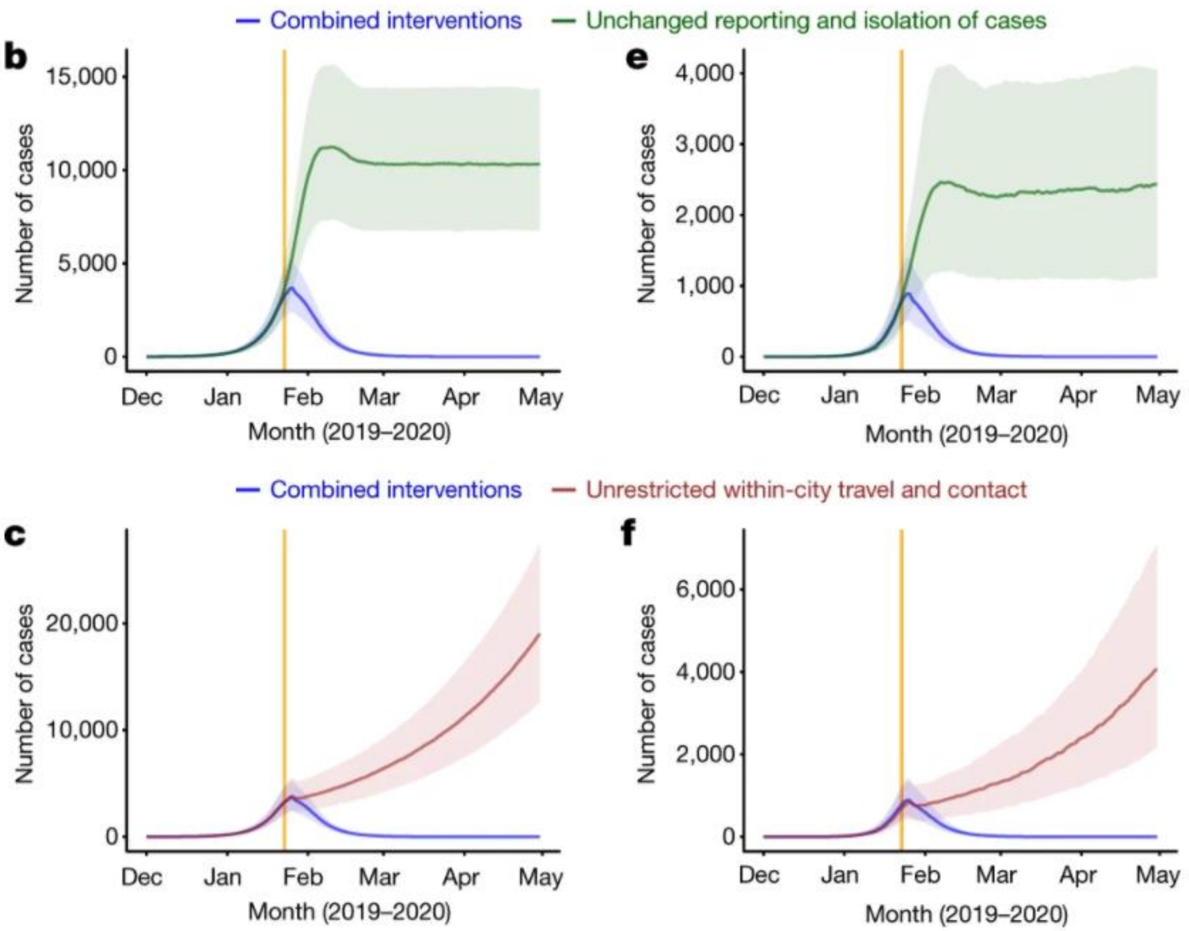
Shengjie Lai, Nick W. Ruktanonchai, Liangcai Zhou, Olivia Prosper, Wei Luo, Jessica R. Floyd, Amy Wesolowski, Mauricio Santillana, Chi Zhang, Xiangjun Du, Hongjie Yu & Andrew J. Tatem

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

72k Accesses | **193** Citations | **1354** Altmetric | Metrics

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).

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

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

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

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

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

248 Accesses | **2** Altmetric | Metrics

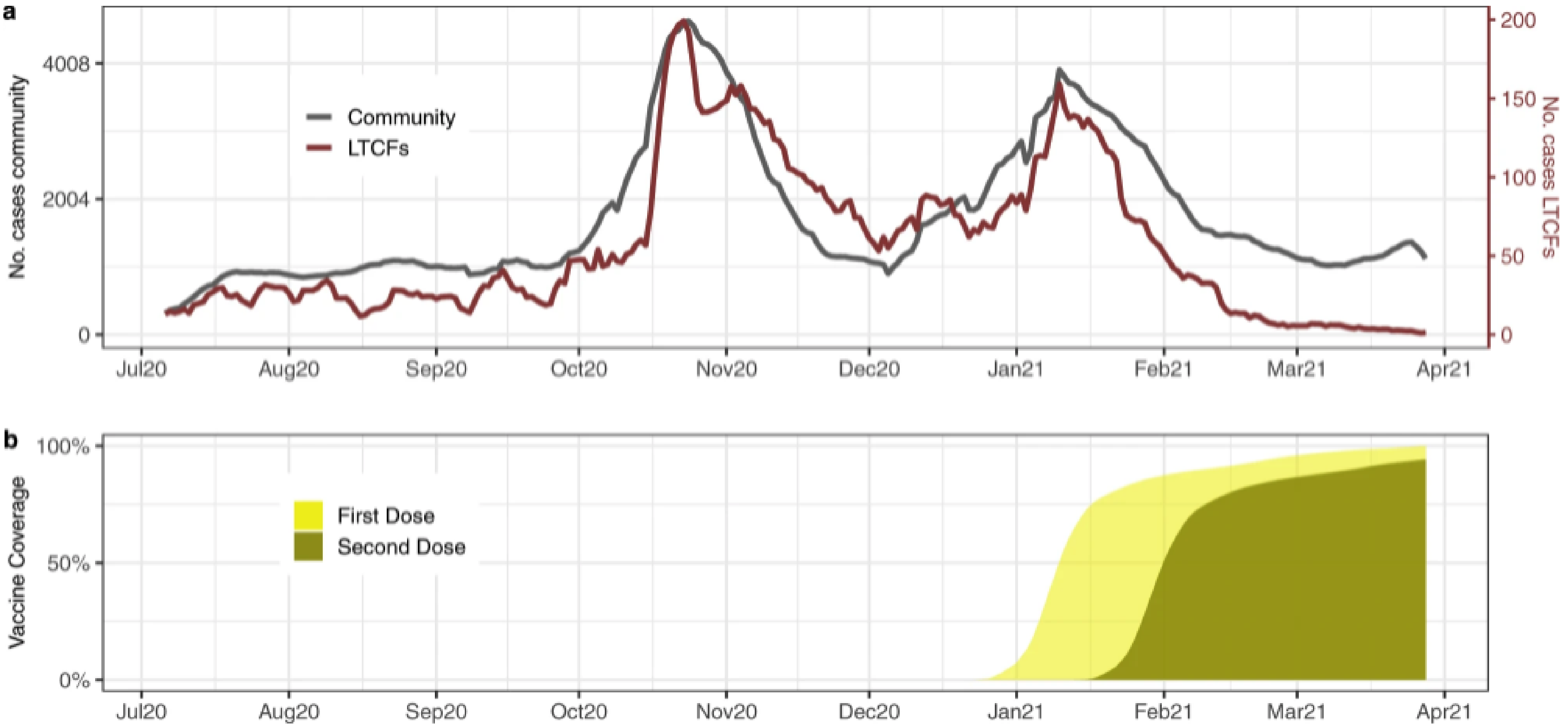
Abstract

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.

