Modeling and forecasting using genomic surveillance: lessons from wastewater and COVID-19 variants

Scott Olesen
US CDC Center for Forecasting & Outbreak Analytics

Wastewater data

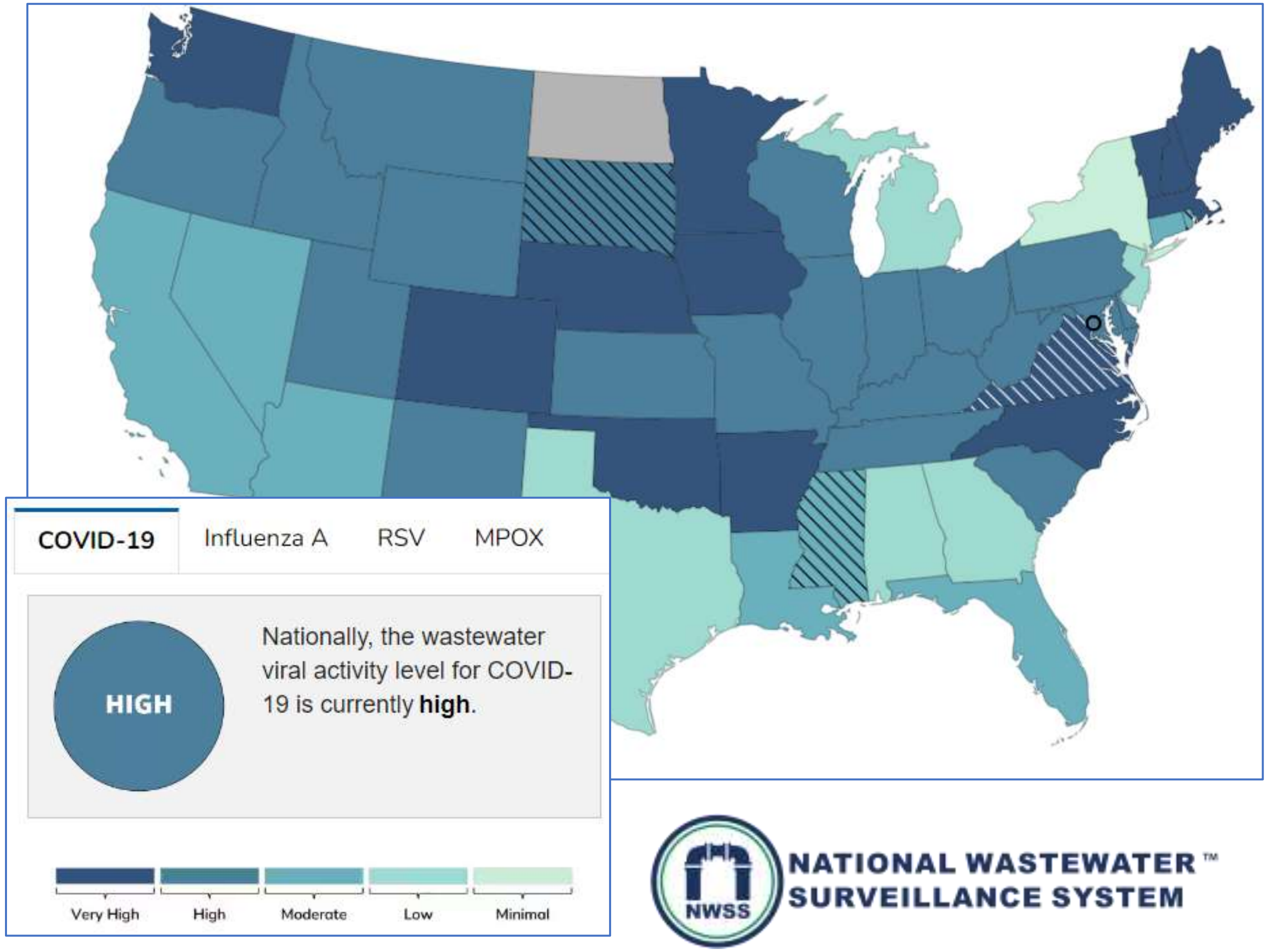
Wastewater surveillance is a promising technology

Potential for **early warning**

Infection → *Shedding* → *Healthcare encounter* → *Reporting*

Potential for parallel quantification of **dozens of pathogens**

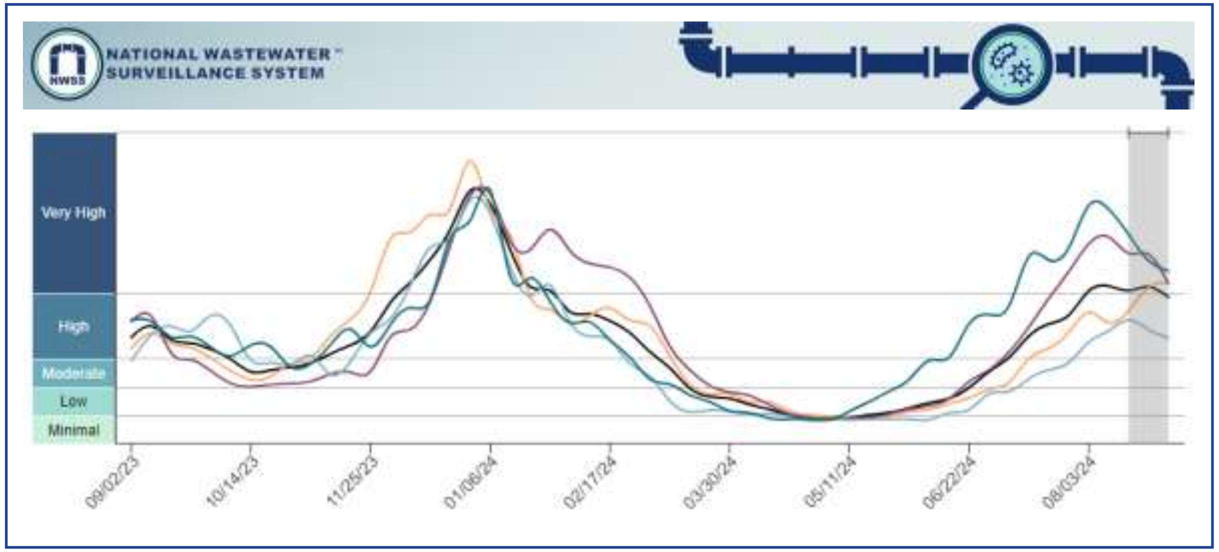
Representative, efficient, **community-wide sampling**



Nationally, the wastewater viral activity level for COVID-19 is currently **high**.

NWSS NATIONAL WASTEWATER™ SURVEILLANCE SYSTEM

In practice, wastewater data-to-action has been challenging



https://www.cdc.gov/nwss/rv/COVID19-nationaltrend.html

New York City
DC
Guam
Puerto Rico
US Virgin Islands

SARS-CoV-2 RNA wastewater levels, United States
15-day percent change by sewershed

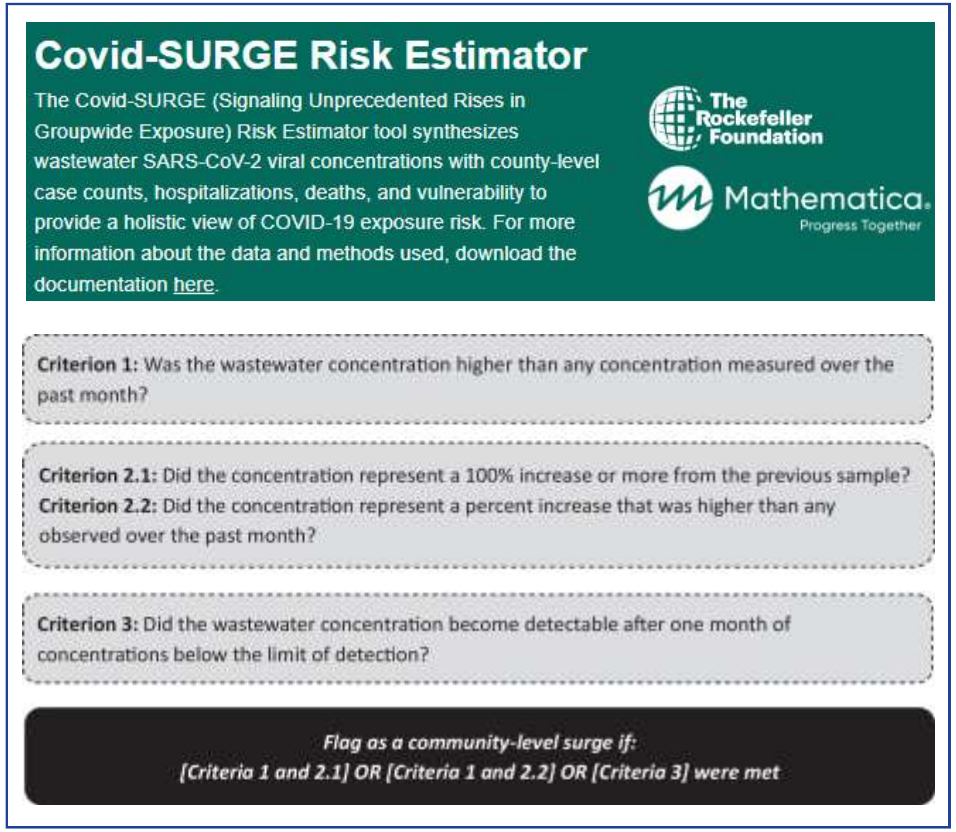
15-day % change category	Num. sites	% sites	Category change in last 7 days
-100%	31	7	33%
-99% to -10%	187	47	-22%
-9% to 0%	26	6	8%
1% to 9%	13	3	-7%
10% to 99%	38	9	-21%
100% to 999%	60	14	-6%
1000% or more	53	13	50%

Total sites with current data: 410
Total number of wastewater sampling sites: 698

Select legend categories to filter points on the map.

