BioSignal

Wastewater Anomaly Contextualization for Pandemic Early Warning

Jack Lakkapragada · AIxBio Hackathon — Apart Research x BlueDot Impact x Cambridge Biosecurity Hub · April 2026

Track 2: Pandemic Early Warning · Sponsored by Measuring AI Progress

The CDC National Wastewater Surveillance System (NWSS) already detects anomalies in viral wastewater signal. The gap is what happens next: public health officials receive a percentile number with no actionable intelligence about why it matters or what to do. BioSignal addresses this data-action gap by building a three-layer pipeline on top of existing CDC infrastructure. A deterministic scoring layer ranks sites by population-weighted priority. An LLM intelligence layer — activated only after the statistics confirm an anomaly — generates structured 3-paragraph situational reports (SITREPs) identifying catchment type, signal drivers, and specific recommended actions. Validated against the December 2023 JN.1 variant surge, BioSignal correctly surfaced the New Jersey Northeast corridor, Las Vegas metropolitan area, and Boston metro as URGENT/HIGH priority sites 1-2 weeks before national hospitalization peaks. The system is fully open-source, runs on public CDC data, and requires no specialized biosecurity infrastructure to deploy.

1. The Problem

Wastewater-based epidemiology has proven itself as an early warning signal — the CDC's NWSS detected SARS-CoV-2 trends in community wastewater weeks before equivalent signals appeared in clinical case counts. As of 2024, NWSS covers hundreds of sites representing over 140 million Americans, with data updated weekly.

The problem is not detection. The CDC's percentile metric — a site-normalized measure of current viral concentration relative to historical maximum — already identifies when a site is behaving anomalously. The problem is the data-action gap: a public health official in Essex County, New Jersey receives a notification that their wastewater site is at the 98th percentile. They have no context for what is driving it, no structured assessment of which population subgroups are at elevated risk, and no specific recommended actions. The signal exists. The intelligence layer does not.

Alarm fatigue — where practitioners receive too many undifferentiated alerts without actionable guidance — reduces system utility and delays response. BioSignal is designed to close this gap without replacing existing CDC infrastructure.

2. Approach

BioSignal implements a three-layer pipeline. The core design principle is strict separation between statistical detection and LLM contextualization: the model does not decide whether an anomaly exists. The statistics do.

Figure 3 — BioSignal System Architecture

Figure 3 — BioSignal system architecture. Each layer has a distinct, non-overlapping role.

Layer 1 — Data Pipeline

Raw NWSS data contains significant data quality issues not documented in the public-facing dataset description. Of 837,382 rows, 186,213 (22%) were removed after applying sentinel value filters: percentile values exceeding 100 (including a 999 error code), ptc_15d values reaching 2,147,483,647 (32-bit integer overflow artifact), and out-of-range percentage changes. Failure to apply these filters would produce false URGENT alerts from database artifacts rather than biological signal.

Layer 2 — Population-Weighted Priority Score

After cleaning, each site receives a priority score:

```
Score = (percentile x 0.5) + (log10(population_served) x 10)
```

The percentile is the primary signal — a CDC-normalized metric accounting for baseline variability. Log-scale population weighting serves as a tiebreaker: large catchments serving major urban corridors receive priority over small sites with identical percentile scores, reflecting that transmission potential scales with population density.

Layer 3 — LLM Intelligence Layer

When a site exceeds the alert threshold (percentile >= 80), the system calls a large language model acting as a Public Health Intelligence Officer. The system prompt explicitly states: Your job is NOT to re-evaluate whether the signal is real. The math has already confirmed it. The model produces a structured 3-paragraph SITREP covering: (1) catchment profile — density, critical infrastructure, vulnerable populations; (2) signal drivers — events, travel patterns, seasonal factors; (3) recommended actions — concrete, jurisdiction-specific. Each SITREP concludes with a priority tier: URGENT, HIGH, or ELEVATED.

> This architecture prevents the circularity failure mode where an LLM both detects and explains an anomaly. The LLM layer has no access to raw signal data and no ability to override the statistical trigger.

