Multi Stream Trajectory Scoring Pandemic Detection System

Alex Cheng, Dario Panerati

Apart Research AIxBio Hackathon, April 2026

Abstract

Current epidemic early warning systems (EWS) detect outbreaks when surveillance signals cross magnitude thresholds, by which time the epidemic is already underway. We present Symphonon, a trajectory-based EWS that scores each surveillance signal by the consistency of its directional change rather than its absolute level. The system fuses heterogeneous signals (wastewater, clinical cases, deaths, excess mortality, and syndromic surveillance) through per-signal independent scoring with asymmetric temporal parameters, runtime quality weighting, and magnitude-trajectory alerting gated by an adaptive threshold.

We validate the system retrospectively on two pathogens without structural recalibration: COVID-19 in the United States (321 weeks, 2019–2026; NWSS + OWID) and seasonal influenza (914 weeks, 2008–2026; ILINet + WHO FluNet). On influenza, the fusion channel issues alerts 4–6 weeks before a z-score baseline on gradual-onset seasons (H3N2 2014–15, H3N2 2017–18), with inter-wave false positive rate (FPR) 12.3% vs 14.4% for the baseline. On COVID-19, fusion FPR is 5.6% vs 9.1% baseline (a 38% relative reduction).

The method underperforms on abrupt-onset waves (BA.5, –3 weeks vs baseline) and on novel spillover events without pre-transition drift (H1N1 2009). Two additional channels (rigidity from regional decoupling, critical slowing down from AR(1)+variance) were implemented and did not add value in the tested configuration; failure modes are documented. The contribution is an engineering integration, not a theoretical novelty: existing components (trajectory scoring, quality weighting, adaptive thresholds) are combined and applied to multi-signal public-health surveillance with cross-pathogen transfer demonstrated.

1. Introduction

At the syndromic surveillance layer pandemic early detection one of the most widespread deployed some variation of z-scores. Once a z-score threshold is passed for some signal an alert is triggered to notify relevant personnel. The existing method works best under ideal circumstances of normal distribution, and constant signal mean and variance. These ideals rarely occur in the context of a pandemic.

Signals such as confirmed deaths or waste water detection rarely follow normal distributions. Surveillance count data is typically Poisson during early spread. Epidemic signals are not stationary. They have seasonal cycles, long-term trends, and structural breaks. The means and standard deviations shift so the baseline shifts as well, but the threshold does not follow.

This paper posits an alternative method using trajectory scoring with an adaptive threshold. It uses non-parametric detection and attempts to adjust for changes in baseline using a trailing window for mean and standard deviation.

Our main contributions are:

1. A pandemic early warning system using multi stream trajectory scoring for detection of score growth
2. Evaluation of this system on real historical data
3. Comparison to hypothetical univariate alternative

2. Related Work

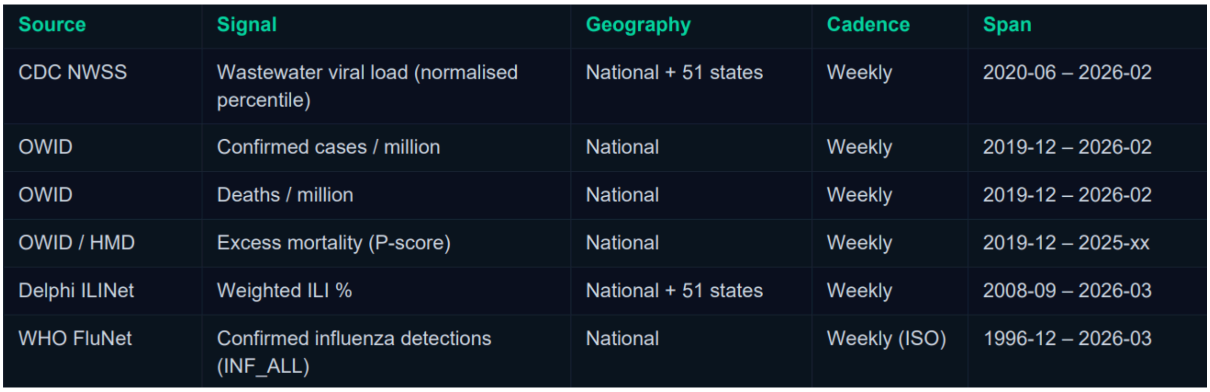
Multi Stream Surveillance. Decision fusion methods (Texier et al. 2014) fuses output of multiple different algorithms (CUSUM, EWMA, EARS, Farrington) on the same data. This paper is an engineering endeavor that uses a single algorithm that fuses trajectory scores across different data streams (deaths, waste water, etc).