● -100% ● -99% to -10% ● -9% to 0% ● 1% to 9% ● 10% to 99% ● 100% to 999% ● 1000% or more ● No recent data

https://data.ohio.gov/wps/portal/gov/data/projects/wastewater+surveillance



Covid-SURGE Risk Estimator

The Covid-SURGE (Signaling Unprecedented Rises in Groupwide Exposure) Risk Estimator tool synthesizes wastewater SARS-CoV-2 viral concentrations with county-level case counts, hospitalizations, deaths, and vulnerability to provide a holistic view of COVID-19 exposure risk. For more information about the data and methods used, download the documentation here.

The Rockefeller Foundation

Mathematica. Progress Together

Criterion 1: Was the wastewater concentration higher than any concentration measured over the past month?

Criterion 2.1: Did the concentration represent a 100% increase or more from the previous sample?
Criterion 2.2: Did the concentration represent a percent increase that was higher than any observed over the past month?

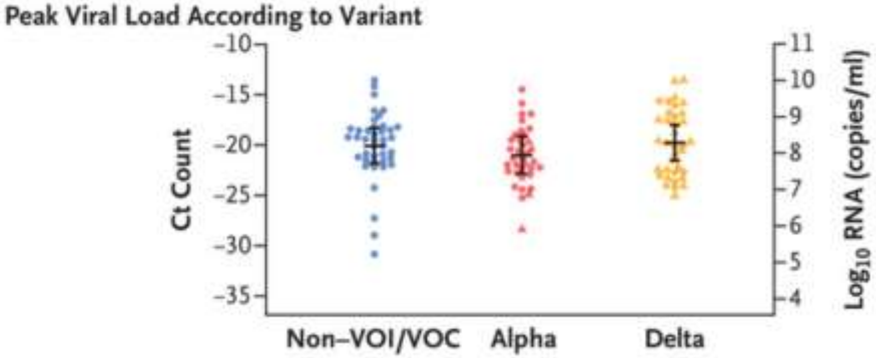
Criterion 3: Did the wastewater concentration become detectable after one month of concentrations below the limit of detection?

Flag as a community-level surge if:
[Criteria 1 and 2.1] OR [Criteria 1 and 2.2] OR [Criteria 3] were met

Keshaviah *PNAS* (2023) DOI: 10.1073/pnas.2216021120

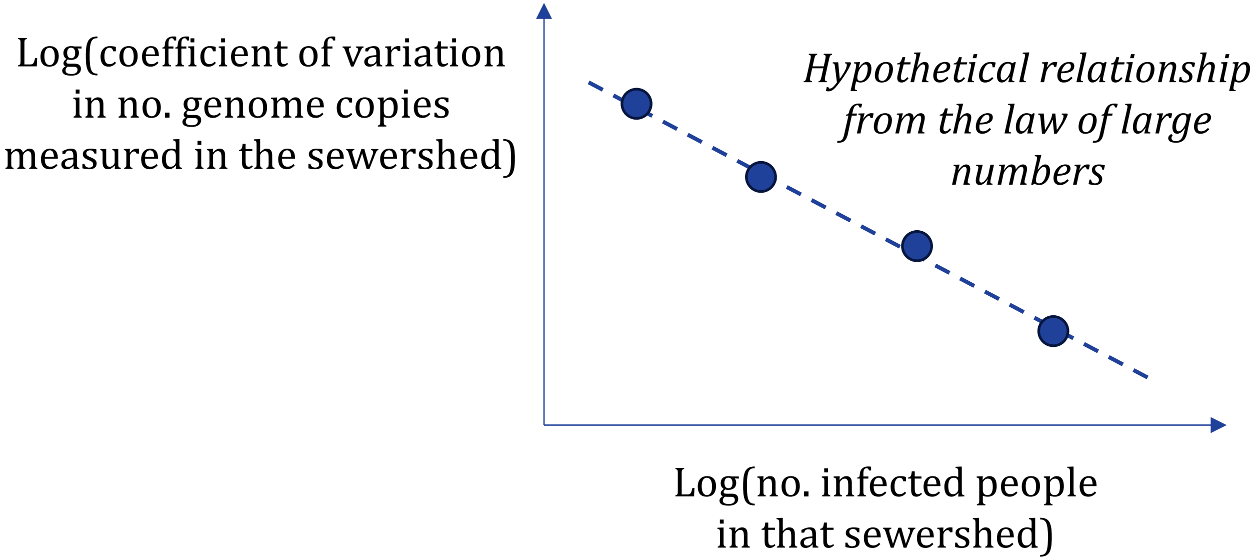
Wastewater data has intrinsic, biological noise

Rates of shedding into wastewater vary between individuals. For comparison, peak nasal viral loads vary >100x between individuals.



Kissler *NEJM* (2021) doi: 10.1056/NEJMc2102507

If the wastewater signal were the mean shedding across infected individuals in a sewershed, then variance in the wastewater signal would decline with higher **disease prevalence** and higher **population size**.

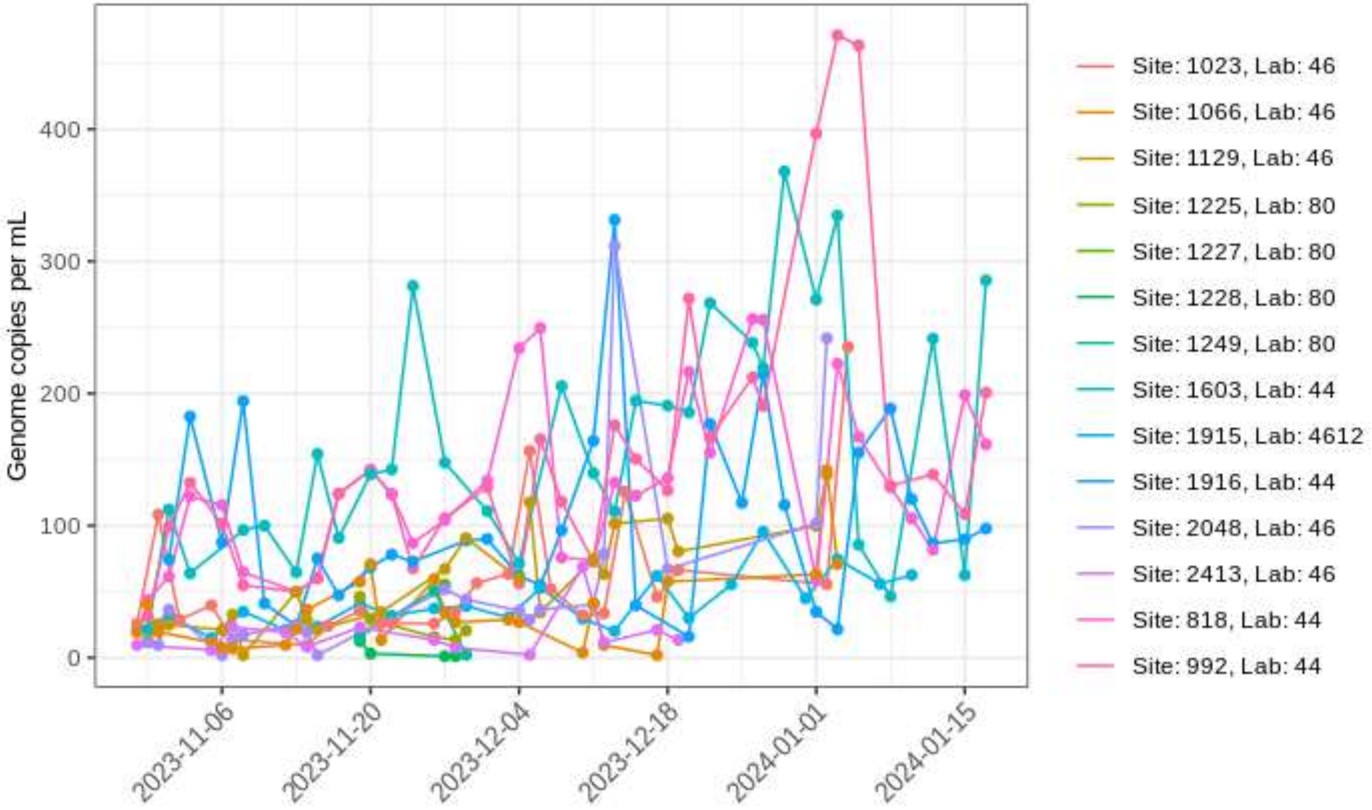


Log(coefficient of variation in no. genome copies measured in the sewershed)