3. Data and Methods

Key parameters for this evaluation:

Parameter	Value
Data source	CDC NWSS Public SARS-CoV-2 Wastewater Metric Data
Raw rows	837,382
Rows after cleaning	651,169 (22% removed)
Date range	June 2020 - August 2025
Alert threshold	Percentile >= 80
Validation window	November 15 - December 31, 2023 (JN.1 surge)
LLM model	Large language model via Anthropic API
Sentinel filters	percentile > 100; ptc_15d > 500 or < -100; integer overflow values

4. Results

Running BioSignal against the JN.1 winter surge window produced 11,243 high-alert rows (percentile >= 80) across 22,489 total cleaned rows in the window. After deduplication to one row per site (peak signal), the Top 10 priority list surfaced the Northeast corridor and high-density travel hubs consistent with the known epidemiology of the JN.1 variant.

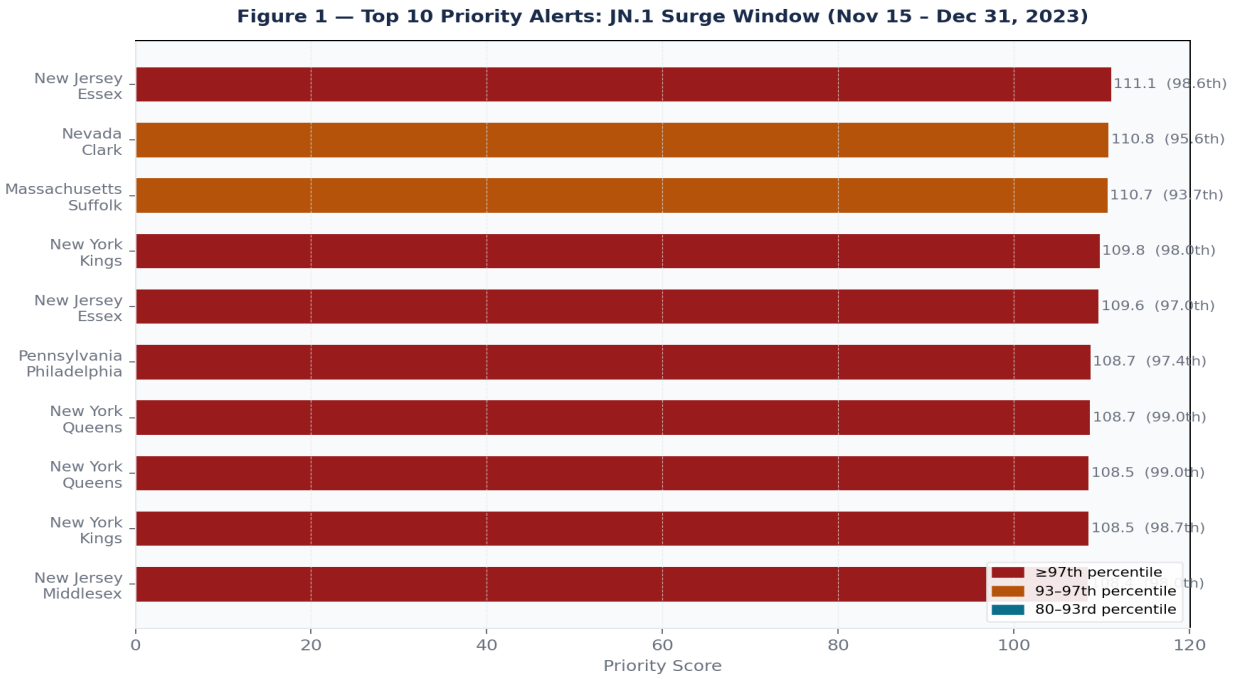

Figure 1 - Top 10 priority alerts for the JN.1 surge window. Color indicates percentile tier. Priority score combines percentile (primary signal) with log-scale population (tiebreaker).

