Sentinel Atlas: A Centralized Platform for Multi-Source Epidemic Surveillance Data[1]

Chris Harig
Independent

YM Liaw
Independent

Vidur Kumar
Johns Hopkins University

Umar Zafar
Johns Hopkins University

With
Apart Research

Abstract

As AI capabilities grow, they may accelerate biological risks, making robust detection and early warning systems vital defenses. Currently, early warning systems for pandemics are hampered by data that is hard-to-find, fragmented, non-navigable, and isolated from the institutions responsible for high-stakes public health decision-making. To address this, we propose a comprehensive dataset collection and aggregation framework to facilitate crowdsourcing epidemiological forecasts. We aim to accelerate defenses by providing a platform for early detection of pandemics predicated on open access to data and global collaboration. Our system currently hosts 21 datasets spanning 200 countries and territories, along with companion prediction datasets. We find that the primary bottleneck in pandemic intelligence is not lack of data or models, but the lack of standardized, accessible infrastructure to integrate and use them effectively.

[1] Research conducted at the AIxBio Hackathon, April 2026

1. Introduction

Current pathogen surveillance data exists in high volume but remains functionally inaccessible due to siloed hosting platforms and inconsistent data schemas. Wastewater monitoring programs, clinical case reporting systems, and other signals such as news and search trends are maintained by separate institutions, hosted on different platforms, and reported in incompatible formats. This fragmentation forces researchers and forecasters to spend critical time manually harmonizing data from multiple sources before any analysis can begin, creating a significant barrier to real-time pandemic intelligence.

AI can accelerate threats (making it easier to create pathogens) but it can also accelerate defenses. As models improve, access to data, and platforms for finding, using and publishing predictions could be key infrastructure for accelerating defense. As models get better, more people can create predictive models and work to stop pandemics. Crowdsourcing infrastructure could play a key role in accelerating defense, but it is underdeveloped and leaves many frictions for developers and models.

We developed EPI-Eval and a dashboard for navigating data to reduce this friction. The platform provides a unified entry point for diverse public health surveillance datasets, with a consistent schema, lightweight visualization and predictive modeling tools.

Our main contributions are:

1. A catalog of 21 datasets hosted on Hugging Face under a consistent schema, covering clinical treatments, wastewater, and behavioral signals across multiple pathogens and geographies
2. An ingestion pipeline that normalizes heterogeneous source formats into a common structure, with live/stale status tagging to make data freshness explicit, open source for anyone to commit datasets and predictions with
3. A live crowdsourcing platform for aggregating relevant datasets and predictions of clinical counts (which we hope, could detect seasonal illness rates or pandemics, early, anywhere in the world)

2. Related Work

Predictive epidemiological models have existed as an important tool for health institutions, governments and citizens to use and learn from, and are improving in accuracy and breadth over time.

Forecasting competitions such as CDC FluSight and the COVID-19 Forecast Hub have demonstrated the benefits of providing data and aggregating predictions, but both are competition infrastructure with a narrow scope, focusing on a single country, season and illness. There are

various other public data sources not linked to competitions. Wastewater surveillance efforts like WastewaterSCAN track various pathogens, but at a country level. Wastewater SPHERE and COVID-19 wastewater dashboard catalog country-specific dashboards without integrating the underlying data for others to easily access and develop models with. Both form a proof of concept for something larger.

Our work is complementary: we provide the missing upstream data layer, consolidating diverse global streams, including wastewater, clinical, and mobility data into a single, schema-consistent repository. By standardizing these disparate sources, we enable concerted global surveillance, visibility into coverage blindspots, and more robust inputs for pandemic forecasting and preparedness.

3. Methods

Data collection and upload followed a source-by-source adapter approach. Each dataset was fetched from its original host (CDC, WHO, UKHSA, and others), processed through a source-specific ingestion script, and uploaded to the EPI-Eval organization on Hugging Face. The schema.yaml provides a minimal framework for all datasets to conform to. The target schema standardizes each record to: location, date, pathogen, metric type, value, unit, and source. Scripts and the schema are available in the *upload_pipeline* directory of the project repository.