Hypothetical relationship from the law of large numbers

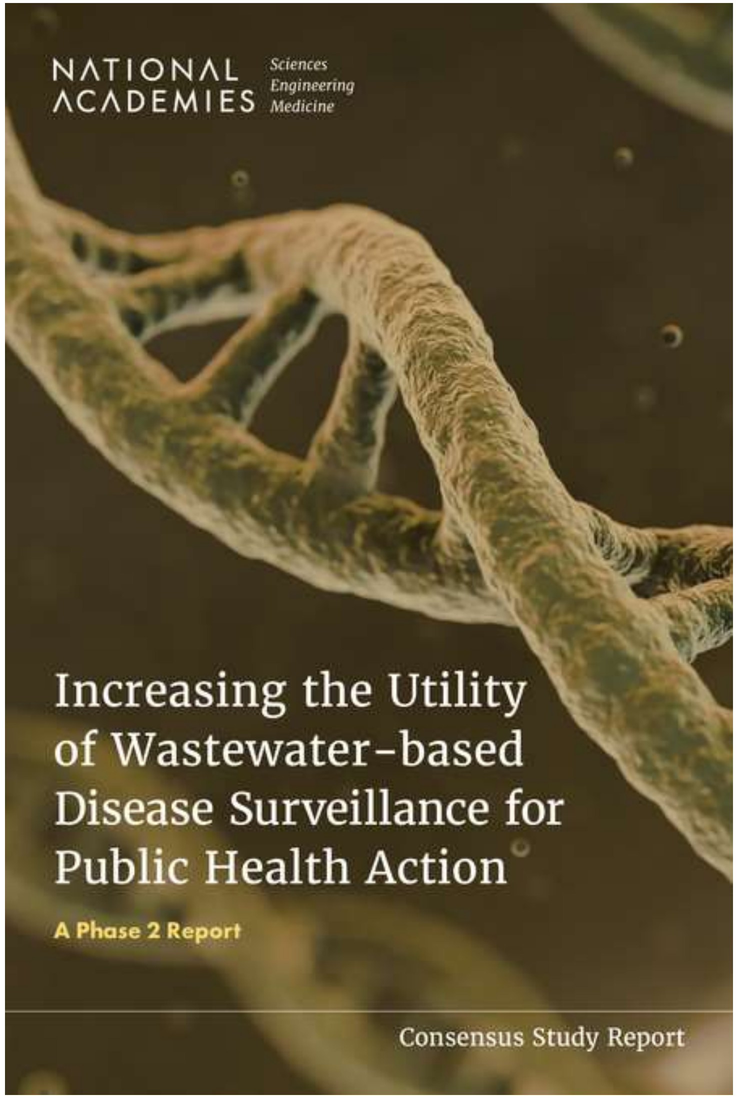
Log(no. infected people in that sewershed)

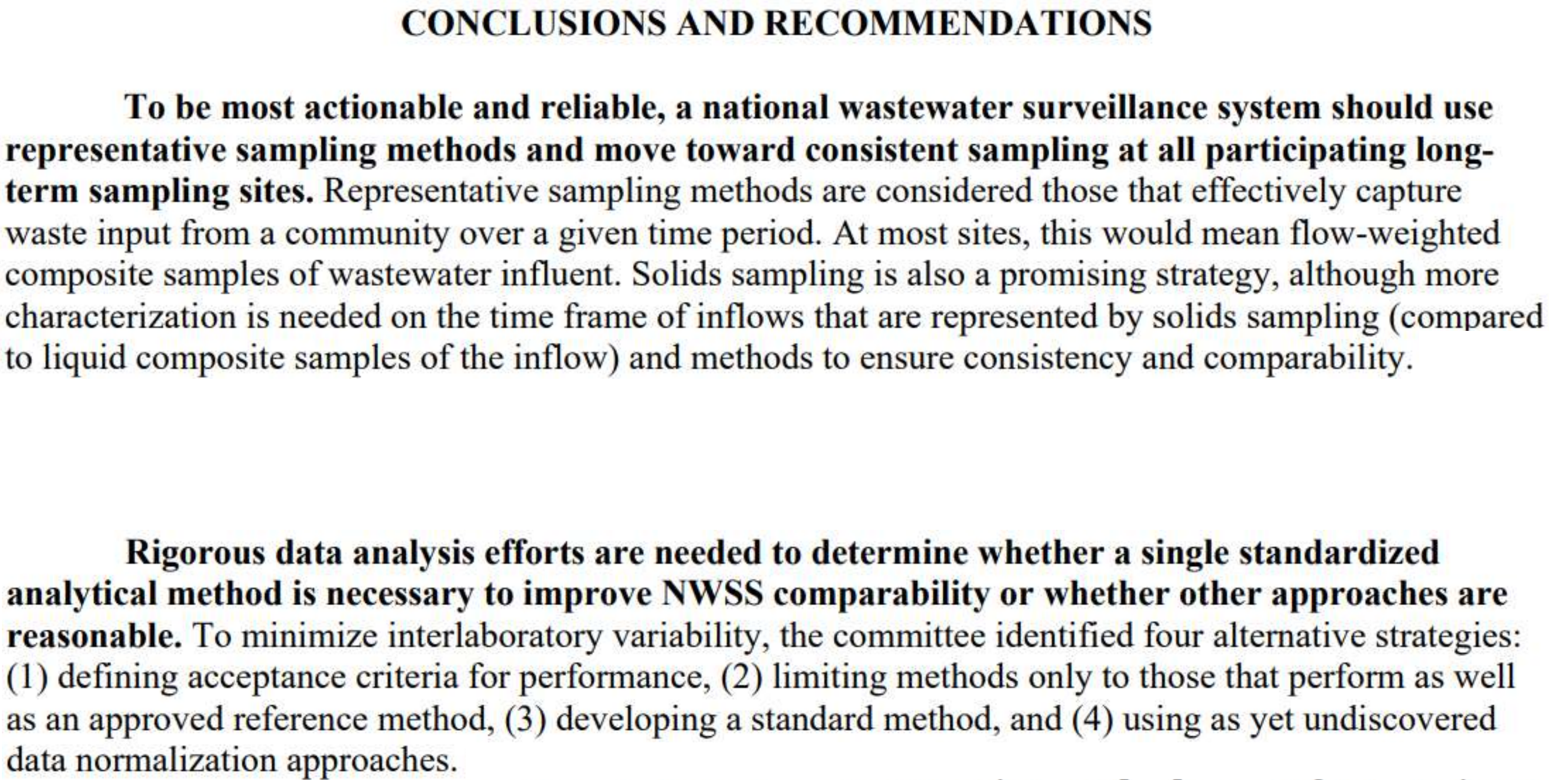
Wastewater has noise due to sites, sampling, and labs



- Some sites report nearly daily, others less than once per week
- Some sites have high sample-to-sample variability, others much less
- Some sites report data within the week, others report weeks later
- Sites can drop in and out
- Sites switch sampling method, lab, or lab method

Where possible, decrease variability in methodology


NATIONAL ACADEMIES
Sciences
Engineering
Medicine
Increasing the Utility of Wastewater-based Disease Surveillance for Public Health Action
A Phase 2 Report
Consensus Study Report