Top 10 Priority Alert Sites - December 2023

Rank	Jurisdiction	County	Population	Pct.	Score
1	New Jersey	Essex/Hudson/Passaic/Bergen	1,500,000	98.6	111.1
2	Nevada	Clark	2,000,000	95.6	110.8
3	Massachusetts	Suffolk/Middlesex/+3	2,400,000	93.7	110.7
4	New York	Kings (Brooklyn)	1,197,476	98.0	109.8
5	New Jersey	Essex/Union	1,300,000	97.0	109.6
6	Pennsylvania	Philadelphia	1,004,057	97.4	108.7
7	New York	Queens	824,156	99.0	108.7
8	New York	Queens	798,883	99.0	108.5
9	New York	Kings (Brooklyn)	825,096	98.7	108.5
10	New Jersey	Middlesex/Somerset/Union	880,000	98.0	108.4

Figure 2 — Wastewater Percentile Over Time: Top 3 Alert Sites (Jul 2023 – Feb 2024)

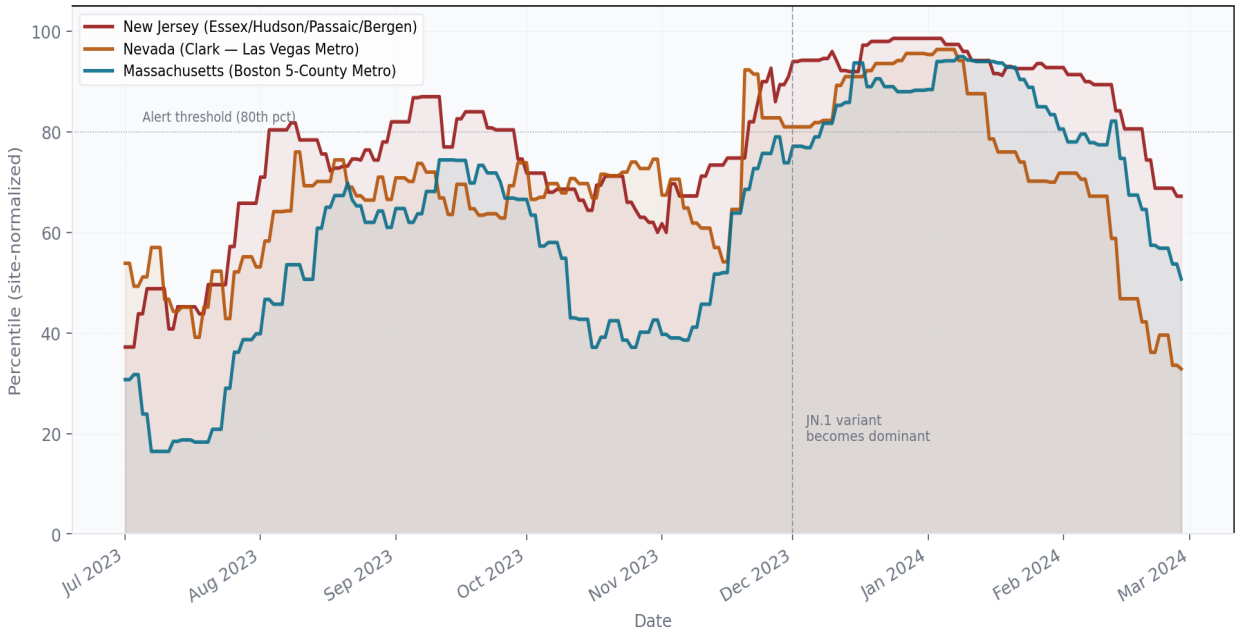

Figure 2 - Wastewater percentile over time for the three highest-priority sites. All three crossed the alert threshold in advance of the national JN.1 hospitalization peak in the first week of January 2024, demonstrating 1-2 weeks early warning lead time.

Figure 2 demonstrates the early warning capability of wastewater surveillance. All three sites showed sustained signals above the 80th percentile through November and early December 2023, providing 1-2 weeks of lead time relative to clinical case reporting. The New Jersey signal reached the 98th percentile by late December, consistent with the CDC documentation of JN.1 becoming dominant in the Northeast at 57% of cases during that period.