After data collection, we designed an interactive dashboard that periodically (or manually if the user chooses) from the huggingface datasets and displays key data points. Graph and Map features allow for in depth exploration of large time series datasets key for developing epidemiological models. None of the data is stored locally on the user's device unless they want to download it for further inspection and development.

Furthermore, the dashboard supports personal data sets and prediction data upload. The user can choose to upload their data, compare it to existing data, and push to Huggingface. There is a short pull request form that asks the user to verify their association, and if the administrator approves it, the prediction is added to any given dataset's sister -predictions dataset. This dataset can then be pulled down like any other, and predictions can be compared with a standard set of metrics (MSE, RMSE, WIS, etc.), and visualized next to the ground truth data.

4. Results

21 datasets currently active, with live/ stale status visible across all sources, covering 200 countries and territories. 21 companion datasets are also included to store predictions.
Global datasets:

- COVID
- Monkey pox
- Mobility
- Tuberculosis
- World population

Country-specific:

- COVID
- Flu
- RSV
- Wastewater

A working crowd sourcing pipeline for pulling datasets down from hugging face and pushing datasets up to be added to the pool, and relevant tooling for exploring data sets and comparing predictions.

Site: epi-eval.com
Datasets: https://huggingface.co/EPI-Eval
Dashboard Code: https://github.com/ChrisHarig/apart-forecasting-tool

5. Discussion and Limitations

A key challenge encountered during development was the lack of consistent data platforms and formats across surveillance sources, particularly in wastewater reporting. Differences in spatial resolution, naming conventions, reporting unit and frequency required manual normalization before integration. Several countries' surveillance data are reported in terms of dashboard and PDF files that do not provide an option for data download thus hindering integration. Differing reporting units also made data interpretation difficult as different countries or regions may have different thresholds for pathogen levels.

This highlights a broader gap in current surveillance infrastructure. While within the US, efforts by CDC have standardized data platforms for different states and countries, organizations such as WHO could benefit from more standardized reporting schemas to improve interoperability across regions and data types.

Limitations

The platform detects only known, reported pathogens. A novel pathogen or stealth outbreak would not generate signal in any of the current data streams; this is an inherent limitation of surveillance systems dependent on identified pathogens and structured reporting. Finally, the dashboard supports tooling for making and publishing predictions but does not make predictions itself.

Future Work

Future work could explore defining common data standards or coordination mechanisms to enable seamless integration of heterogeneous surveillance signals at a global scale. More datasets from many sources could be included. Hosting models on Huggingface could give under-resourced organizations a chance to gain an edge on pandemic preparedness they otherwise wouldn't have access to.

Infrastructure to streamline competition and forecast models could also be explored.

6. Conclusion

We found dataset collection and prediction aggregation was the main bottleneck to creating useful epidemiological forecasts. We created a collection of 21 datasets for training predictive models for seasonal illness prevention and pandemic prevention. We also propose a crowdsourcing of data and predictions as an effective tool for developing better pandemic preparedness. We hope our tool can lead to democratization of predictive tooling and datasets, and therefore more high quality predictions for global preparedness.

Code and Data

- **Code repository**: https://github.com/ChrisHarig/apart-forecasting-tool
- **Data/Datasets**: https://huggingface.co/EPI-Eval
- **Other artifacts** : epi-eval.com