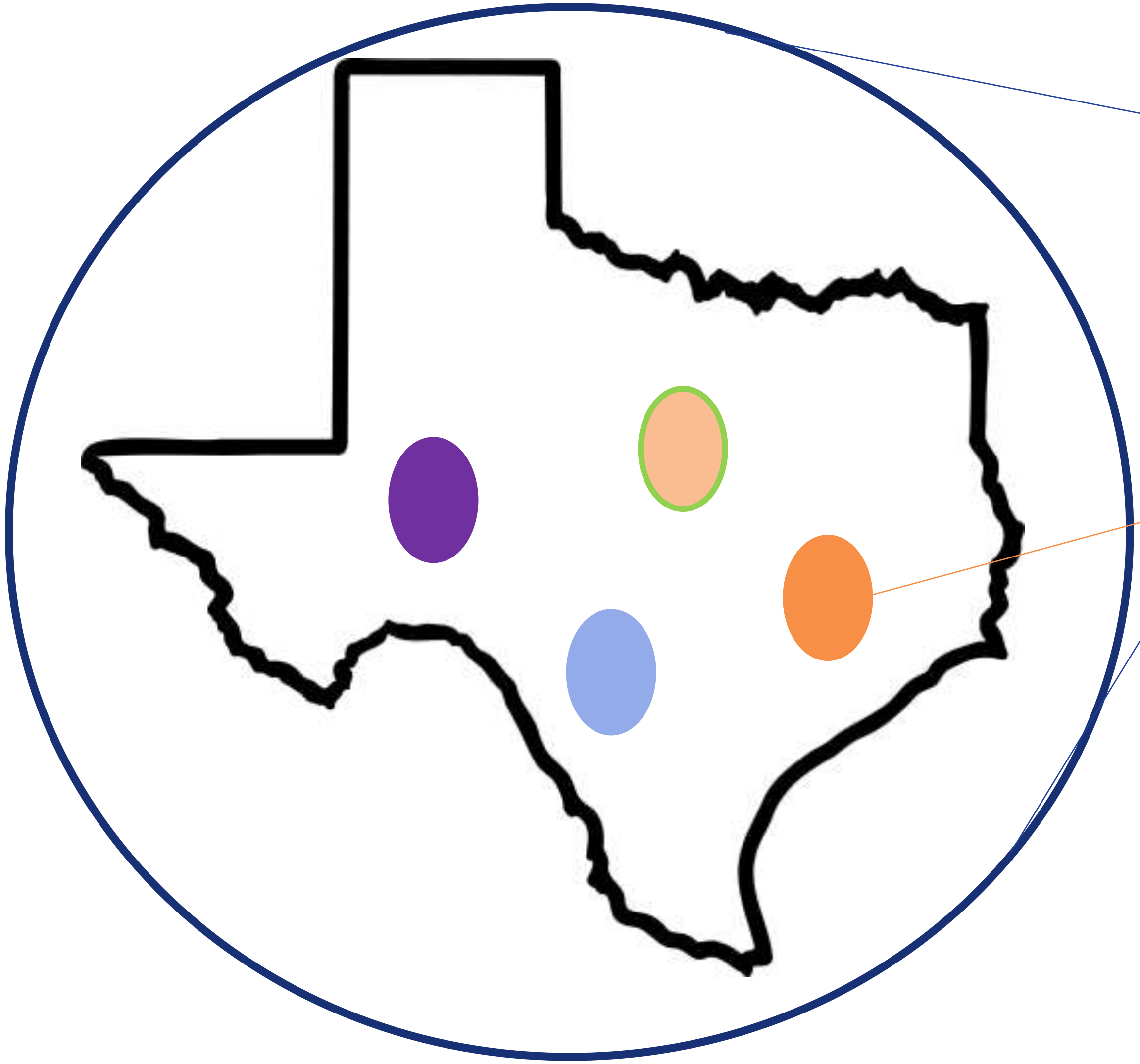


CONCLUSIONS AND RECOMMENDATIONS

To be most actionable and reliable, a national wastewater surveillance system should use representative sampling methods and move toward consistent sampling at all participating long-term sampling sites. Representative sampling methods are considered those that effectively capture waste input from a community over a given time period. At most sites, this would mean flow-weighted composite samples of wastewater influent. Solids sampling is also a promising strategy, although more characterization is needed on the time frame of inflows that are represented by solids sampling (compared to liquid composite samples of the inflow) and methods to ensure consistency and comparability.

Rigorous data analysis efforts are needed to determine whether a single standardized analytical method is necessary to improve NWSS comparability or whether other approaches are reasonable. To minimize interlaboratory variability, the committee identified four alternative strategies: (1) defining acceptance criteria for performance, (2) limiting methods only to those that perform as well as an approved reference method, (3) developing a standard method, and (4) using as yet undiscovered data normalization approaches.

Build analytical methods that work around known variabilities

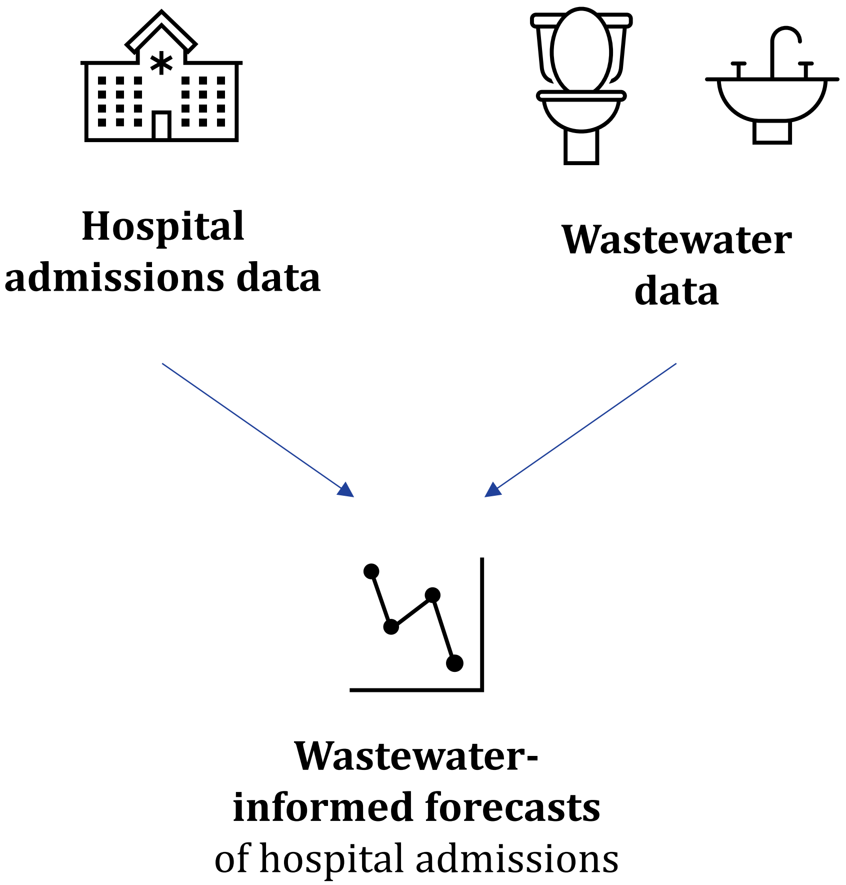


Entire state population contributes to **state-level hospital admissions**

Populations in each wastewater catchment area contribute to **site-level wastewater concentration**

Account statistically for variations between data sites, including: wastewater facilities, wastewater lab methods, and hospital admissions reporting systems

Use Bayesian signal fusion to combine wastewater data with existing data streams



Hospital admissions data

Wastewater data

Wastewater-informed forecasts of hospital admissions

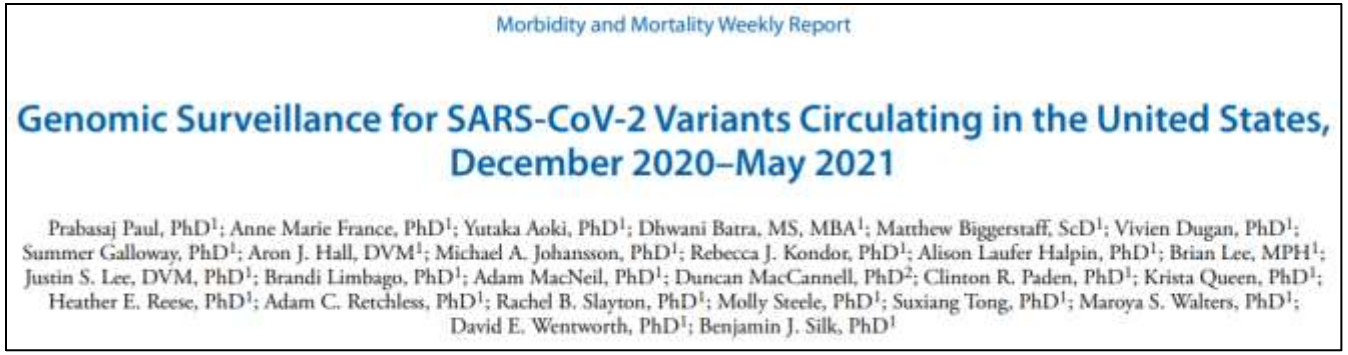
Bayesian hierarchical approach for wastewater data:

- Each site has a true number of people infected, connected to an epidemiological dynamics model
- There is some function that relates people infected with genomes shed into wastewater
- Each site has some adjustment factor, between true genomes shed and observed concentrations
- Adjustment factors are partially pooled

https://github.com/cdcgov/wastewater-informed-covid-forecasting

SARS-CoV-2 variant prevalences

CDC's variant nowcasting methodology has stayed mostly constant

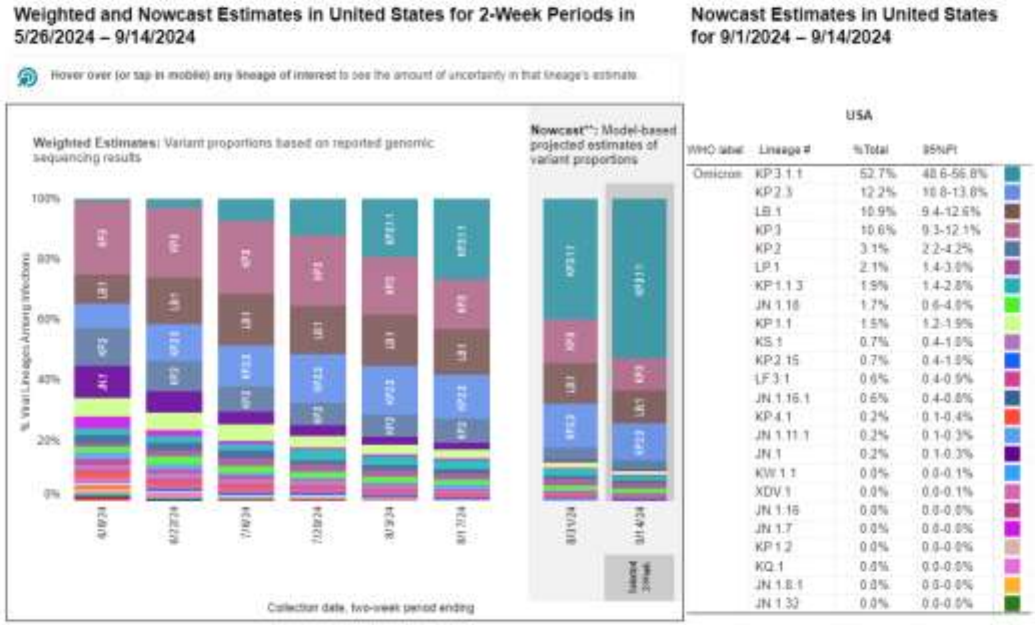


Morbidity and Mortality Weekly Report

Genomic Surveillance for SARS-CoV-2 Variants Circulating in the United States, December 2020–May 2021

Prabasaj Paul, PhD[1]; Anne Marie France, PhD[1]; Yutaka Aoki, PhD[1]; Dhwani Batra, MS, MBA[1]; Matthew Biggerstaff, ScD[1]; Vivien Dugan, PhD[1]; Summer Galloway, PhD[1]; Aron J. Hall, DVM[1]; Michael A. Johansson, PhD[1]; Rebecca J. Kondor, PhD[1]; Alison Laufer Halpin, PhD[1]; Brian Lee, MPH[1]; Justin S. Lee, DVM, PhD[1]; Brandi Limbago, PhD[1]; Adam MacNeil, PhD[1]; Duncan MacCannell, PhD[2]; Clinton R. Paden, PhD[1]; Krista Queen, PhD[1]; Heather E. Reese, PhD[1]; Adam C. Retchless, PhD[1]; Rachel B. Slayton, PhD[1]; Molly Steele, PhD[1]; Suxiang Tong, PhD[1]; Maroya S. Walters, PhD[1]; David E. Wentworth, PhD[1]; Benjamin J. Silk, PhD[1]

https://www.cdc.gov/mmwr/volumes/70/wr/mm7023a3.htm



Weighted and Nowcast Estimates in United States for 2-Week Periods in 5/26/2024 – 9/14/2024

Hover over (or tap in mobile) any lineage of interest to see the amount of uncertainty in that lineage's estimate.

Weighted Estimates: Variant proportions based on reported genomic sequencing results

Nowcast**: Model-based projected estimates of variant proportions

100%
80%
60%
40%
20%
0%
% Viral Lineages Among Infections
Collection date, two-week period ending

Nowcast Estimates in United States for 9/1/2024 – 9/14/2024

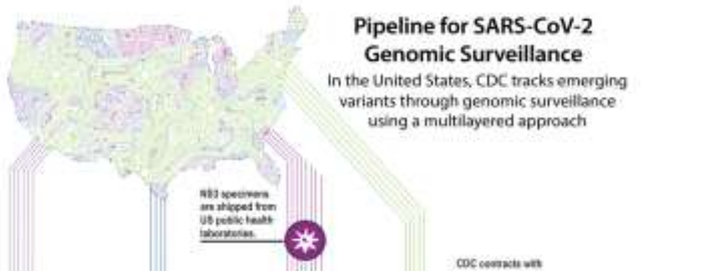
USA

WHO label	Lineage #	%Total	95%PI
Omicron	KP.3.1.1	52.7%	48.6-56.8%
	KP.2.3	12.2%	10.8-13.8%
	LB.1	10.9%	9.4-12.6%
	KP.3	10.6%	9.3-12.1%
	KP.2	3.1%	2.2-4.2%
	LP.1	2.1%	1.4-3.0%
	KP.1.1.3	1.9%	1.4-2.8%
	JN.1.18	1.7%	0.6-4.0%
	KP.1.1	1.5%	1.2-1.9%
	KS.1	0.7%	0.4-1.0%
	KP.2.15	0.7%	0.4-1.0%
	LF.3.1	0.6%	0.4-0.9%
	JN.1.16.1	0.6%	0.4-0.8%
	KP.4.1	0.2%	0.1-0.4%
	JN.1.11.1	0.2%	0.1-0.3%
	JN.1	0.2%	0.1-0.3%
	KW.1.1	0.0%	0.0-0.1%
	XDV.1	0.0%	0.0-0.1%
	JN.1.16	0.0%	0.0-0.0%
	JN.1.7	0.0%	0.0-0.0%
	KP.1.2	0.0%	0.0-0.0%
	KQ.1	0.0%	0.0-0.0%
	JN.1.8.1	0.0%	0.0-0.0%
	JN.1.33	0.0%	0.0-0.0%

- Project forward in time using a multinomial regression approach: the probability that a sample is in a particulate taxon is varies in time (with log-odds linear in time)
- Account for heterogeneity in sampling practices:
  - w_i = probability an infected person gets a PCR
  - w_p = probability a positive PCR is sequenced
- Means and variances computed using a survey approach with weights $1/(w_i w_p)$

Using multiple genomic signals could improve forecasts

Clinical genomic data



Pipeline for SARS-CoV-2 Genomic Surveillance

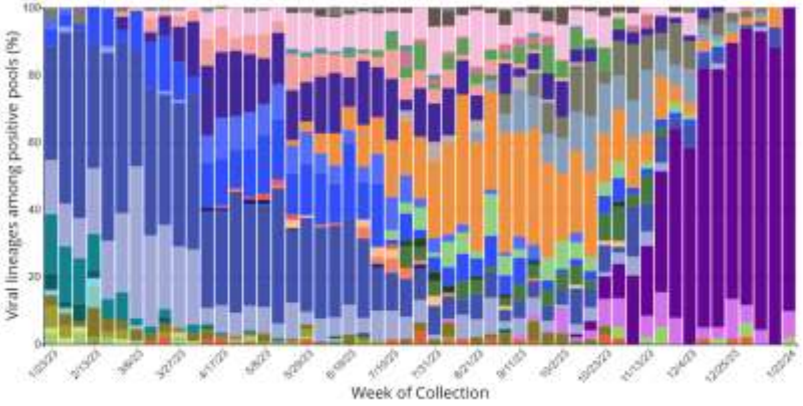
In the United States, CDC tracks emerging variants through genomic surveillance using a multilayered approach