Mann-Kendall. Mann-Kendall (Zheng et al. 2023) counts concordant minus discordant pairs across the entire window - every earlier point compared to every later point. This is used extensively for retrospective testing in epidemiology, but it has never been applied for live testing. What this paper proposes is mathematically similar, but modified to work over a trailing window so it can be deployed live.

3. Methods

Symphonon takes in raw signals from different sources. Trajectory scoring is run on each source individually first before being combined into a weighted fusion score in order to account for differences in signal lag.

Table 3.0.1



Source	Signal	Geography	Cadence	Span
CDC NWSS	Wastewater viral load (normalised percentile)	National + 51 states	Weekly	2020-06 – 2026-02
OWID	Confirmed cases / million	National	Weekly	2019-12 – 2026-02
OWID	Deaths / million	National	Weekly	2019-12 – 2026-02
OWID / HMD	Excess mortality (P-score)	National	Weekly	2019-12 – 2025-xx
Delphi ILINet	Weighted ILI %	National + 51 states	Weekly	2008-09 – 2026-03
WHO FluNet	Confirmed influenza detections (INF_ALL)	National	Weekly (ISO)	1996-12 – 2026-03

Symphonon is back tested on Covid and Flu data, and compared to z-score back testing.

3.1. Normalisation

Each raw signal gets normalised using a **rolling 104-week trailing 95th percentile**:

$$\text{Normalised}(t) = \text{raw}(t) / \text{P95}(\ \text{raw}[t-103 : t]\ )$$

The window only looks backward so no future data leaks in, this is for backtesting purposes to simulate real life conditions.

Normalisation forces signals into comparable scales. 95th percentile was selected rather than max to avoid anchoring to a single extreme peak.

A trailing window avoids failure modes: rolling-max normalisation anchors everything to 1.0 post-peak, and expanding-window normalisation permanently compresses signals after a big event like Omicron.

3.2. Per-Signal Trajectory Scoring

For each signal independently compute:

> Score(t) = (#{weeks w in [t-w_r+1, t] : signal(w) > signal(w - w□)}) / w_r

For each of those weeks, ask whether the signal is higher than it was w□ weeks before that. Count the fraction of weeks over w_r weeks where the answer is yes. Score ∈ [0, 1].

A score near 1.0 means the signal beat its own value from w□ weeks ago in almost every one of the last w_r weeks.

w_r and w□ are hardcoded values.

Table 3.2.1

Signal	w_L	w_R	α	w_S	Rationale
NWSS wastewater	4	4	0.45	0.45	Leading indicator; short windows for reactivity
Cases / million	6	6	0.55	0.30	Mid-lag; moderate smoothing
Deaths / million	10	8	0.70	0.15	Lagging; stability prioritised over reactivity
Excess mortality	12	10	0.75	0.10	Most lagging; confirmation-like signal
ILI (ILINet)	3	3	0.35	0.60	Leading syndromic indicator
FluNet confirmed	4	4	0.45	0.40	Mid-lag laboratory confirmation

3.3. Quality Weighting

Each signal also carries a quality multiplier computed as:

> quality(t) = fraction of real (non-interpolated or non-imputed) observations in a trailing 12-week window
>
> base_weight(i) are fixed weights used to account for lag.