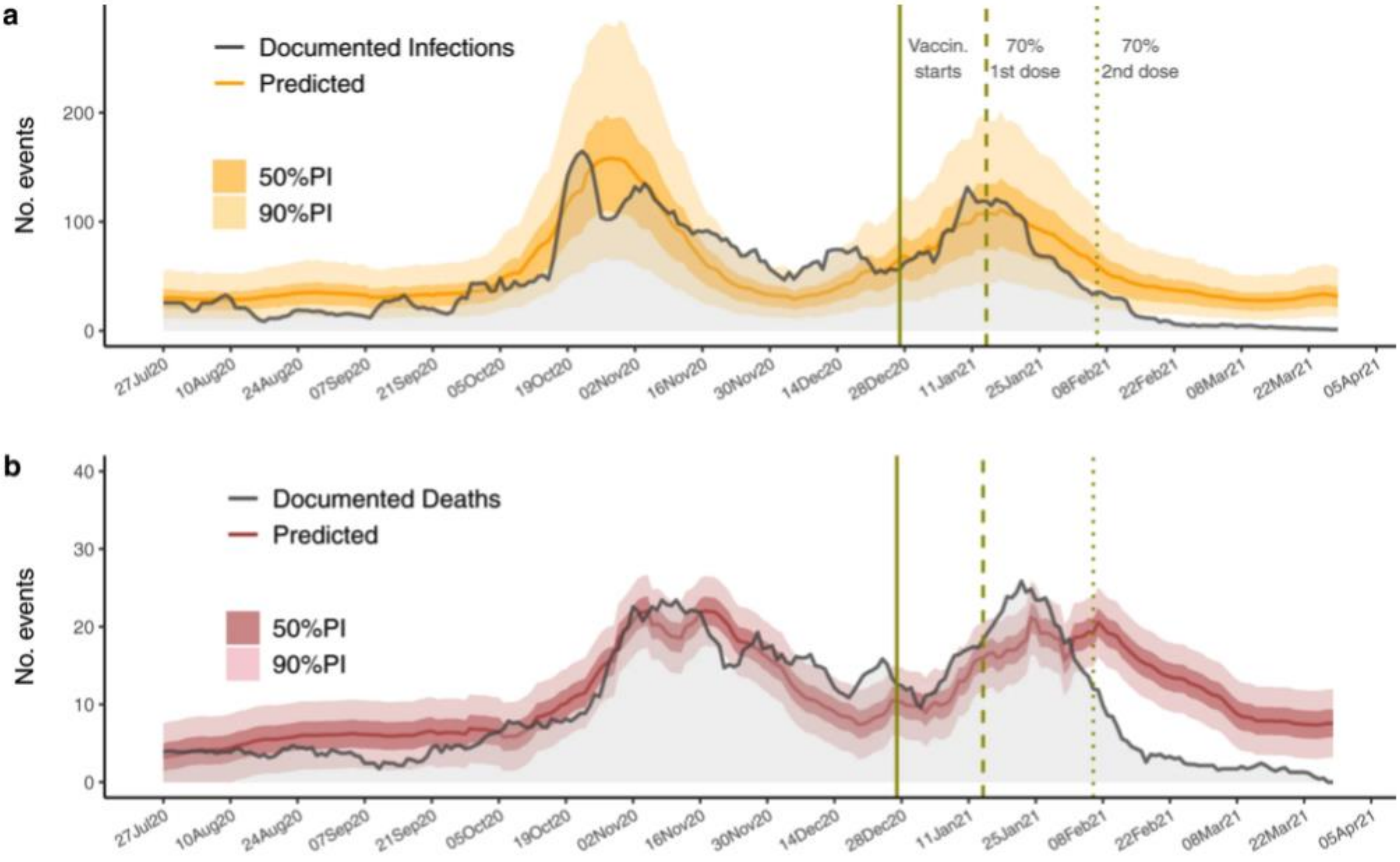
From: High coverage COVID-19 mRNA vaccination rapidly controls SARS-CoV-2 transmission in long-term care facilities

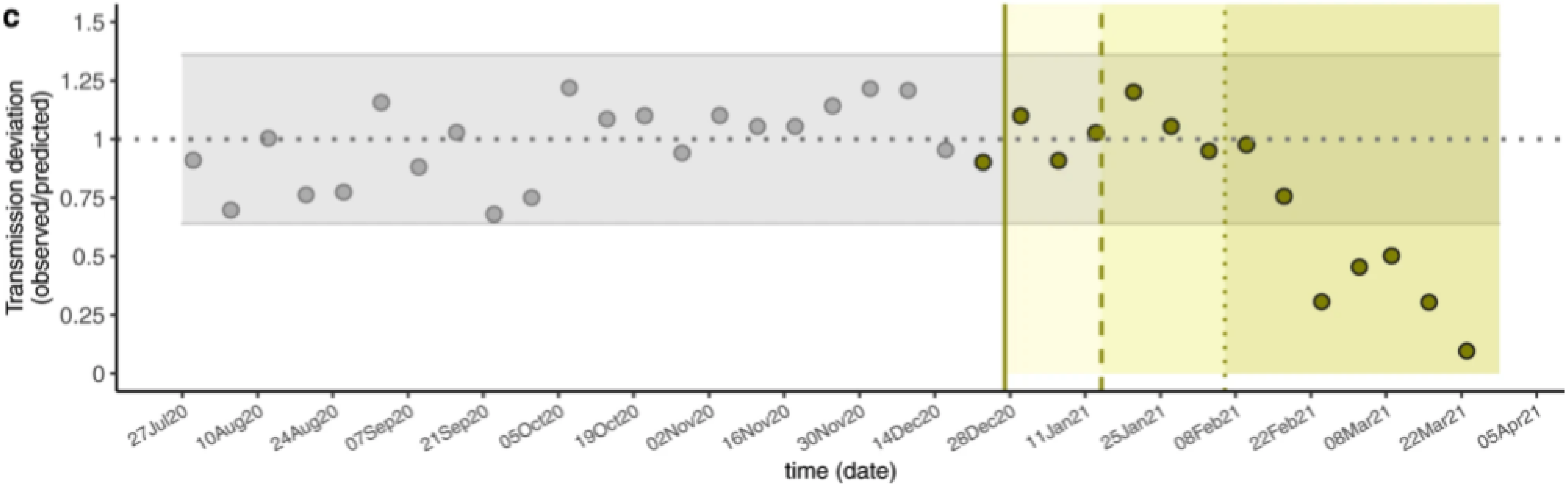


a Comparison of the total community (gray) and LTCFs' documented infections (red) trajectories in Catalonia, Spain. **b** First and second dose vaccine coverage among LTCFs' residents.