NS3 specimens are shipped from US public health laboratories.
CDC contracts with

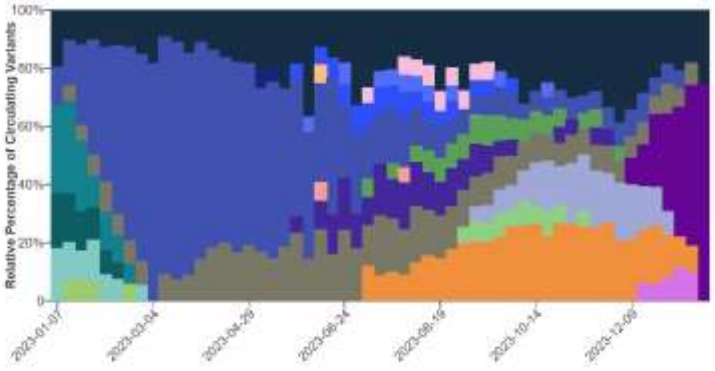
Legacy model

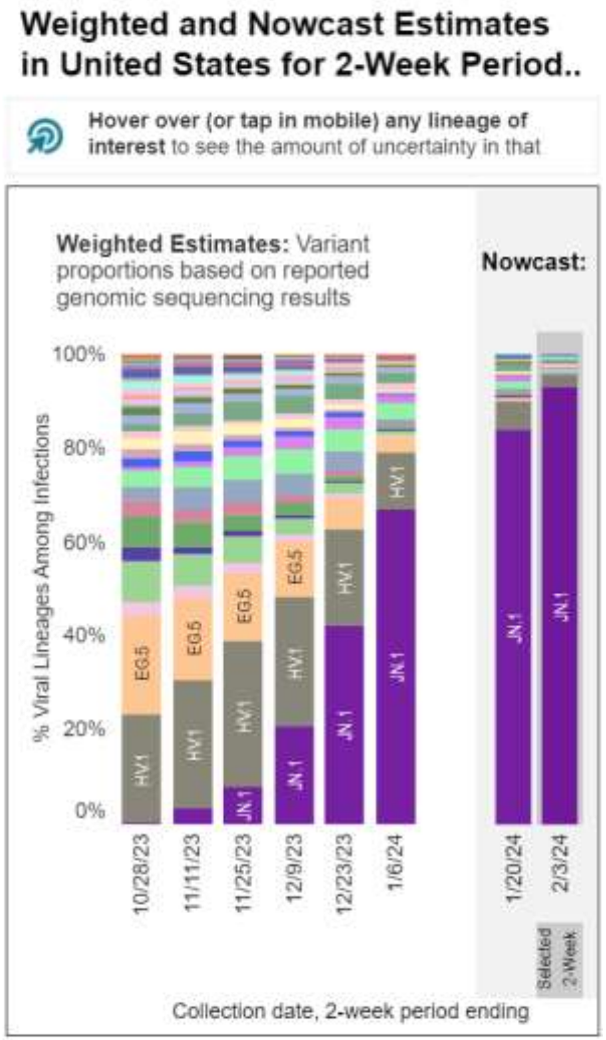
Data fusion model

Traveler-based genomic data



Wastewater genomic data





Weighted and Nowcast Estimates in United States for 2-Week Period..

Hover over (or tap in mobile) any lineage of interest to see the amount of uncertainty in that

Weighted Estimates: Variant proportions based on reported genomic sequencing results

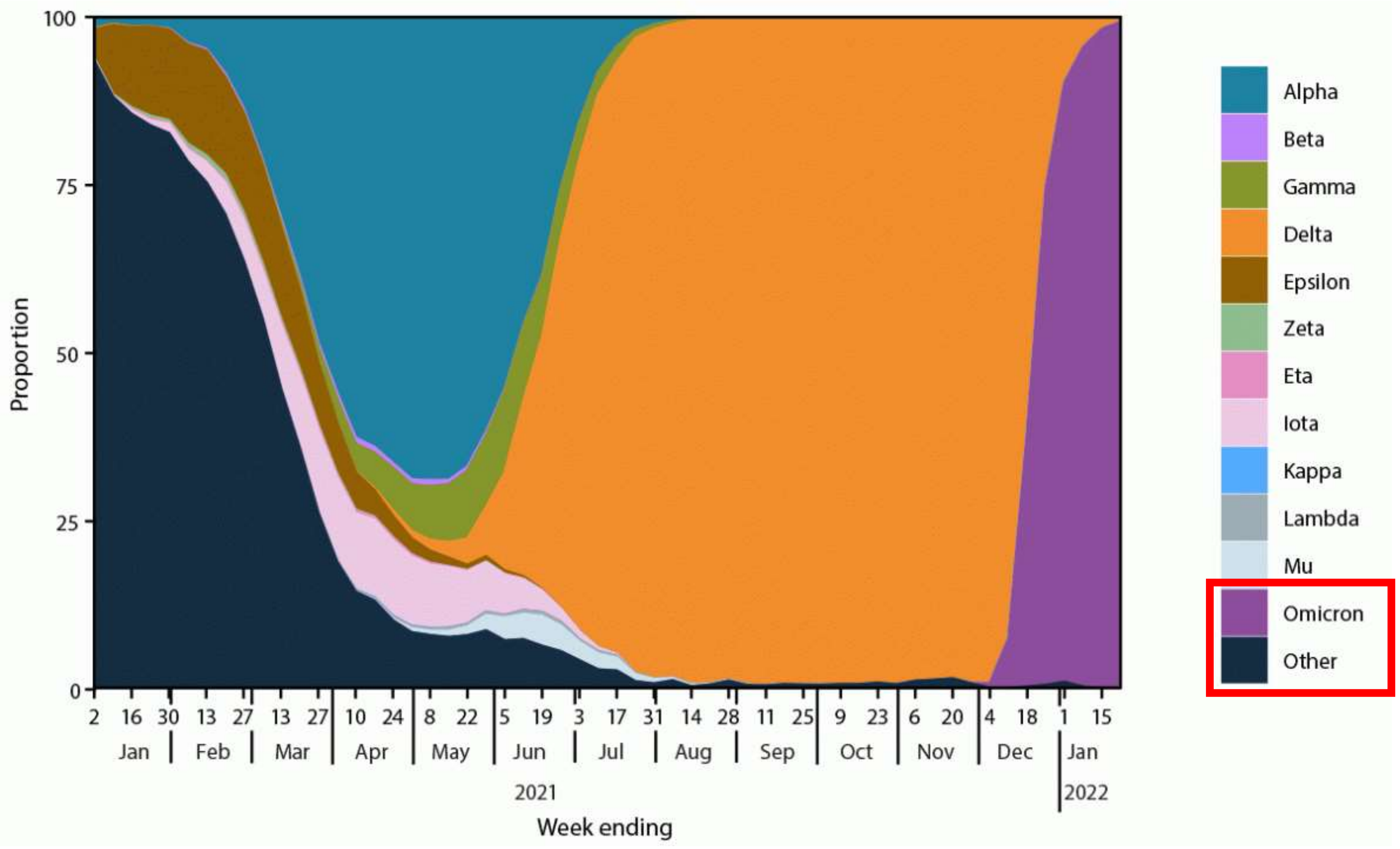
Nowcast:

Different genomic signals cannot be naively combined

Data stream	Key strengths	(Human) populations	Delay	Data type
National SARS-CoV-2 Strain Surveillance (NS3)	Genome quality, sample size	Hospitalized cases	Weeks	Counts
National Wastewater Surveillance System (NWSS)	Scope of monitoring, turnaround time	Hundreds of communities across the US	Days	Proportions
Traveler-based Genomic Surveillance (TGS)	Genome quality, turnaround time	International travelers	~1 week	Counts/pools

The relevant entities to be modeled change over time

Genomic Surveillance for SARS-CoV-2 Variants: Predominance of the Delta (B.1.617.2) and Omicron (B.1.1.529) Variants — United States, June 2021–January 2022



For example,

- In the US in January 2022, “Omicron” could reasonably mean B.1.1.529
- A month later, it was important to distinguish BA.1 and BA.2
- Later, BA.4, BA.5, XBB, etc.

