An AI-Based Public Health Data Monitoring System

Ananya Joshi[1] , **Nolan Gormley**[1] , **Richa Gadgil**[1] , **Tina Townes**[1] , **Roni Rosenfeld**[1] , **Bryan Wilder**[1]

[1]Carnegie Mellon University
{aajoshi, ngormley}@andrew.cmu.edu

Abstract

Public health experts need scalable approaches to monitor large volumes of health data (e.g., cases, hospitalizations, deaths) for outbreaks or data quality issues. Traditional *alert-based* monitoring systems struggle with modern public health data monitoring systems for several reasons, including that alerting thresholds need to be constantly reset and the data volumes may cause application lag. Instead, we propose a *ranking-based* monitoring paradigm that leverages new AI anomaly detection methods. Through a multi-year interdisciplinary collaboration, the resulting system has been deployed at a national organization to monitor up to 5,000,000 data points daily. A three-month longitudinal deployed evaluation revealed a significant improvement in monitoring objectives, with a 54x increase in reviewer speed efficiency compared to traditional alert-based methods. This work highlights the potential of human-centered AI to transform public health decision-making.

1 Introduction

Public health data monitoring systems are critical for detecting outbreaks and ensuring data quality for downstream applications (Fig 1), but traditional alert-based systems struggle to detect these outlying data events with modern, high-volume, heterogeneous data streams Shmueli and Burkom [2010].

To improve timely and informative public health actions, data modernization efforts in the past decade have led to an exponential increase in data volume, complexity, and heterogeneity [WHO, 2022], especially at national-level data curation organizations like **Delphi**. Ironically, these investments in data modernization have been *incompatible* with existing public health monitoring systems, reducing the utility of this investment[1]. Specifically, most organizations used alerting-based monitoring systems – where the computational data processing methods often involved defining specific types of alerts for data patterns that matched data quality issues or outbreaks Dong *et al.* [2022]. However, these definitions could not adapt to the wide variety of public health data sources, fluctuating quality, and increasing volume in the modern data —- unlike the relatively small, static data pipelines of the past. After observing the failure modes using the alerting paradigm, as we describe in this paper, we, a group of researchers at Delphi, designed a novel monitoring system that addresses the core limitations of the prior system via a newly available AI-based ranking paradigm.

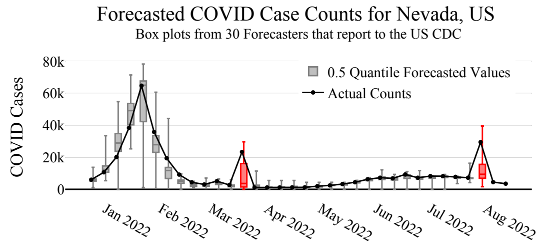


Figure 1: Data monitoring catches the data quality changes in case counts, shown by the large spikes when cases were trending down, which resulted in similar spikes for predicted counts (red) from multiple forecasts.

Goal: The 'gold standard' of a successful public health data monitoring system is when data reviewers 1) can quickly find as many high-priority events and 2) increase their effeciency.

Limitations of Current Alerting Paradigm Existing alerting-based monitoring systems, where data that meets a statistical definition produces an alert, are incompatible with modern data because they:

1. Are stagnant to data changes (Computational) : At scale, the alerting algorithm parameters needed to be manually adjusted across millions of data streams because of the fluctuating public health volume and measurement standards. This requirement for manual tuning negates the scalability benefits of automation.
2. Constrain situational awareness (Engineering): At scale, individual data points need to be inspected in the *context* of the larger public health data available [Burkom,

[1] Ethical Approval / Approving Organization: This study was IRB approved.