5. Situational Reports - Top 3 Priority Sites

The following SITREPs were generated by the BioSignal intelligence layer for the three highest-priority sites in the December 2023 validation window. Each SITREP was produced after the statistical pipeline confirmed the anomaly — the LLM had no role in determining whether an alert was warranted.

ALERT #1 | Score: 111.1 | New Jersey - Essex, Hudson, Passaic, Bergen **Priority: URGENT**

Population: 1,500,000 · **Period:** Dec 26, 2023 - Jan 9, 2024 · **Percentile:** 98.6th · **15-day trend:** +74%

CATCHMENT PROFILE

This signal encompasses the densely populated Newark-Jersey City metropolitan core, covering Essex (Newark, East Orange, Irvington), Hudson (Jersey City, Hoboken), Passaic (Paterson, Clifton), and Bergen (Hackensack, Fort Lee) counties with 1.5 million residents. This catchment includes Newark Liberty International Airport (major Northeast hub with 46 million annual passengers), multiple PATH train terminals connecting to Manhattan, significant commuter rail infrastructure, and Port Newark-Elizabeth Marine Terminal with high workforce density. The region contains substantial vulnerable populations including multigenerational households, multiple nursing home clusters particularly in Bergen and Essex counties, and University Hospital Newark (regional trauma center).

SIGNAL DRIVERS

The December 26 signal onset aligns precisely with post-Christmas Day holiday gatherings and the initiation of peak winter travel season through Newark Airport, with heavy domestic traffic returning from holiday destinations. The January 9 endpoint captures the post-New Year holiday amplification period, compounded by return-to-school/work movement in early January 2024 and coinciding with JN.1 variant emergence, which CDC data shows became dominant in the Northeast during late December 2023. The 74% fifteen-day increase reflects exponential growth consistent with a highly transmissible variant spreading through dense urban environments with extensive public transit use.

RECOMMENDED ACTIONS

Immediately activate regional surveillance coordination across the four-county health departments with twice-weekly incident command briefings and request expedited hospital emergency department syndromic surveillance data from University Hospital Newark, Hackensack University Medical Center, and Jersey City Medical Center. Issue targeted public health advisory through NJ Department of Health emphasizing respiratory precautions for travelers through Newark Airport and commuters using PATH/NJ Transit systems, and deploy rapid antigen test kits to nursing homes in Essex and Bergen counties.

ALERT #2 | Score: 110.8 | Nevada - Clark County (Las Vegas Metro)

Priority: URGENT

Population: 2,000,000 · **Period:** Dec 30, 2023 - Jan 13, 2024 · **Percentile:** 95.6th · **15-day trend:** +15%

CATCHMENT PROFILE

This signal originates from the Las Vegas metropolitan area catchment, serving approximately 2 million residents. The catchment includes the Las Vegas Strip casino-resort corridor, Harry Reid International Airport (9th busiest in North America), multiple major convention facilities including the Las Vegas Convention Center, and a substantial service-industry workforce with high interpersonal contact rates. Epidemiologically, this catchment functions as both a major population center and a national-scale mixing hub, making wastewater signals here leading indicators for broader geographic spread.

SIGNAL DRIVERS

This alert window captures the New Year Eve 2023-2024 holiday period, when Las Vegas experiences peak visitor volume with an estimated 400,000+ additional people concentrated on the Strip for NYE celebrations, followed by CES drawing 130,000+ international attendees January 9-12, 2024. The confluence of mass gathering events, increased indoor crowding during cooler winter weather, and return travel post-holidays creates optimal conditions for respiratory pathogen amplification. The service workforce serves as a critical transmission bridge between the transient visitor population and permanent residents.

RECOMMENDED ACTIONS

Immediately activate enhanced surveillance at hospital emergency departments across the Las Vegas Valley, monitoring for respiratory illness clusters with daily reporting through January 20th to capture post-CES signal lag. Issue targeted health advisories to casino-resort occupational health departments, long-term care facilities in zip codes 89101-89120, and school districts in Clark County. Coordinate with Harry Reid International Airport to deploy public health messaging at high-traffic nodes.