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.

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]

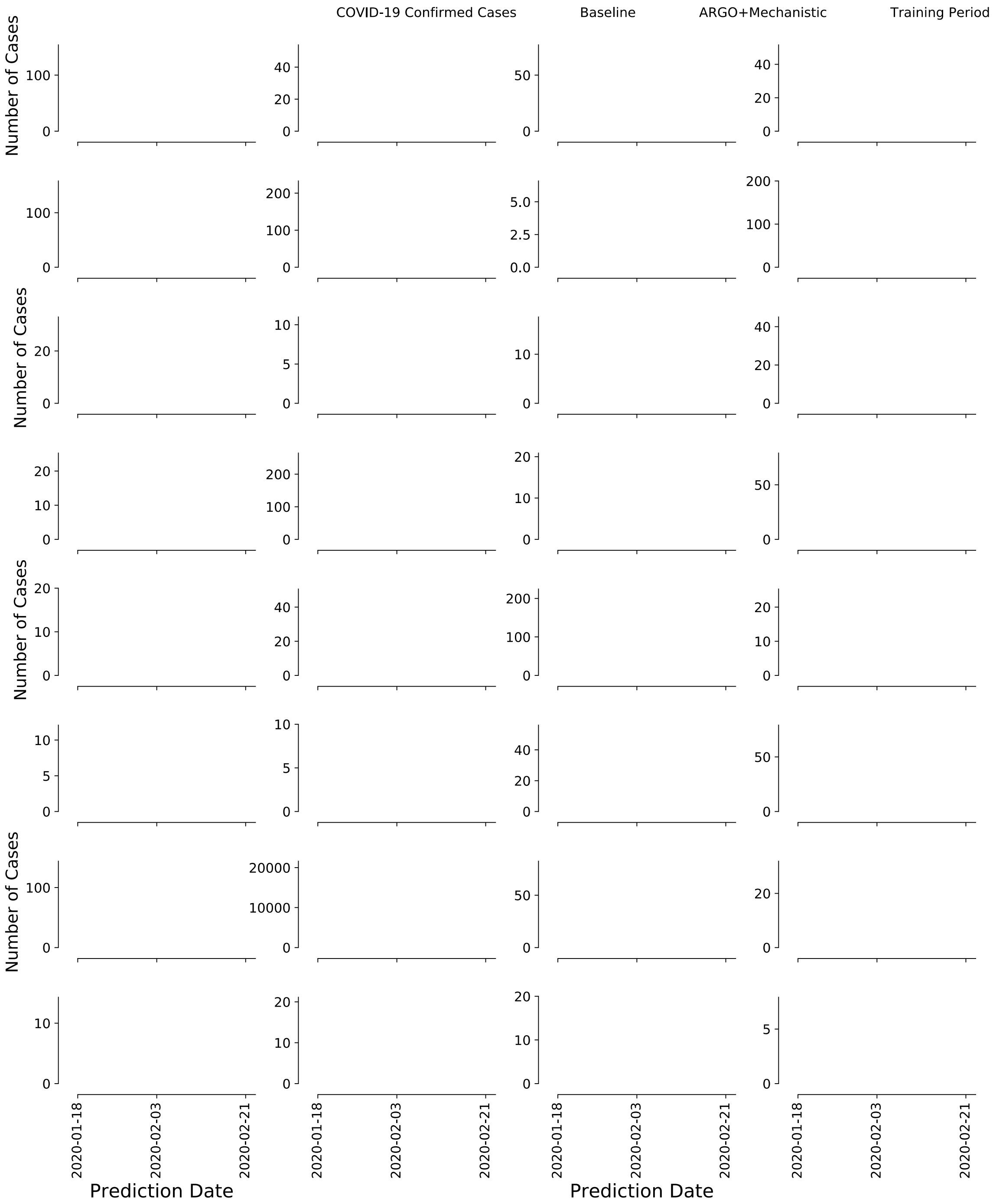
Related Article

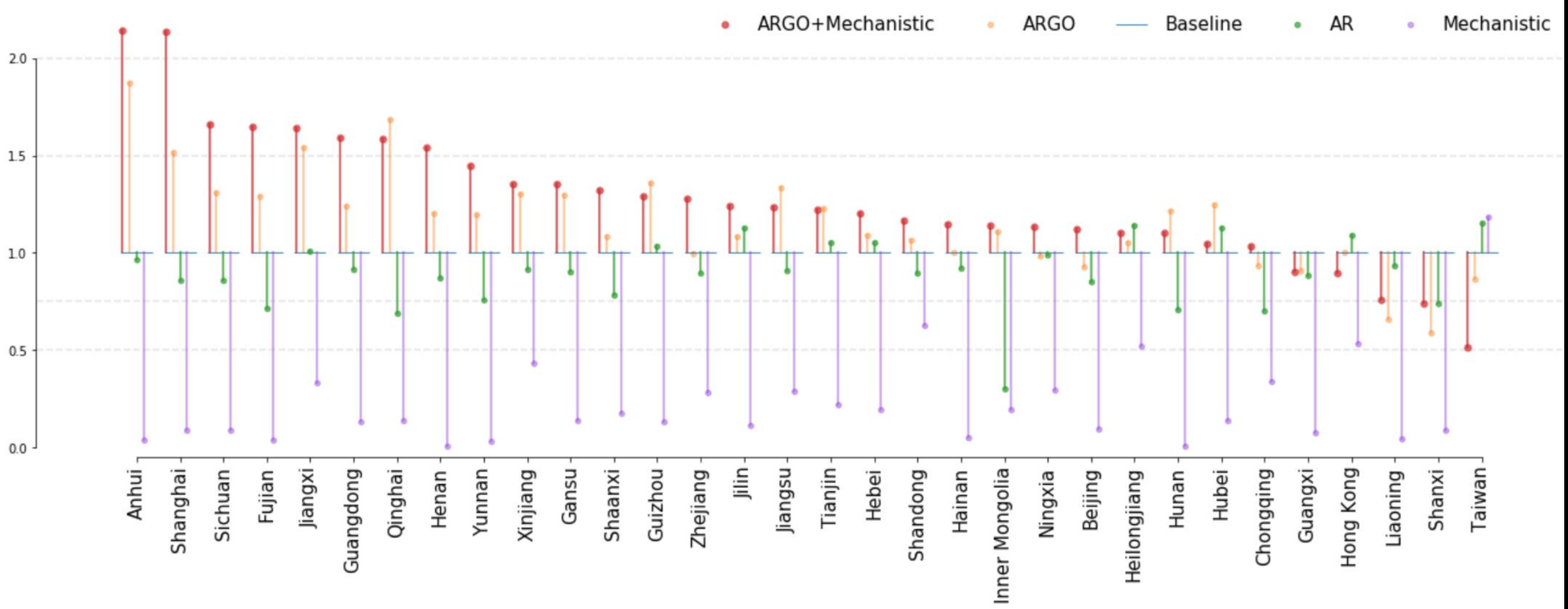
This is a corrected version. See correction statement in: https://www.jmir.org/2020/9/e23996/