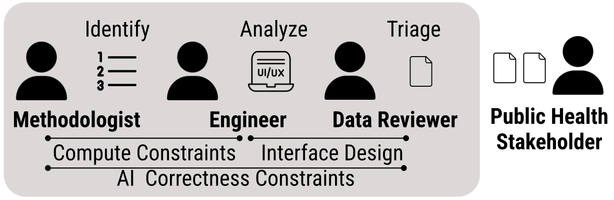


Figure 2: Data reviewers triage notable events from public health data for stakeholders. Our system is powered by an AI-ranking method that identifies relevant events and formalizes how different stakeholders interact directly or indirectly with the AI method.

2017]. However, the volume of alerts and constraints on relevant context (e.g. alerts from cases and deaths cannot be combined because cases usually lead deaths so the alerts may be unrelated) increase engineering complexity and introduce application latency.

3. **Disincentivize engagement** (Reviwer Actions): Most importantly, reviewers using these applications often had to sort through an overwhelming and unusable number of data alerts (e.g. 35,000+ alerts [Coletta and Zhou, 2019]). Without the context they need to engage with the system from 2., they cannot meet their minimum goal of correctly analyzing high priority events.

In response to these challenges, we, a research team at Delphi, designed a ranking-based AI system that leverages the strengths and constraints of multiple public health stakeholders, that has been used for the past two years. In the process, we conducted a multi-year collaboration involving public health data reviewers, engineers, and computer scientists, identified aspects of the process that were well suited for an AI approach, and developed an approach to evaluate it in practice. In the following sections, we describe the process of developing this system via:

- **Sec 4:** Modernizing Data Monitoring with AI: opportunities and constraints on automation.
- **Sec 5:** Goal-oriented system design guidelines
- **Sec 6,7:** Measurement, evaluation, and deployment strategies for empirical validation

The final system (Fig. 2) centers improved human interactions through an unsupervised, ranking-based, AI paradigm. To test the impact of this approach on the system over time, we documented the performance of this human-AI collaboration over 3 months, where it was subject to the fluctuations and changes typical in public health data. During this time, we also added interaction modalities for stakeholders that interact with the AI outputs to test and support the data monitoring system. We *preregistered* the Github commits and evaluation strategy on OSF before experiments began. Our results show that the ranking paradigm outperformed the alerting paradigm on core qualitative and quantitative metrics designed by reviewers that demonstrate the improved 1) effectiveness of the ranking paradigm in practice, 2) efficiency of reviewers using the system, and 3) overall experience with the system, which support data monitoring in the modern setting.

2 Related Work

The challenges encountered in public health data monitoring closely resemble those in other critical domains requiring large-scale data monitoring (e.g., environmental, financial, and agricultural sectors), especially because these fields depend on domain expertise for reasons such as liability and contextual interpretation. Numerous systems exist to help these types of experts understand relationships within large volumes of temporal data [Kesavan *et al.*, 2020]. However, most of these systems focus on surfacing aggregate trends rather than monitoring individual data points in the *context* of these trends. As modern data streams become more complex, these types of general-purpose data analysis systems become inadequate for the following reasons:

Stagnant to Data Changes: Alerting systems at scale suffer from well-documented issues, including statistical inaccuracies [Joshi *et al.*, 2024; Shmueli and Burkom, 2010] and overwhelming alert volumes [Hurt-Mullen and Coberly, 2005]. Even the most sophisticated adaptive mechanisms [Burkom, 2017] require infeasible, constant retuning to provide a reasonable set of high quality alerts due to the fluctuating quality of data.

Missing Situational Awareness: Some systems fuse data across visual interfaces [Deng *et al.*, 2023; Cao *et al.*, 2017; Vajiac *et al.*, 2022]. While generally effective in typical HCI and visualization settings, these approaches are incompatible with public health data due to its shifting lead times across indicators and complex correlation structures [Lakdawala and Joshi, 2023]. Static correlation-based techniques often fail in this context [Fan *et al.*, 2014; Hilda *et al.*, 2016].