The effective fusion weight for signal i becomes:

> w_eff(i, t) = base_weight(i) × quality(i, t)

Then weights are re-normalised to sum to 1 before fusion. If quality falls below a minimum threshold, the entire alert is suppressed.

3.4. Score Fusion

Scores are fused into a single composite:

$$\text{fusion\_score}(t) = \Sigma_i \left[\, \text{w\_eff}(i,t) \times \text{Score}(i,t) \,\right] / \Sigma_i\, \text{w\_eff}(i,t)$$

Scores are fused rather than relying on raw signals to account for lagging signals (deaths, excess mortality) which have long $w_{□}/w_r$ windows in the score structure, so they no longer drag down the early-warning contribution from leading signals like wastewater. Each signal has been "translated" into a common [0,1] trajectory language before fusion in 3.2.

3.5. Adaptive Alert Threshold

The alert threshold is not fixed - it adapts to recent baseline behaviour:

$$\text{threshold}(t) = \max(\, \text{floor}, \text{mean}(\text{fusion\_score}[t-25:t]) + k \times \text{std}(\text{fusion\_score}[t-25:t]) \,)$$

- 26-week trailing window (roughly one semester of recent history)
- k = 2.0 for COVID, k = 1.5 for flu (flu has higher seasonal variance so a tighter multiplier is used)
- The floor prevents the threshold from collapsing to near-zero during quiet periods

Fusion alert fires when:

- fusion_score(t) > threshold(t) for 3 consecutive weeks
- quality ≥ 0.50

The 3-week persistence requirement is what crushes the false positive rate - a single noisy spike doesn't fire.

The comparison baseline is a rolling z-score on the primary signal (NWSS wastewater for COVID-19, ILI for influenza): at each week, $z(t) = (x(t) - \text{mean\_8w}) / \text{std\_8w}$; alert fires when $z(t) > 2.5$.

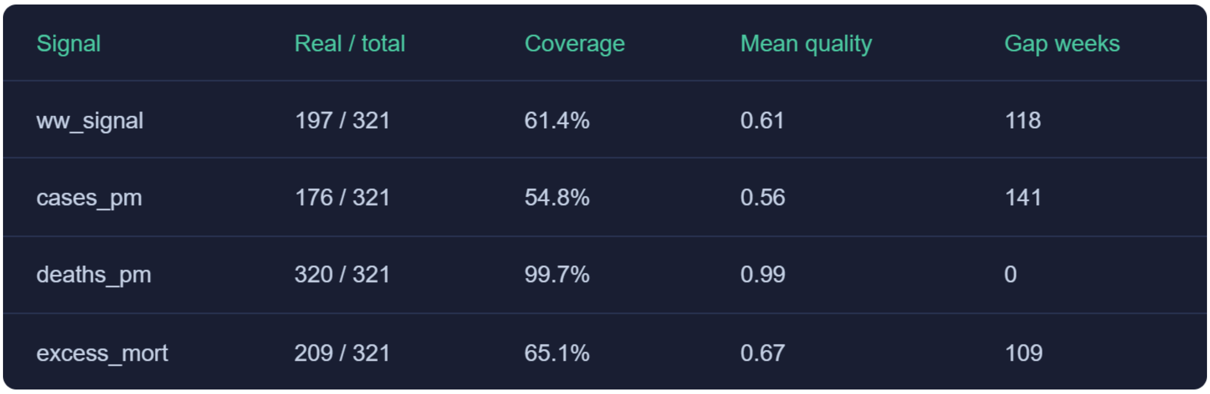
3.6 Validation protocol

For each known wave w with onset date t w , we search a window [t w − 8, t w + 18] weeks for the first alert from each channel. Onsets for COVID-19 are taken from CDC variant tracking archives (Alpha, Delta, Omicron, BA.5, XBB). Onsets for influenza are defined as the CDC's own FluView standard: the first week ILI % exceeds the national baseline for ≥ 2 consecutive weeks. False positive rate is computed on inter-wave periods only, excluding a ±8-week buffer around each onset and a 52-week burn-in at the start of the series. This yields 197 inter-wave weeks for COVID-19 and 763 for influenza.