Abstract

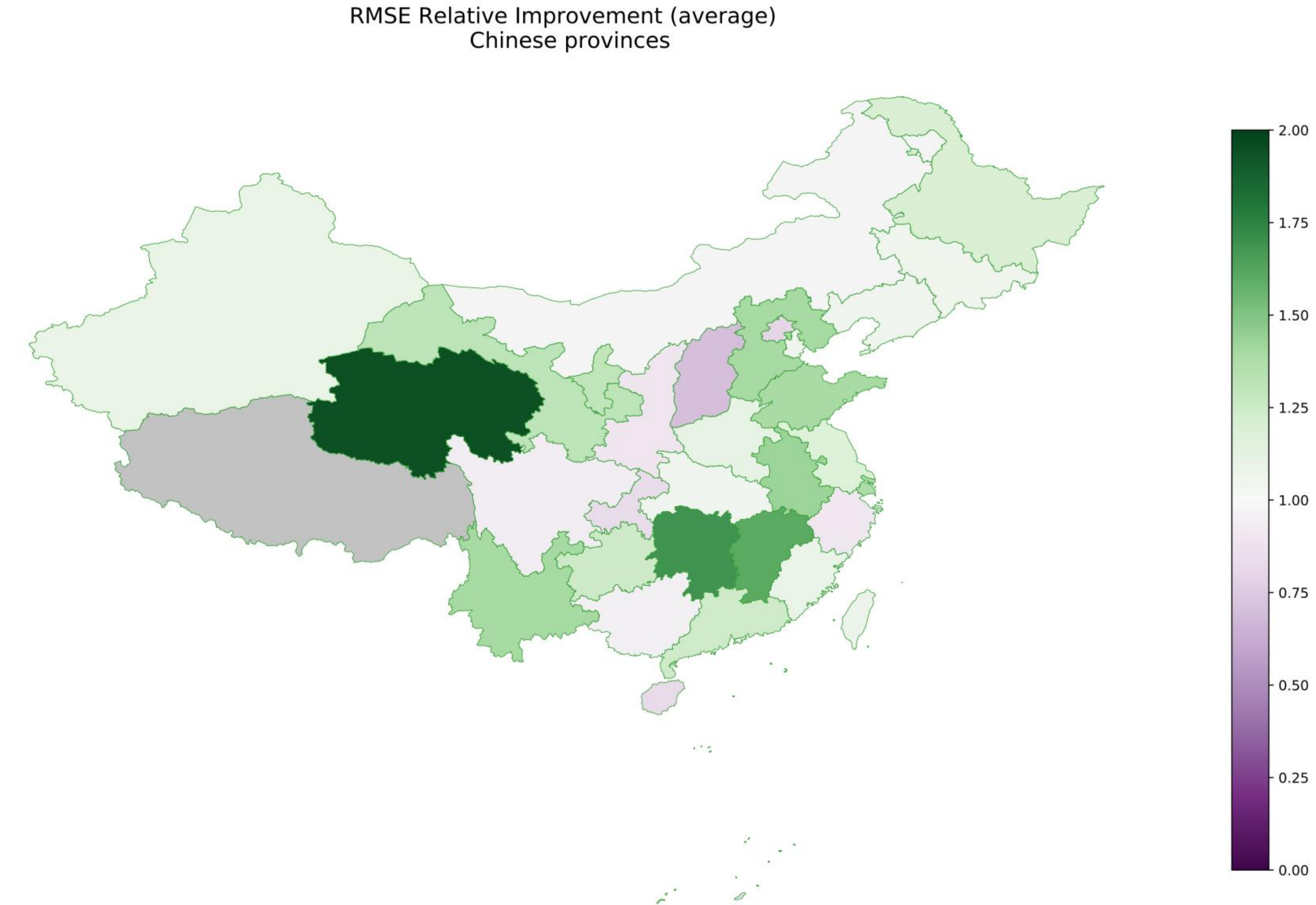
Background:
The inherent difficulty of identifying and monitoring emerging outbreaks caused by novel pathogens can lead to their rapid spread; and if left unchecked, they may become major public health threats to the planet. The ongoing coronavirus disease (COVID-19) outbreak, which has infected over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

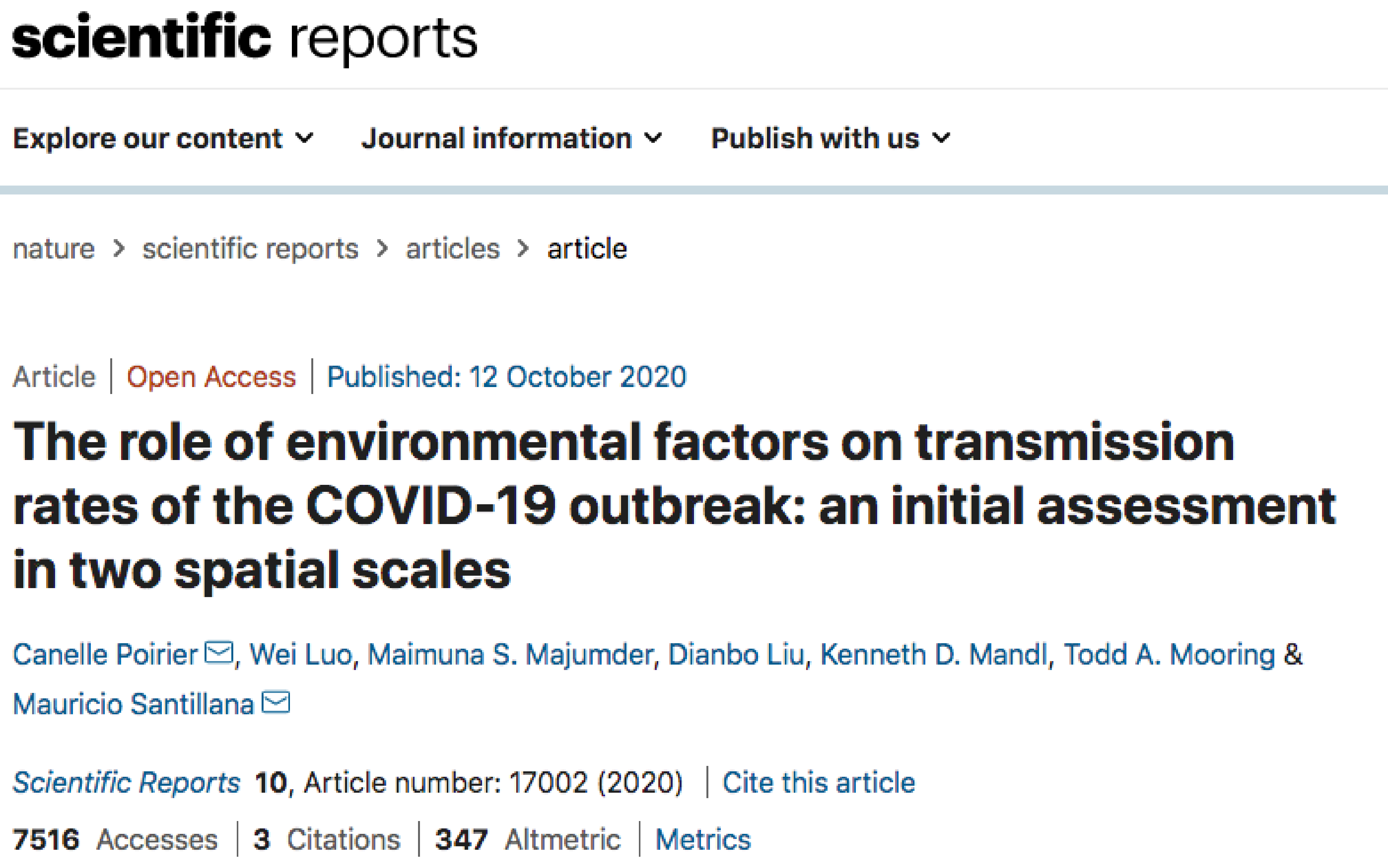
Objective:
We present a timely and novel methodology that combines disease estimates from mechanistic models and digital traces, via interpretable machine learning methodologies, to reliably forecast COVID-19 activity in Chinese provinces in real time.