For modelers, the applicable modeling units (i.e., taxa) could be driven by cladistics or epidemiology

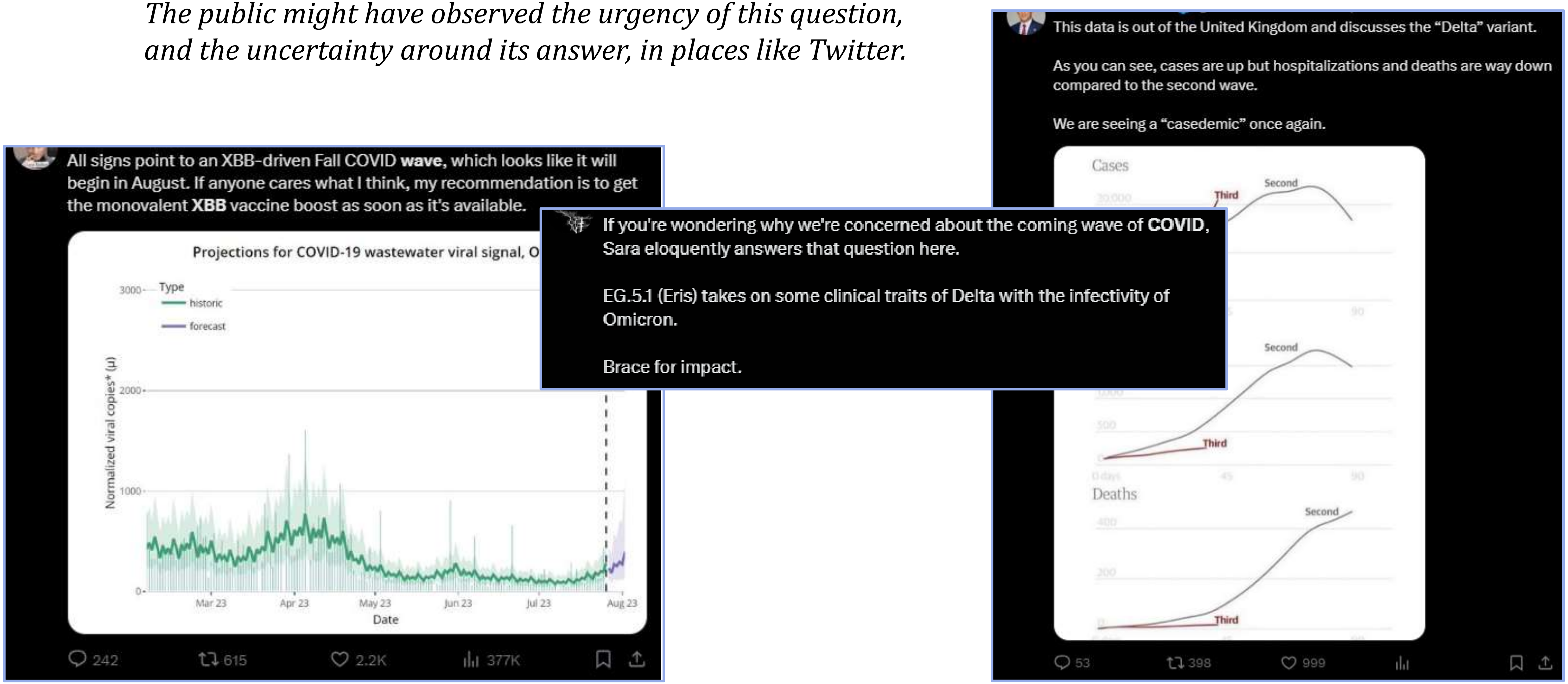
https://www.cdc.gov/mmwr/volumes/71/wr/mm7106a4.htm

Relevant forecasting targets will differ by application

Question	Timing	Data quantity & quality	Utility of current variant nowcasting methods
Will this new taxon (e.g., variant) trigger a wave?	Early, at variant emergence	Relatively poor	Relatively poor
When will taxon X achieve Y% prevalence?	Early to peak	Relatively poor to relatively high	Relatively high

The most urgent question was "will this variant drive a wave?"

The public might have observed the urgency of this question, and the uncertainty around its answer, in places like Twitter.



All signs point to an XBB-driven Fall COVID **wave**, which looks like it will begin in August. If anyone cares what I think, my recommendation is to get the monovalent **XBB** vaccine boost as soon as it's available.

If you're wondering why we're concerned about the coming wave of **COVID**, Sara eloquently answers that question here.

EG.5.1 (Eris) takes on some clinical traits of Delta with the infectivity of Omicron.

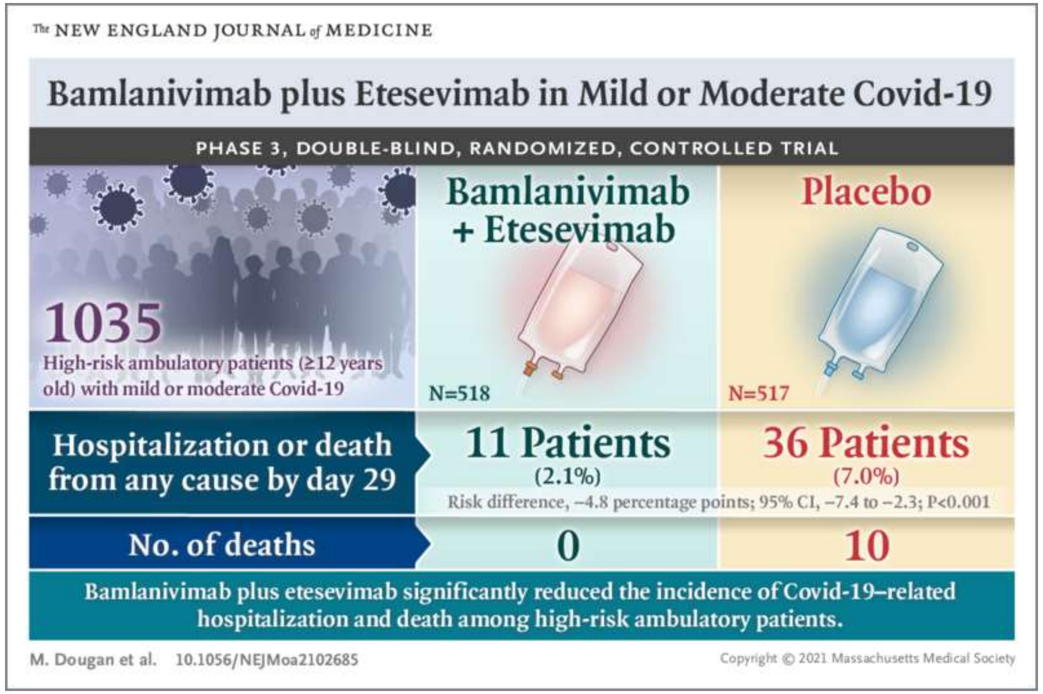
Brace for impact.

This data is out of the United Kingdom and discusses the "Delta" variant.

As you can see, cases are up but hospitalizations and deaths are way down compared to the second wave.

We are seeing a "casedemic" once again.

Variant forecasting informed monoclonal antibody deployment



The NEW ENGLAND JOURNAL of MEDICINE

Bamlanivimab plus Etesevimab in Mild or Moderate Covid-19

PHASE 3, DOUBLE-BLIND, RANDOMIZED, CONTROLLED TRIAL

1035
High-risk ambulatory patients (≥12 years old) with mild or moderate Covid-19

	Bamlanivimab + Etesevimab N=518	Placebo N=517
Hospitalization or death from any cause by day 29	11 Patients (2.1%)	36 Patients (7.0%)
	Risk difference, −4.8 percentage points; 95% CI, −7.4 to −2.3; P<0.001	
No. of deaths	0	10

Bamlanivimab plus etesevimab significantly reduced the incidence of Covid-19–related hospitalization and death among high-risk ambulatory patients.