4. Results

4.1 Covid

Table 4.1.1



Signal	Real / total	Coverage	Mean quality	Gap weeks
ww_signal	197 / 321	61.4%	0.61	118
cases_pm	176 / 321	54.8%	0.56	141
deaths_pm	320 / 321	99.7%	0.99	0
excess_mort	209 / 321	65.1%	0.67	109

Sources mentioned in figure 3.0.1. Covid uses CDC NWSS, OWID, and OWID / HMD. Where CDC NWSS includes waste water viral load. OWID for confirmed cases/million and deaths/million. OWID/HMD for excess mortality.

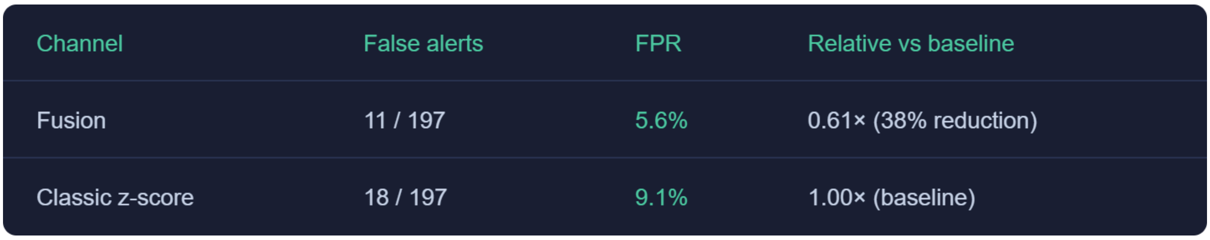
NWSS reached meaningful national coverage only in 2022. The runtime quality layer (§3.5) automatically reduces wastewater contribution to the fusion score during the sparse-coverage period.

Wave	Onset	Fusion	Classic (z)	Adv. fusion
Alpha	2021-03-01	—	—	n/a — coverage
Delta	2021-07-05	—	2021-07-12	n/a — coverage
Omicron	2021-12-06	—	2021-12-06	n/a — coverage
BA.5	2022-06-27	2022-05-23	2022-05-02	–3 weeks
XBB	2022-12-05	—	2022-12-05	n/a

Alpha was first detected in 2019-2020. The onset date given above uses CDC variant tracking archives for America in particular. 2021-03-01 was the first date recognized by the CDC for Covid in America. Neither traditional z-score nor trajectory scoring is capable of detecting novel diseases without a baseline signal.

Delta and Omicron were rightfully suppressed due to sparse coverage of waste water data. XBB did not fire. We interpret this as either a byproduct of the signal being contaminated by earlier BA.5, or caused by the signal cases/million being degraded in the latter half of the pandemic.

The single wave detected by both channels is BA.5. Fusion fires 3 weeks after the classic detector. This is an expected result. BA.5 rose abruptly over a short period, which is the profile where a magnitude spike (z-score) has a structural advantage over trajectory scoring. Trajectory detection requires sustained directional movement across multiple consecutive weeks; a fast-rising signal is already high before that pattern can accumulate.

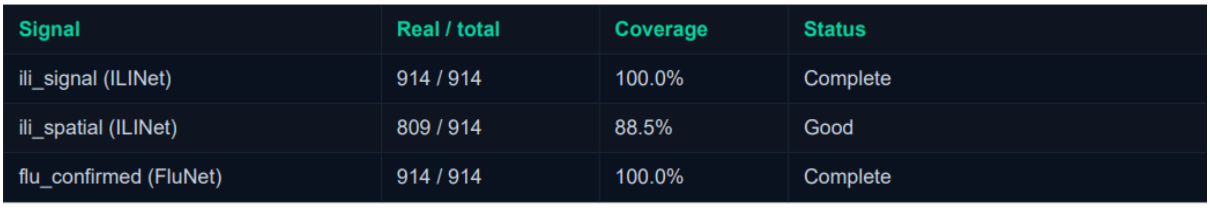


Channel	False alerts	FPR	Relative vs baseline
Fusion	11 / 197	5.6%	0.61× (38% reduction)
Classic z-score	18 / 197	9.1%	1.00× (baseline)