COVID-19 Confirmed Cases
Baseline
ARGO+Mechanistic
Training Period
Number of Cases
Number of Cases
Number of Cases
Number of Cases
0
100
0
20
40
0
50
0
20
40
0
100
0
100
200
0.0
2.5
5.0
0
100
200
0
20
0
5
10
0
10
0
20
40
0
10
20
0
100
200
0
10
20
0
50
0
10
20
0
20
40
0
100
200
0
10
20
0
5
10
0
5
10
0
20
40
0
50
0
100
0
10000
20000
0
50
0
20
0
10
0
10
20
0
10
20
0
5
2020-01-18
2020-02-03
2020-02-21
Prediction Date
Prediction Date



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 | Open Access | Published: 12 October 2020

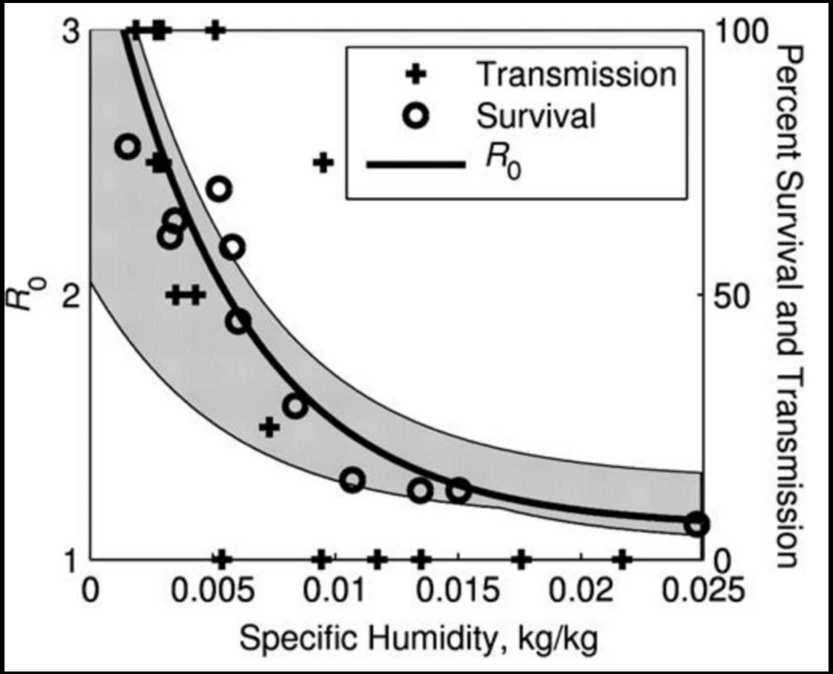
The role of environmental factors on transmission rates of the COVID-19 outbreak: an initial assessment in two spatial scales

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

7516 Accesses | **3** Citations | **347** Altmetric | Metrics

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."*

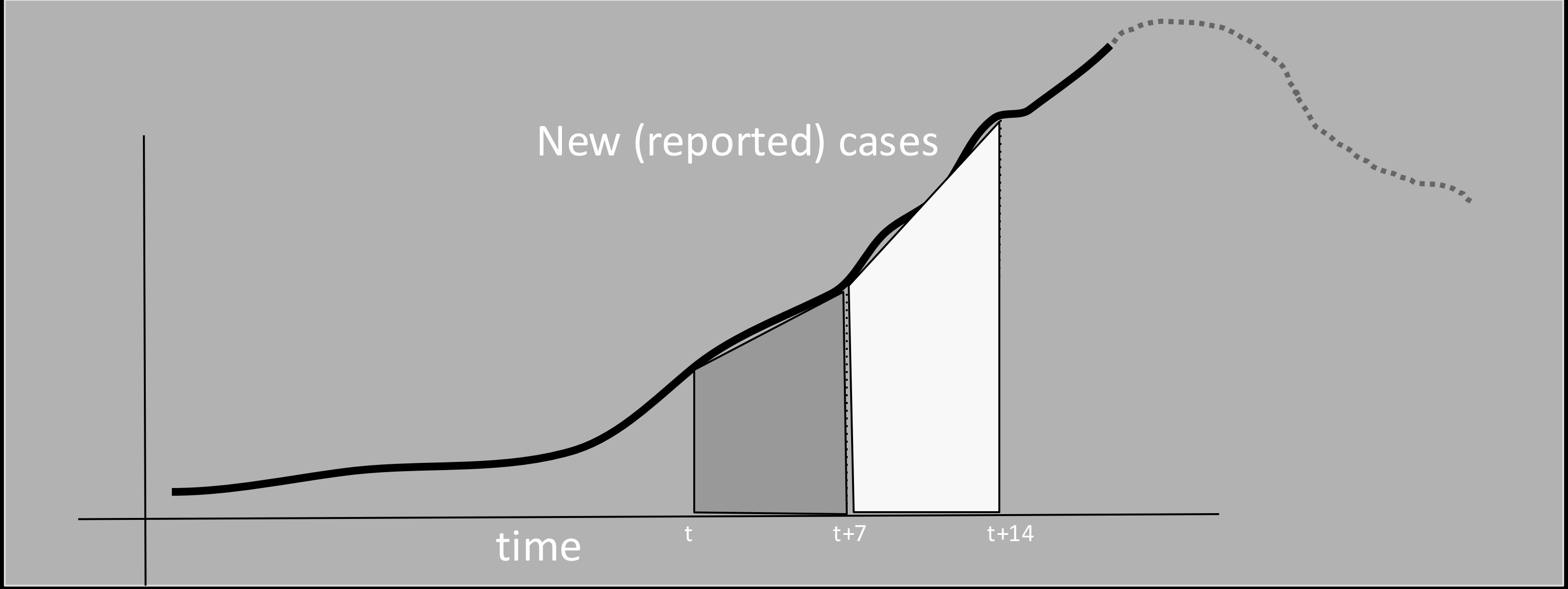
These observed relationships in influenza transmission have been assumed (without any evidence) for the ongoing COVID-19 outbreak. It has been stated that COVID-19 transmission will decrease as warmer temperatures (leading to higher absolute humidity conditions) are experienced in the upcoming spring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