Additionally, the nonstationary and hierarchical nature of public health data [Reinhart *et al.*, 2021] introduces constrained relationships that must be accounted for to prevent spurious correlations. This failure mode is very likely in public health given that data quality varies across univariate data streams, within indicators across regions, and between different indicators for the same region. This variability also complicates **directly** comparing across data dimensions, impacting which limited subset of events are detected by these methods. Consequently, many existing methods that are highly parameter-sensitive, computationally expensive [Teng *et al.*, 2017; Kesavan *et al.*, 2020], or require high visual literacy cannot work in these setting. These limitations are prominent even in more recent ranking-based approaches [Ding *et al.*, 2023; Montambault *et al.*, 2022], which often treat all data dimensions similarly—an inappropriate assumption for public health data, where temporal and geospatial dimensions are far more important than other dimensions.

Disincentivizes Engagement: Systems that generate overwhelming numbers of alerts can lead to information overload, preventing thorough review. Exploratory filtering across dimensions and indicators (e.g., cases, deaths, hospitalizations) [Preim and Lawonn, 2020; Chen *et al.*, 2010b; Maciejewski *et al.*, 2009] requires users to navigate millions of filter combinations to inspect relevant data points. While batching alerts [for Disease Control and Prevention, 2023][2] can reduce alert

[2] For example, public health alerts can be aggregated over data provider, indicator, geography, time, revision history, or population.

volume, this approach risks improper aggregation. For example, alerts for cases and deaths cannot be combined due to their distinct lag structures and event correlations. Further, contemporary drill-down techniques demand extensive manual inspection [Hurt-Mullen and Coberly, 2005], further slowing the review process.

These challenges are exacerbated in modern public health monitoring. First, when combined with high-volume, low-quality data—characterized by nonstationarity, noise, missing values, and inconsistencies over long timespans – the aforementioned theoretical limitations of alerting methods are realized. Second, as existing systems using the alerting paradigm were *designed* to only be used on a local scale [Chen *et al.*, 2010a], anecdotally where **all** data could still be analyzed manually, they were never built to support situational awareness. As a result, the need for improved monitoring systems has been widely recognized by public health practitioners [Hopkins *et al.*, 2017], with limitations persisting across both historical [Chen *et al.*, 2010a] and modern public health monitoring systems.

A final challenge in designing monitoring systems is that academic research on disease monitoring systems has predominantly focused on improving methods for data identification in isolation [Buckeridge, 2007], including recent AI approaches. Unfortunately, this has not always translated into practical adoption [Shmueli and Burkom, 2010]. Instead, we adopt an interdisciplinary systems design approach, inspired by: "The current disconnect among algorithm developers, implementers, and users has ... foster[ed] distrust in statistical monitoring and in biosurveillance itself" [Shmueli and Burkom, 2010].

3 Monitoring and Study Background

We identified the interests of the following involved parties through interviews and discussions over 2 years. These following insights were synthesized from interviews and interactions with 5 public health departments.

A core responsibility of the Delphi Research Group at Carnegie Mellon University is to its public health stakeholders. Since 2020, these stakeholders, including public health decision makers, modelers, and journalists wanted to find the 'needle in the data haystack' – or important public health events from Delphi's data relevant to *their* regions. They were concerned that a black-box AI approach might overlook their regions. This was particularly important since smaller populations have historically been deprioritized because the underlying alerting methods penalized small counts, leading to systematic underreporting.

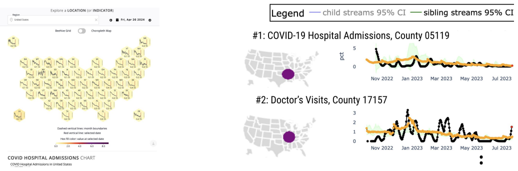

Explore a LOCATION (or INDICATOR)
Region
United States
Fri, Apr 26 2024
Beehive Grid
Choropleth Map
Dashed vertical lines: month boundaries
Red vertical line: selected date
Hex fill color: value at selected date
0.0
2.0
4.0
6.0
8.0
COVID HOSPITAL ADMISSIONS CHART
COVID Hospital Admissions in United States
Legend
child streams 95% CI
sibling streams 95% CI
#1: COVID-19 Hospital Admissions, County 05119
pct
5
0
Nov 2022
Jan 2023
Mar 2023
May 2023
Jul 2023
#2: Doctor's Visits, County 17157
3
2
1
0
Nov 2022
Jan 2023
Mar 2023
May 2023
Jul 2023