False positive rates are taken by looking at all alerts that do not correspond to an observed pandemic wave within the validation period. The results are not analytical but rather empirical. Fusion has substantially lower false positive rates.

4.2 Flu

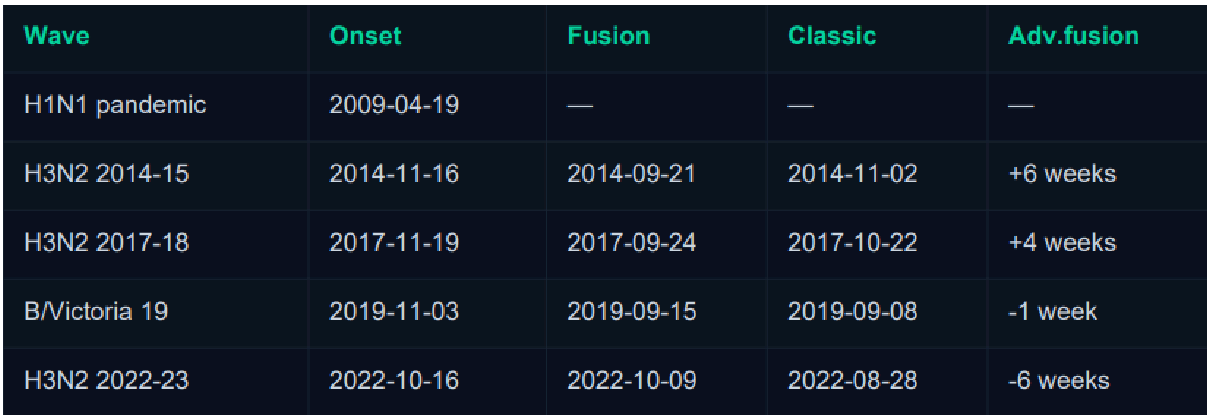
Table 4.2.1



Signal	Real / total	Coverage	Status
ili_signal (ILINet)	914 / 914	100.0%	Complete
ili_spatial (ILINet)	809 / 914	88.5%	Good
flu_confirmed (FluNet)	914 / 914	100.0%	Complete

Substantially higher data coverage compared to Covid.

Table 4.2.2



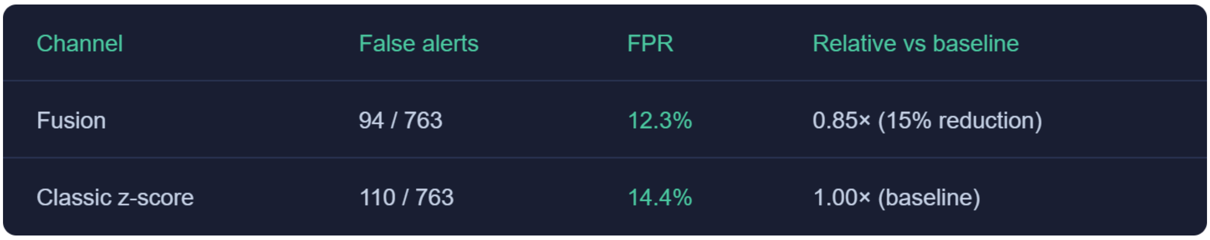
Wave	Onset	Fusion	Classic	Adv.fusion
H1N1 pandemic	2009-04-19	—	—	—
H3N2 2014-15	2014-11-16	2014-09-21	2014-11-02	+6 weeks
H3N2 2017-18	2017-11-19	2017-09-24	2017-10-22	+4 weeks
B/Victoria 19	2019-11-03	2019-09-15	2019-09-08	-1 week
H3N2 2022-23	2022-10-16	2022-10-09	2022-08-28	-6 weeks

H1N1 is the first onset of the Flu.