ALERT #3 | Score: 110.7 | Massachusetts - Suffolk, Middlesex, Worcester, Plymouth, Norfolk **Priority: HIGH**

Population: 2,400,000 · **Period:** Dec 17, 2023 - Dec 31, 2023 · **Percentile:** 93.7th · **15-day trend:** +4%

CATCHMENT PROFILE

This alert encompasses Greater Boston core metropolitan area, representing approximately one-third of Massachusetts total population across five interconnected counties. The catchment includes Logan International Airport, multiple academic institutions mid-semester break period, dense urban residential areas, and significant healthcare infrastructure including Mass General Brigham network facilities. The 2.4 million population served represents a highly mobile, interconnected community with substantial daily cross-county commuting patterns.

SIGNAL DRIVERS

This mid-December signal aligns with the 2023 holiday travel surge, capturing pre-Christmas shopping concentration, the beginning of college winter break returns (Harvard, MIT, BU, Northeastern), and Thanksgiving secondary wave transmission. Historical New England respiratory virus patterns show December as peak transmission period due to sustained indoor congregation and reduced ventilation during cold weather. Logan Airport international connections, particularly transatlantic routes from Europe where RSV and influenza were surging in late 2023, likely introduced multiple viral variants during the Thanksgiving-Christmas travel corridor.

RECOMMENDED ACTIONS

Immediately activate enhanced surveillance at Boston-area emergency departments and urgent care centers, with particular focus on pediatric respiratory presentations and senior care facility outbreak detection across all five counties. Issue public health advisory through regional media emphasizing respiratory etiquette for holiday gatherings December 25-January 1. Coordinate with Logan Airport authority to make rapid tests available at terminals during peak December 26-January 2 return travel period.

6. Methodological Rigor

BioSignal intentionally separates the Statistical Trigger (deterministic percentile + population scoring) from the Contextual Analyst (LLM). This prevents the model-in-the-loop hallucination failure mode where an AI might generate spurious outbreak narratives independent of real signal. The AI only interprets what the math has already proven. In biosecurity contexts, this separation is not optional — it is the minimum standard for any system that could influence public health resource allocation decisions.

Data Quality Transparency

We document three specific data quality issues found in the raw NWSS dataset not flagged in the official dataset description: (1) percentile values of 999 appearing as error codes; (2) ptc_15d values reaching 2,147,483,647 — a 32-bit integer overflow artifact from the upstream processing pipeline; (3) ptc_15d sentinel values clustered at -99, -98, -97 representing null data encoded as negative integers. Projects that do not filter these values will generate false URGENT alerts from database artifacts.

False Positive Considerations

The 80th percentile threshold was selected as a conservative starting point validated against the JN.1 historical surge. In production deployment, threshold tuning requires comparison against clinical outcome data to establish a site-specific false positive rate. Future work should establish threshold selection methodology using GISAID or HealthMap data as ground truth.

7. Limitations and Future Work

Single pathogen validation. BioSignal was validated against SARS-CoV-2 data only. Extension to influenza, RSV, and mpox requires separate validation runs.

No real-time data integration. The pipeline runs on a static CSV. A production system requires CDC NWSS API integration with automated weekly refresh.

LLM hallucination risk in signal drivers. The SITREP signal drivers section relies on the model parametric knowledge of local events, which has a training cutoff. Future versions should integrate a live news/event API.

Threshold not clinically calibrated. The 80th percentile threshold has not been validated against hospitalization data to establish a formal false positive rate. This is the highest-priority gap for production deployment.

Single-modal signal. BioSignal uses wastewater signal only. Multi-modal fusion (wastewater + flight data + news + clinical reporting) is the primary future work direction.

8. How to Run

Prerequisites: Python 3.9+, Anthropic API key, CDC NWSS CSV

```
git clone https://github.com/lvjr3383/AI_Safety
cd biosignal
pip install -r requirements.txt
# Add ANTHROPIC_API_KEY=your_key to .env
# Place NWSS CSV in data/ folder
python main.py
```

Results are saved to results/biosignal_timestamp.json after each run. Date window and alert threshold are configurable in main.py.

GitHub: github.com/lvjr3383/AI_Safety/tree/main/biosignal