Figure 3: **A.** Baseline 1 (**deployed**) used to document events for stakeholder manually, but subject to the drill-down fallacy across dimensional layers. **B.** Baseline 2 showed a limited set of ranked data streams but was not used to document events.

Additionally, a fully closed-looop AI based approach for monitoring was unacceptable. Human reviewers for public health data provide critical context for missing and rapidly changing data in public health settings, especially in cases where the reporting structure fundamentally changes in a way that would take several days for an AI model to automatically adapt. For instance, a sharp rise in hospitalizations may appear to signal an outbreak to an AI method, but a human could contextualize this given other changes, like a new policy about how hospitalizations are recorded. Given that liability was also a concern, the AI method could not operate independently and all outputs needed to be 'human-verified'. The following parties (number involved) were then tasked with meeting stakeholder requirements for monitoring.

- **Data Reviewers (3):** They perform the daily data monitoring. To meet their aformentioned output goal from the monitoring system their key performance indicators (KPIs) around efficiency (how quickly they identify events) and efficacy (accuracy of events identified). They need interfaces that are responsive, provide necessary context, and are easy to learn with minimal onboarding [Ooge *et al.*, 2022].
- **Engineers (3):** Building a functional monitoring system is considerable effort, especially one that serves and visualizes large volumes of data quickly. The three engineers on this project needed to find a balance between data processing and system responsiveness so that data reviewers would not lose focus or attention[3].
- **Computer Scientist (3):** Few monitoring methods are appropriate for public health data because it is noisy, nonstationary, and has dynamic correlation structures. Existing methods generally have core statistical problems, which should be avoided in new design.

After this initial information-gathering phase, a core team of 2 data reviewers, 2 engineers, 2 computer scientists was tasked with deployment. This core group collaborated for 18 months, meeting at least biweekly to refine the **deployed** system and interface.

4 Modernizing Data Monitoring with AI

Based on our initial observational interviews, public health data reviewers perform three actions when monitoring data (Fig. 2):

- *Identify unexpected (anomalous) data points.* These may indicate either reporting errors or early signs of an outbreak. Traditionally, this has been done manually or by using an alerting method.

[3] The past 200 days of data are queried from a virtual database in MySQL v.8. The event ranking method, written in Python3.10, is run separately on a 3.3GHz Intel Core i7 machine with Pop!OS daily with a cron job; the results are stored in a different MySQL v.8 database.

- *Contextualize anomalies.* Reviewers assess anomalies within a broader situational context to ensure meaningful events are not overlooked. This triaging process helps them prioritize data for stakeholders.
- *Decide whether to continue reviewing.* This decision balances the reviewing urgency with the reviewer's attention capacity and capability to perform high-quality triaging. This is often based on the severity of recent triaged data points.

As described, alerting systems fail to support these tasks effectively at scale.

4.1 Alerting-Paradigm Approach (Baseline 1)

It was not immediately clear that an AI approach was appropriate or needed given that public health stakeholders wanted contextualized, expert-verified data. Our tested alerting based approaches included:

1. An approach similar to DHIS2 where we set alerts for z-scores calculated using an adaptive rolling window. Unfortunately, we ran into the same problem as [Coletta and Zhou, 2019] where there were tens of thousands of alerts daily that, when reviewed, provided little reason to remain engaged with the system (Statistical Limitation).