On H3N2 2014-15 and H3N2 2017-18, the fusion channel issues alerts 6 and 4 weeks respectively before the ILI baseline is crossed - i.e. while clinical burden remains below the CDC's own detection threshold. These are the scenarios where trajectory detection delivers its gradual, consistent directional movement becomes visible before any single week exceeds historical levels.

B/Victoria 19 fusion fired one week late relative to onset compared to the classic detector. We cannot clearly determine if the result represents the detector failing, as the result falls within an expected margin of error.

We speculate fusion failed to fire for H3N2 2022-23 because Flu spread rebounded after Covid lockdown was lifted. This caused the trailing threshold to be abnormally high when H3N2 hit, causing the fusion to fire late.



Channel	False alerts	FPR	Relative vs baseline
Fusion	94 / 763	12.3%	0.85× (15% reduction)
Classic z-score	110 / 763	14.4%	1.00× (baseline)

5. Discussion and Limitations

Limitations

1. **Scope**. The method detects epidemic onsets characterised by a gradual pre-transition phase of consistent directional movement. It is structurally unsuited to novel pathogen spillover events with abrupt emergence (H1N1 2009 being the paradigmatic case in our validation).
2. **Retrospective validation**. All results reported here are retrospective. A true prospective test requires locking the system on historical data and running it forward on future waves as they occur.
3. **Single Country**. Validation is restricted to the United States because the highest-quality multi-signal surveillance data is available there.
4. **Autocorrelation**. Symphonon does not account for autocorrelation of information sources.

5. **Operational Ground-truth**. w_r and w□ are guestimations. The ground truth in epidemiology are reports, but that is not actual beginning of the pandemic.

6. **Baseline comparison**. Our z-score baseline is a widespread but simple aberration detector. We have not compared against more sophisticated surveillance systems (e.g. Farrington-Noufaily, BCP, ensemble nowcasts).

7. **Engineering, not Theoretical**. The individual components (trajectory scoring, quality weighting, adaptive thresholds, multi-channel alerting) are not individually novel. The contribution is the specific integration and its demonstrated cross-pathogen transfer on public data.

8. **Limited Data**. Lack of test data. Waste water detection came online mid-Covid, so the testing is incomplete.

Future Work

Some of these limitations such as 5 is caused by hardware or testing methods, which we do not touch upon in this paper. The lowest hanging fruit for would probably be something closer to adjusting the algorithm for autocorrelation, and comparing trajectory scoring to ensemble models being deployed right now by CDC (CDC, 2026).

6. Conclusion

The results appear to be somewhat inconclusive. Symphonon does not fire for the first few waves of Covid because of incomplete data. It either fails to fire or fires late for BA.5 and XBB respectively. These results imply that trajectory scoring does not work as well as z-score or alternatively better data could assist with fidelity.

Flu testing does not face the same data availability problem. Mixed results suggest the method does not work well, or that more fine tuning may be needed.

References

1. Centers for Disease Control and Prevention. (2026, April 17). *Forecasts of flu hospital admissions*. FluSight. https://www.cdc.gov/flu-forecasting/data-vis/current-week.html

Code and Data

- ***Code repository:*** https://github.com/Brukino/Symphonon/tree/main/symphonon_repo
- ***Data/Datasets:***
  data.cdc.gov/api/views/2ew6-ywp6/rows.csv?accessType=DOWNLOAD
  covid.ourworldindata.org/data/owid-covid-data.csv
  api.delphi.cmu.edu/epidata/fluview
  FluNet is a manual download from the WHO site, no direct URL.

LLM Usage Statement

LLM was used extensively for the following: language translation, search for similar works, and create rough drafts.