Author Contributions

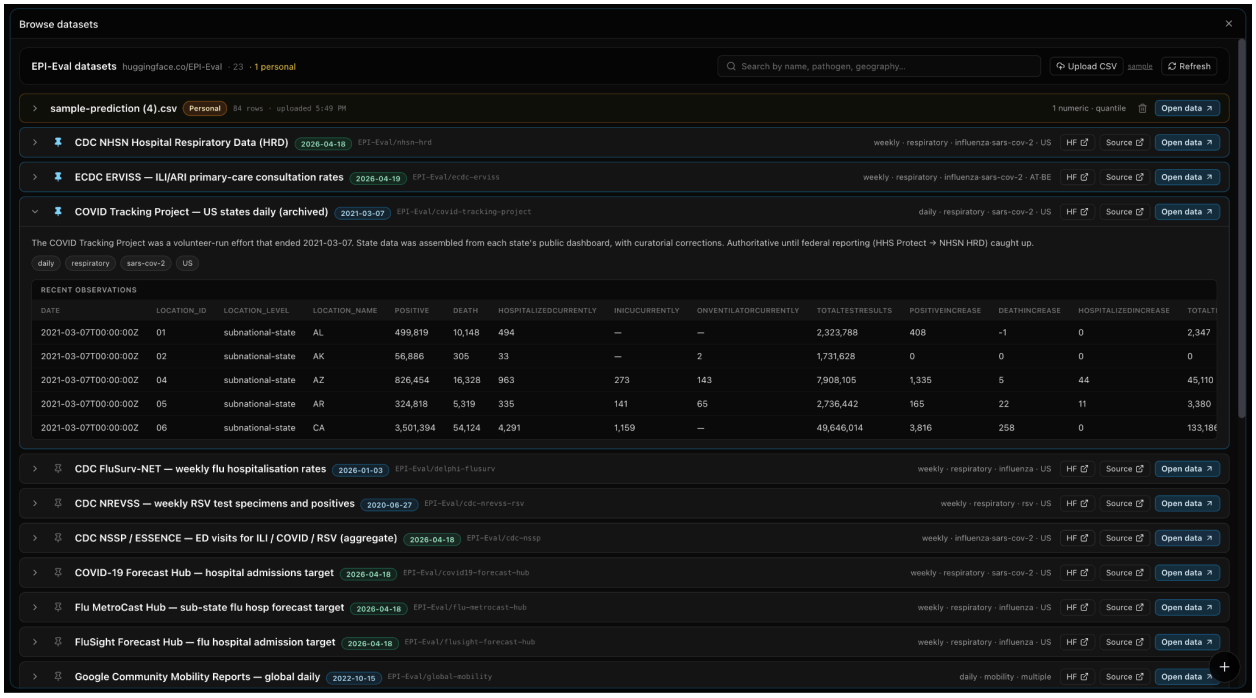
Chris Harig led the implementation of the dashboard, data formats, dataset collection, predictive modeling features, and assisted in the final report.
YM Liaw led wastewater and mobility data collection and the final report.
Vidur Kumar created the initial versions and ideas for the dashboard, mapping features, and predictive modeling features.
Umar Zafar worked on dataset collection and assisted on the dashboard.

References

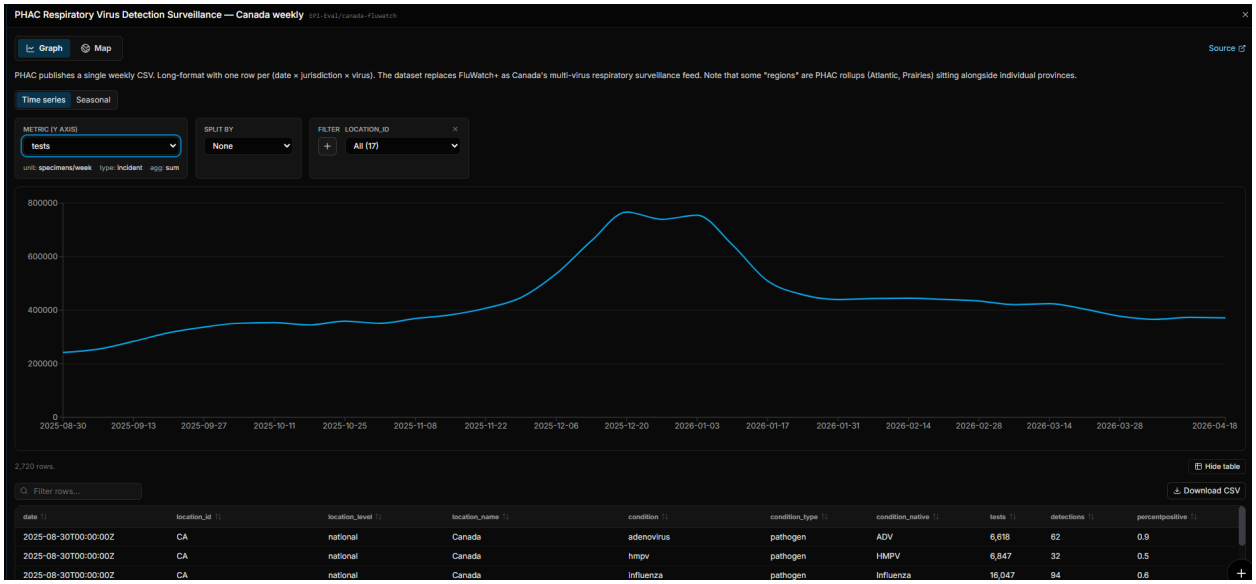
1. *CDC Flusight Competition: https://www.cdc.gov/flu-forecasting/evaluation/2024-2025-report.html*
2. *Wastewater SPHERE: Global Wastewater Dashboard Map*
3. *WHO Global wastewater monitoring sources for SARS-CoV-2 (COVID-19 virus): https://data.who.int/dashboards/covid19/wastewater*
4. *Dataset attributions can be found on Huggingface: https://huggingface.co/EPI-Eval*