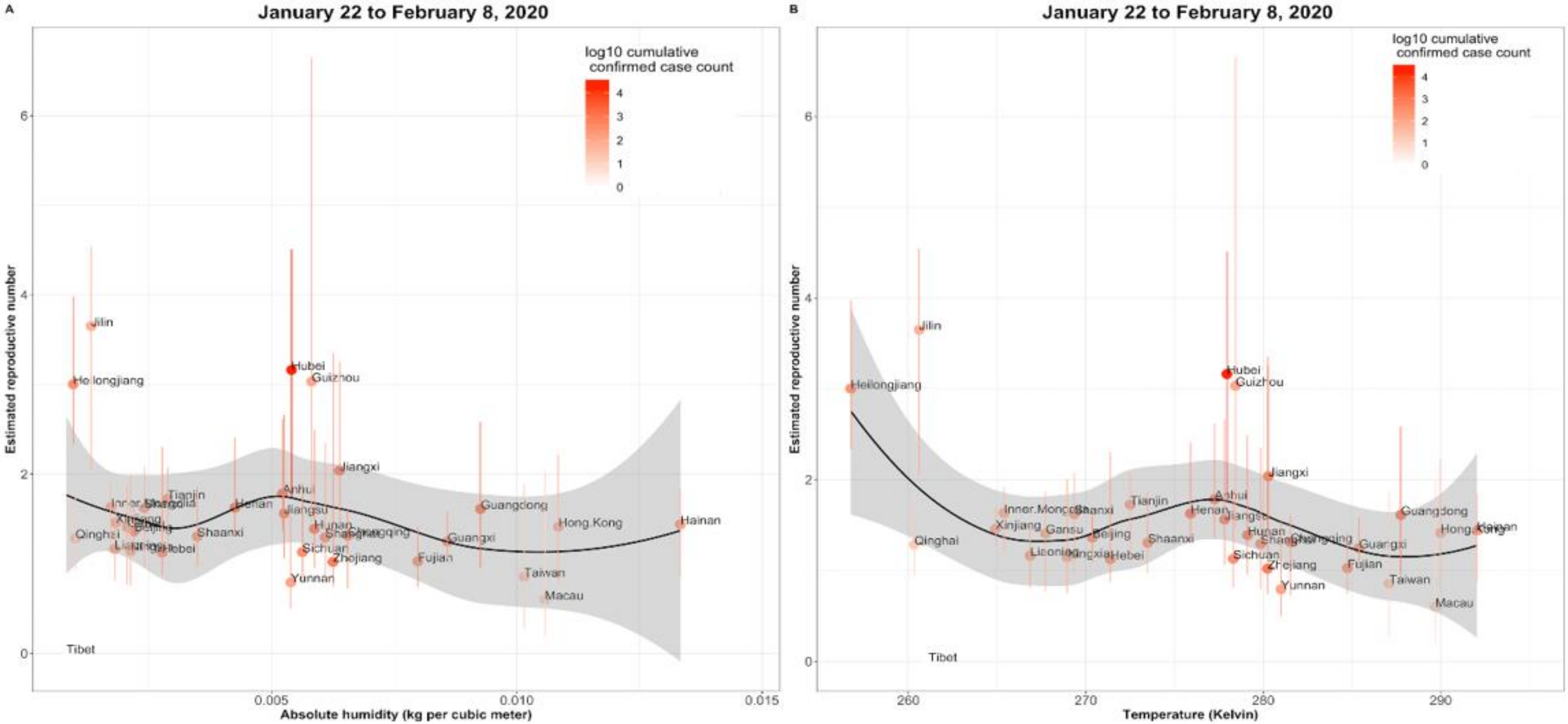
A novel coronavirus (SARS-CoV-2) was identified in Wuhan, Hubei Province, China, in December 2019 and has caused over 240,000 cases of COVID-19 worldwide as of March 19, 2020. Previous studies have supported an epidemiological hypothesis that cold and dry environments facilitate the survival and spread 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.