2. Batching alerts across different contextually meaningful dimensions (usually by indicator [Sarikaya *et al.*, 2018; for Disease Control and Prevention, 2023]) to support situational awareness masked important data [Peña *et al.*, 2016] because it was difficult to know in advance which dimension these alerts should be batched along to be the most meaningful (Reviewer Limitation).

3. **Baseline 1:** ***Deployed Alerting System (120 weeks)*** Without a reliable automated system, reviewers reverted to manual data inspection using exploratory tools (Fig. 3A). However, identifying anomalies through visual inspection was inefficient and error-prone [Joshi *et al.*, 2023; Sibolla *et al.*, 2018; Hurt-Mullen and Coberly, 2005]. Further, aggregation strategies often led to a *drill-down fallacy* [Lee *et al.*, 2019], where reviewers could still end up needing to explore a combinatorial number of dimensions to locate the data event that corresponded to the visual alert.

4.2 Alerting-Based Monitoring Fallacies

Reviewer experiences with Baseline 1 became increasingly negative and echoed those of other public health professionals [Coletta and Zhou, 2019]. As expected by the public health stakeholders, reviewers missed important events, especially those in smaller regions and outside of the indicators that are displayed first on the interface. Then, between a) randomly choosing data filters, b) using multiple clicks to drill down to the raw data level, c) finding the appropriate regional tier responsible for the event by trial and error, and d) scrolling to compare indicator behavior across signals, whether relevant or not, reviewers became fatigued.

Baseline 1's failure was the outcome of an alerting paradigm. When reviewers were presented with thousands of alerts to review daily, reviewers abandoned the statistical alerting approach and instead *preferred the manual inspection method*, where they would review data based on their intuition (e.g. what they heard on the news). This was in *conflict* with some stakeholders expectations that all regions have the opportunity to be inspected in a systematic fashion.

Additionally, the computational scientists could not identify any alerting-based methods that could adapt to the non-stationarity and fluctuation volume of the data consistently enough without requiring constant parameter tuning, let alone one lightweight enough to meet the engineering constraints. After several initial design iterations, it became clear that there were two limitations to alerting:

Perscriptive Alerting Fallacy: Many monitoring methods require parameter tuning over time to reflect changes in data dynamics. This tuning is in an effort to align the alerts with user expectations, but in doing so, often only catches the same types of anomalous data points a reviewer already knows about. In effect by focusing on identifying the known classes of unexpected data points, there are many other types that go undetected.

Instead of deploying methods that require human approaches which are strongly prescriptive for data event detection, these systems *should* be framed as measuring deviance from expected data. This is especially important because data reviewers cannot not determine if the alerts they were inspecting were not interesting because it was a calm day **or** because the alerting thresholds were tuned incorrectly and needed to be fixed, limiting their situational awareness and trust in the system.

Threshold Fallacy: The nature of the alerting paradigm is that while data points in a single stream can be prioritized, there is no standard way to do so across millions of data steams. Thus, the range of high priority and low priority "alarms" are treated the same – even if there is a difference, which there usually is from tens of thousands of alarms. Reviewers, already with limited time and energy need to decide if continuing to inspect data is worth it. Under the alerting paradigm, reviewers could not determine whether they were seeing a low-alert day (i.e., truly calm public health conditions) or if the alerting system was miscalibrated—further diminishing engagement.

4.3 Shift to AI-Based Ranking

To address these issues, we designed a system around **anomaly ranking** rather than threshold-based anomaly alerting. Unlike binary alerts, a ranking-based AI system:

- Scores data points based on deviation from expectation (addressing prescriptive alerting).
- Helps reviewers decide whether further inspection is warranted (addressing threshold).
- Aligns with how human experts naturally assess data (meets observations).