Appendix A.

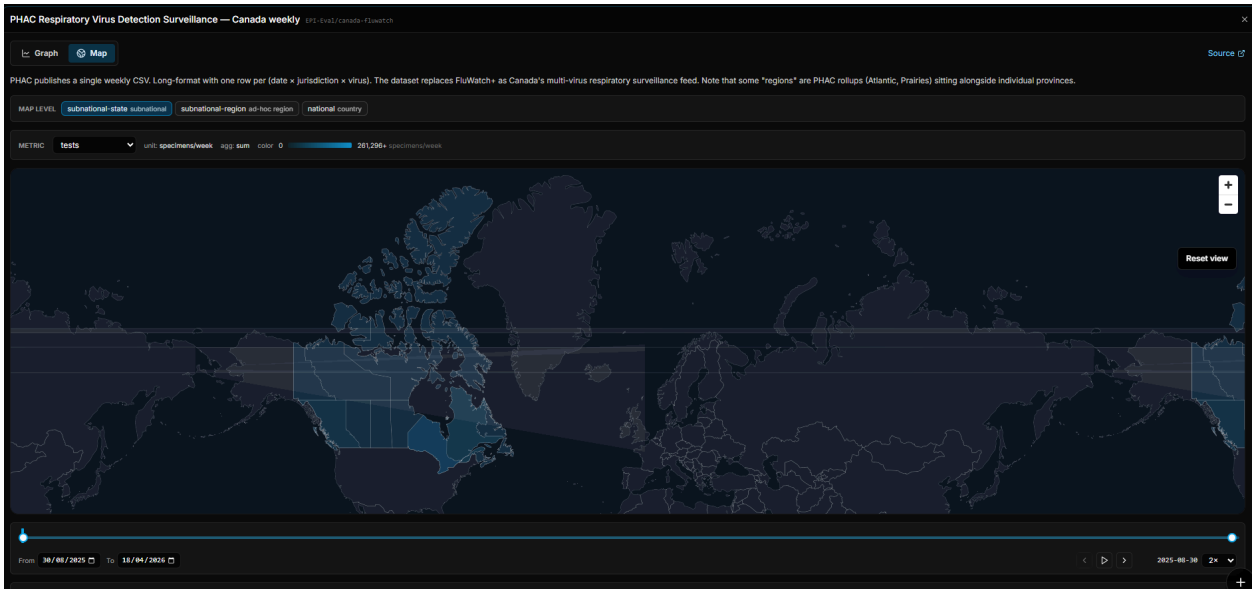
Screenshots of the dashboards main features (dashboard can be viewed at epi-eval.com)



Dashboard interface that shows available data sources (green: live, updated data; blue: historical data)



Integrated dashboard view displaying time-series analysis (above) alongside regional mapping (below). Note: The level of detail automatically adjusts to match the granularity of reported data



LLM Usage Statement

Claude and ChatGPT were used to build the back and frontend of this project, including the UI and ingestion scripts. LLM were also used in brainstorming and wording refinement.