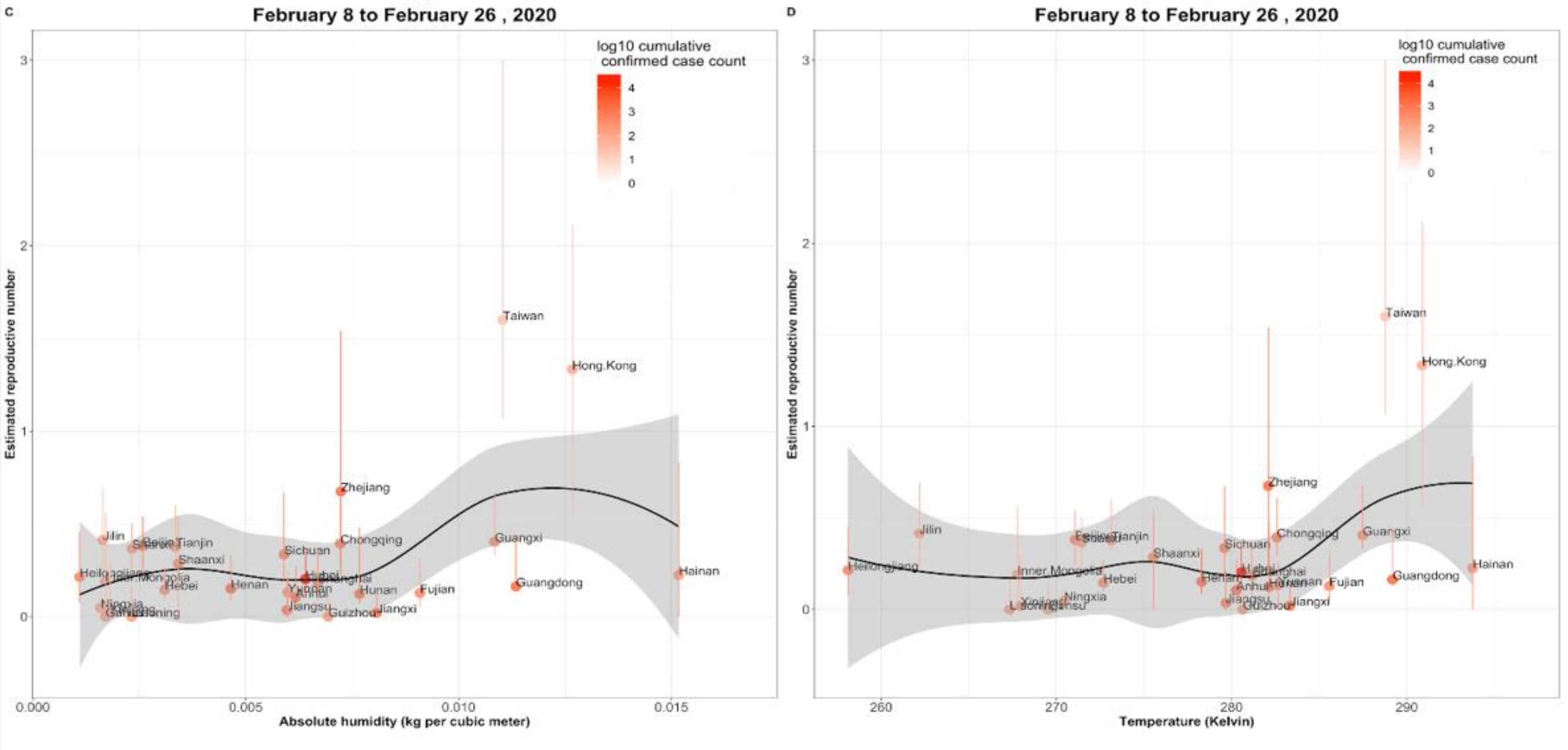
Calculating a proxy for R0



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

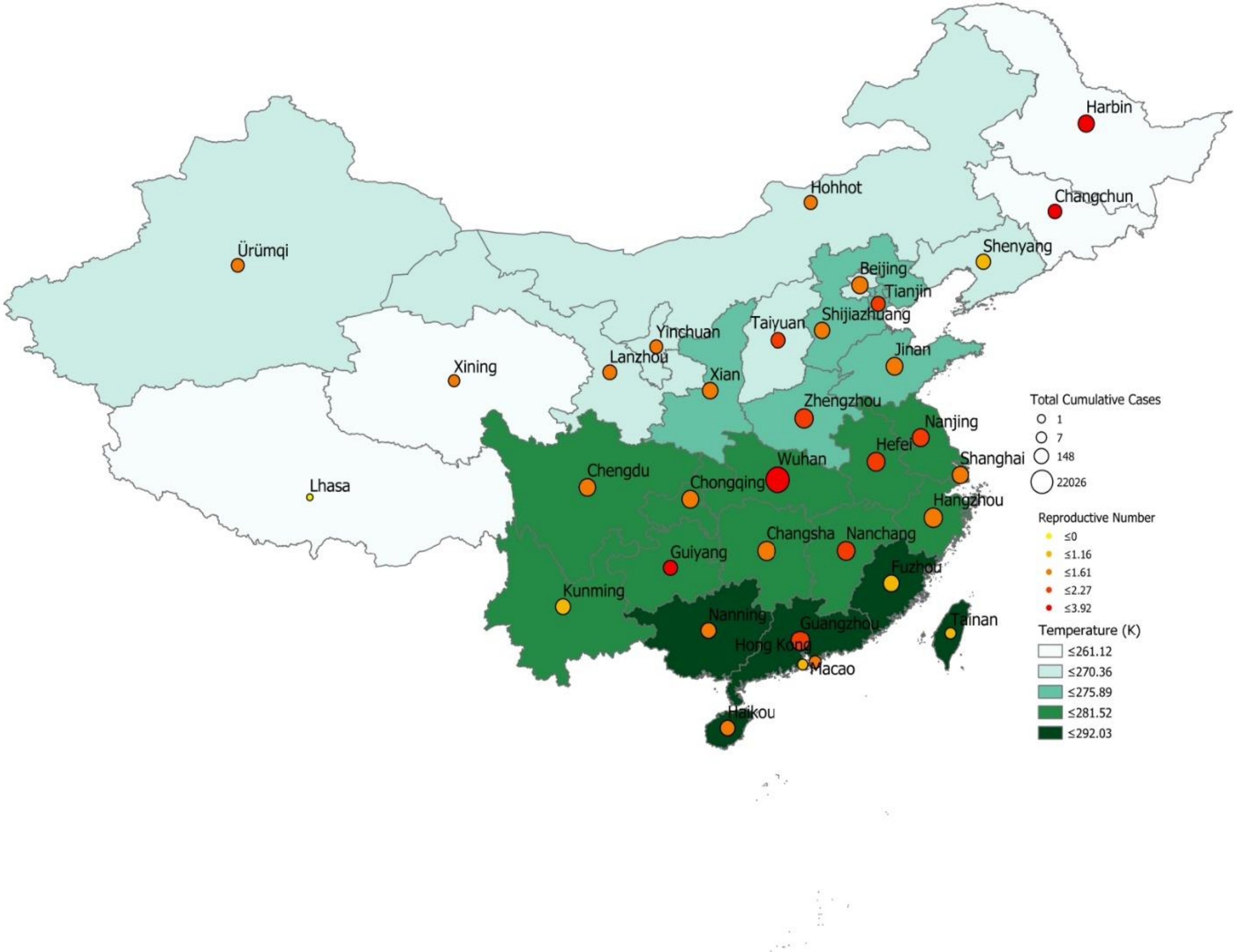
C(t) is the total cumulative cases at time t, and d = 5,6,7 is an estimate of the serial interval


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
Inner Mongolia
Shanxi
Henan
Anhui
Jiangsu
Xinjiang
Gansu
Beijing
Qinghai
Shaanxi
Liaoning
Ningxia
Hebei
Hunan
Shanghai
Chongqing
Sichuan
Zhejiang
Yunnan
Guangdong
Hong Kong
Hainan
Guangxi
Fujian
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
Inner Mongolia
Shanxi
Tianjin
Anhui
Henan
Jiangsu
Xinjiang
Gansu
Beijing
Shaanxi
Qinghai
Liaoning
Ningxia
Hebei
Hunan
Shanghai
Chongqing
Guangxi
Sichuan
Zhejiang
Yunnan
Guangdong
Hong Kong
Hainan
Fujian
Taiwan
Macau
Tibet
260
270
280
290
Temperature (Kelvin)



February 8 to February 26 , 2020

February 8 to February 26 , 2020


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

Towards the Development of Decision-Support Tools for the Pediatric Intensive Care Unit

Mauricio Santillana, PhD (msantill@fas.harvard.edu)

Collaborators: Gaston Fiore, David Castineira (MIT), Brian Walsh, Yuval Barak Corren, Katherine Schlosser, John Arnold, Craig Smallwood, and many others I have not met.

Philosophy: we do not aspire to replace clinicians by “smart automatic systems/robots”, instead, we aspire to help medical teams make better decisions systematically.

Goals:

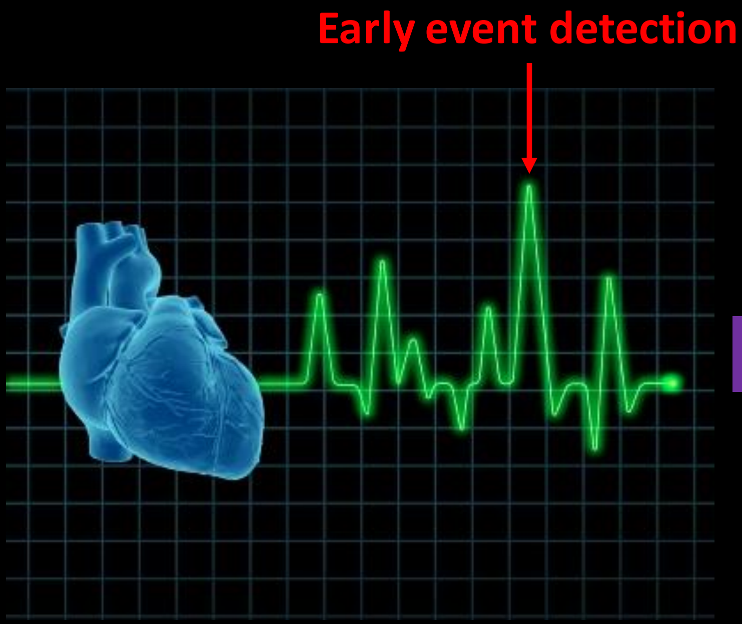
1. Use **real-time** data from monitors (vital signs + mechanical ventilation) to perform **event detection**

2. **Learn** from **historical patterns** to **improve care to children** and potentially **reduce costs** in the most expensive unit in the hospital.