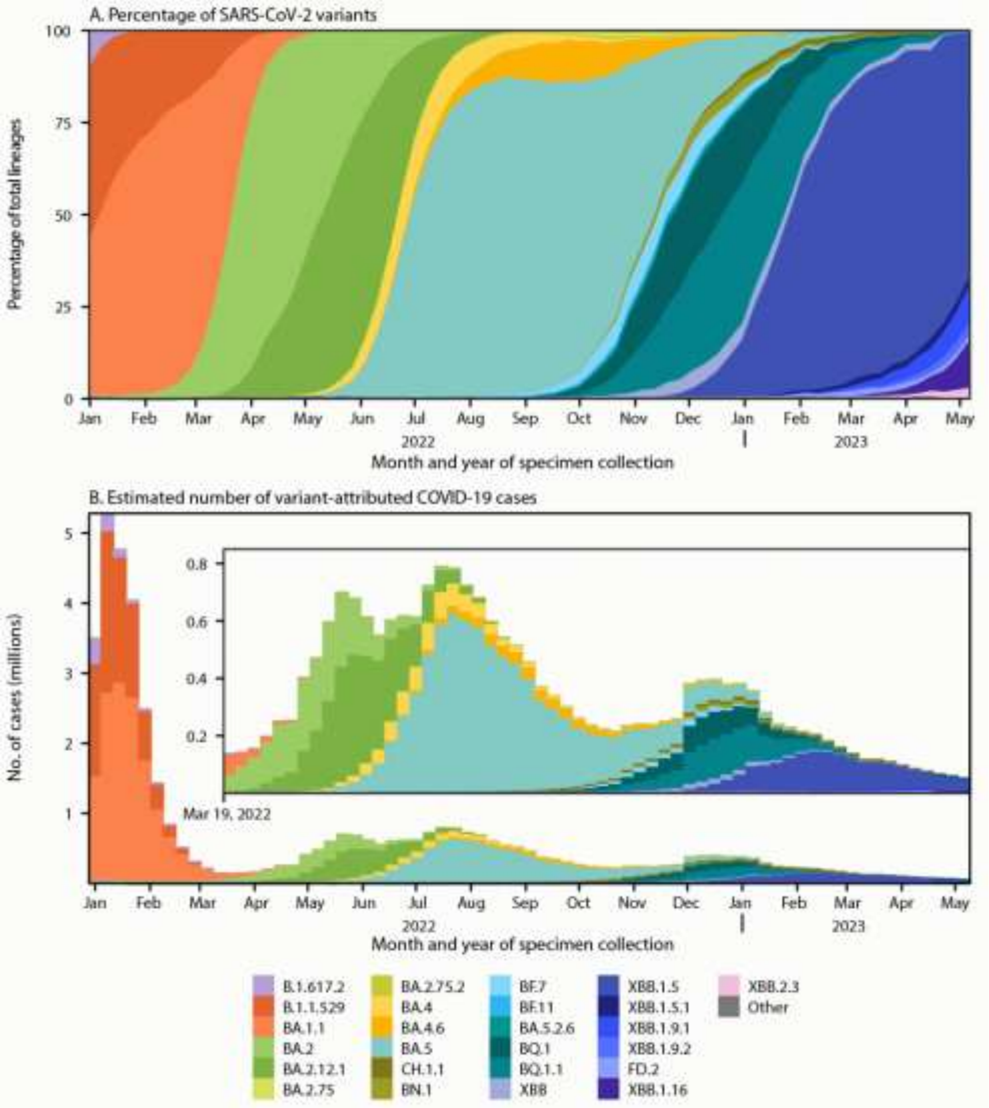
M. Dougan et al. 10.1056/NEJMoa2102685

Copyright © 2021 Massachusetts Medical Society

For example, **bamlanivimab/etesevimab**

- Feb 2021: Approved for use (US FDA Emergency Use Authorization).
- Jun 2021: Distribution paused as **Beta and Gamma variants grew.** Considered ineffective against those variants.
- Sep 2021: Distribution restarted when **Beta and Gamma failed to spread >5%**.
- Oct 2021: Distribution to Hawaii paused because **Hawaii had >5% Delta**. Considered ineffective against Delta.
- Oct 2021: Distribution restarted because determined that Delta was not resistant.
- Jan 2022: Distribution stopped because considered ineffective against **Omicron**.
- EUA later revoked, after **dominance of Omicron** was assured.

There is likely utility in jointly modeling variant prevalences and counts of infections



Growth in numbers of infections is more important than growth in proportional prevalence.

- Proportional prevalence is not the same as total infections.
- High proportional prevalence of a taxon (e.g., “variant”) could be a good thing (e.g., if it’s a low-virulence taxon).

However, it is unclear if multi-strain forecasts of infection counts will outperform a combination of (1) strain-agnostic infection count forecasts, and (2) variant prevalence forecasts.

https://www.cdc.gov/mmwr/volumes/72/wr/mm7224a2.htm

Conclusions

1. **Respect the data generating process.** Build models that account for known sources of noise, either statistically or mechanistically.
2. **Judge models on performance.** Build models based on what will plausibly improve performance.
3. For a model to be useful to public health, it must **demonstrably outperform the methods actually used by public health practitioners**. A wastewater-only-in, hospitalizations-out model for forecasting hospitalizations will not demonstrably outperform eyeballing of hospitalization data trends.
4. Models should be built within frameworks that make **evaluation and comparison** as simple as possible.
5. **Signal fusion** for wastewater modeling is difficult but full of promise.

Acknowledgements and disclaimer

Scott Olesen (CDC/CFA) <ULP7@CDC.GOV>

Andy Magee (CDC/CFA) – multisignal variant nowcasting model development

Thanasi Bakis (CDC/CFA contractor) – variant nowcasting model iteration frameworks

Clint Paden (CDC/NCIRD/CORVD) – CDC variant nowcasting

Dhwani Batra (CDC/NCEZID/DIDRI/OAMD) – CDC genomics data & platforms

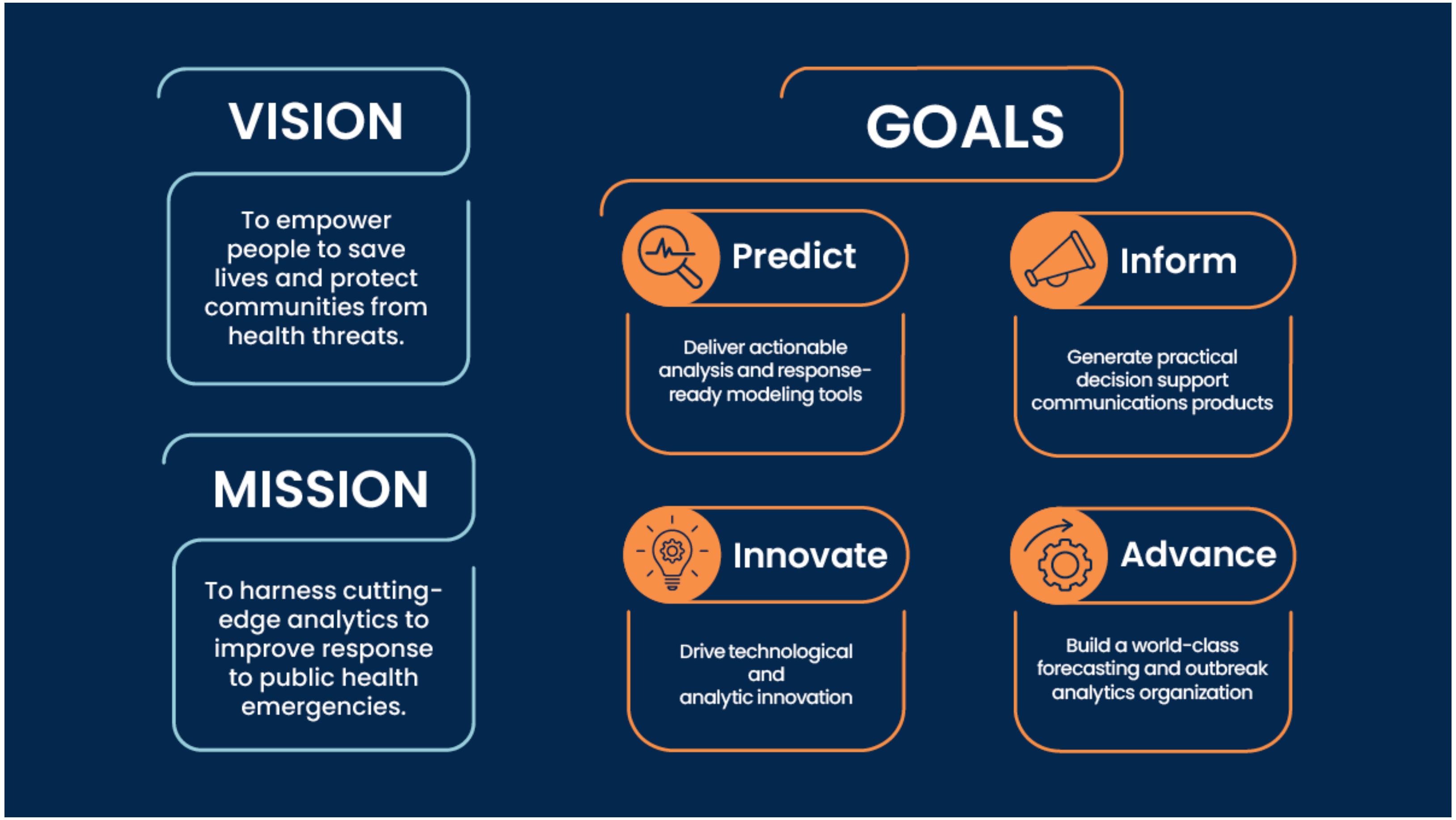
Kaitlyn Johnson (CDC/CFA) – wastewater-based COVID-19 hospitalization forecasts

Dylan Morris (CDC/CFA) – CFA Short Term Forecasting team lead

https://github.com/cdcgov/wastewater-informed-covid-forecasting

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

CFA is hiring!



VISION

To empower people to save lives and protect communities from health threats.

MISSION

To harness cutting-edge analytics to improve response to public health emergencies.

GOALS

Predict

Deliver actionable analysis and response-ready modeling tools

Inform

Generate practical decision support communications products

Innovate

Drive technological and analytic innovation

Advance

Build a world-class forecasting and outbreak analytics organization

Contact Scott Olesen <ULP7@CDC.GOV> for guidance on navigating federal hiring system