Goals:

1. Use **real-time** data from monitors (vital signs + mechanical ventilation) to perform **event detection**

2. **Learn** from **historical patterns** to **improve care to children** and potentially **reduce costs** in the most expensive unit in the hospital.



a) Early detection of Ventilator Associated Conditions (VACs are associated to increased mortality)
b) Continuous assessment of readiness to extubate a patient (reduce re-intubation cases and prevent infections and VACs)
c) Early determination of long lengths of stay in the ICU (improve resource allocation across the hospital)
d) Real-time determination of vital signs percentile curves (improve patient assessment in real-time)
e) Continuous monitoring of sedation levels (may prevent over-sedation and may reduce length of stay)

Relevant publications

1. B. Patel**, F. Sperotto M. Molina**, S. Kimura, M. Delgado, M. Santillana J N Kheir. Computerized prediction of avoidable serum potassium testing in the cardiac intensive care unit. Pediatric Critical Care Medicine, 2020;22 (4)

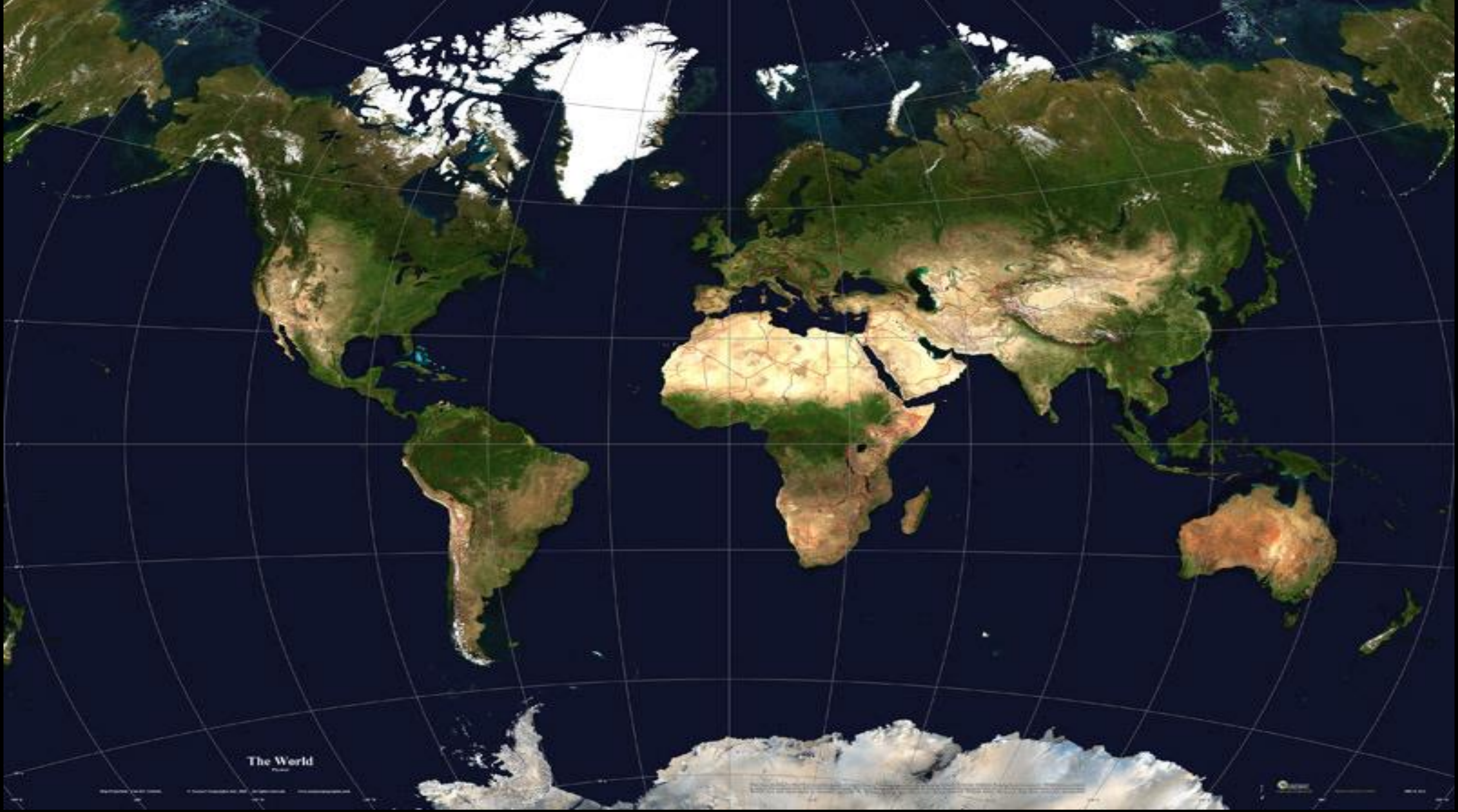
2. S Tideman**, M. Santillana, J. Bickel, B. Reis. Internet search query data improves fore- casts of daily emergency department volume. Journal of the American Medical Informatics Associ- ation. 2019; ocz154, 2019.

3. Smallwood C, Walsh B, Rettig J, Thompson J, M. Santillana, Arnold J. 955: A machine-learning algorithm for oxygenation response prediction in mechanically ventilated children. Critical Care Medicine 2016 Dec; Volume 44, Issue 12, pp:315. doi: 10.1097/01.ccm.0000509631.99570.c2

4. K R. Schlosser, G. Fiore**, C. D. Smallwood, J. Griffin, M. Santillana, J. H. Arnold Non-invasive ventilation is interrupted frequently and mostly used at night in the pediatric intensive care unit. Respiratory Care, 2019.

5. Schlosser K**, Smallwood C, Arnold J, Lee G, Priebe G, Walsh B, M. Santillana 1015: Identification of pediatric ventilator-associated conditions using continuous ventilator data. Critical Care Medicine. 2016 Dec; Volume 44, Issue 12, pp:330. doi: 10.1097/01.ccm.0000509691.59308.ae

6. Walsh B, Smallwood C, Rettig J, M. Santillana, Arnold J. 949: Development of heart, respiratory rate, and oxygenation saturation percentile curves in children. Critical Care Medicine 2016 Dec; Volume 44, Issue 11pp:313. doi: 10.1097/01.ccm.0000509625.30958.52



Thank you!

Contact: msantill@g.harvard.edu