This was especially important as reviewers were not mandated to use the system for a specific amount of time or in any forced way. Instead, their investment in monitoring was directly tied to how important they thought the day's alerts were. Of the few *anomaly ranking methods* that support this setting, the most appporpriate was a multiple-univariate outlier detection approach for monitoring [Blázquez-García *et* *al.*, 2021], specifically, a validated human-in-the-loop, linear-time monitoring method [Joshi *et al.*, 2024]. This method relies on the user to define *data expectations* and scores data points at scale, making it an appropriate class of AI methods to base the monitoring system in.

Baseline 2: Method AI Approach The selected method was run in isolation over 30 weeks with minimal changes to a threadbare system. While this toy approach was not used to document events for stakeholders, it serves as a second type of baseline for just the AI ranking approach in isolation vs. part of a human-in-the-loop system over time. In this approach, we displayed interactive line plots in a static HTML file for the top-k data streams (Fig. 3B).

With the underlying AI method in place, the next step was to design a robust interaction model that met reviewers needs and could document events for stakeholders.

5 Design Guidelines and Expectations

To design interaction modalities with the AI ranking method, the group developed the following design guidelines:

(C1) **Meaningful Interaction with AI Method:** To avoid the circular logic of the prescriptive alerting fallacy, changes computer scientists make to the method had needed to be translated directly from interactions with the engineers or data reviewers. This ensured that changes were grounded in real-world use cases rather than arbitrary threshold adjustments.

(C2) **Providing Situational Awareness:** To address the issue of misleading alert inspection the system interface components must provide context to analyze and triage data. Here it is critical to balance avoiding interface and visualization oversimplification with cognitive overload to prevent data misinterpretation.

(C3) **Engagement.** Engagement is critical to sustained system use, but can be easily diminished. One challenge previously observed was *algorithmic fatigue*, where reviewers become less engaged with the system, e.g. due to the volume of alerts.

6 Human-AI Interaction Modalities

In consideration of the previously outlined constraints and design guidelines, the team developed the following modalities for human interaction with the AI method (Fig. 2). In general, the AI method is designed to incorporate correctness and computational constraints specified by the methodologist. These constraints or parameters are integrated into the method to balance computational efficiency with accuracy. The methodologist collaborates with both engineers and data reviewers to iteratively refine these constraints, ensuring that the system meets their needs. For example, changes in the data quality of public health require modifications to the AI method's inputs (e.g. adjusting aggregation policies, incorporating multivariate forecasting approaches to account for signal lag or correlation, or interpolating missing data). Modifications to the method typically occur on a monthly basis, aligning with the frequency at which public health data collection processes and statistical properties evolve.

Beyond these interaction modalities, reviewer analysis and triaging actions require an *interface* that appropriately reflect the AI method's outputs. Together, the AI method, interaction modalities, and the interface constitute the system design for our modern monitoring system. To evaluate the effectiveness of this system design beyond the AI method evaluation alone (Baseline 2), the group focused on strategies to enhance situational awareness and user engagement. Additionally, the team developed a meta-deployment and evaluation framework to assess Human-AI interaction effectiveness *over time*. In this approach, we make sequential changes to the interface that reflect reviewer needs in ways that ensure reviewers have sufficient time to adapt to system changes and provide informed feedback [Carroll *et al.*, 2014; Janes *et al.*, 2013].

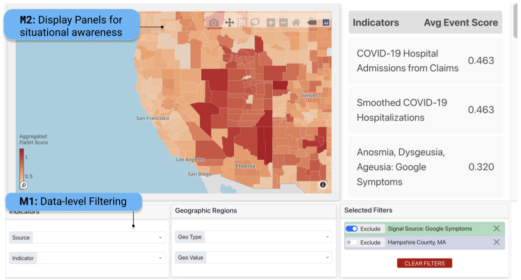
7 Interface & Interaction Experiments

The basic revised monitoring system takes the AI ranked list of data streams and (Fig. 5) segments the data so the temporal dimension is emphasized in analysis. Reviewers can easily expand each data row, which includes a custom map to orient the reviewer and other stream properties. Each row also contains an interactive line plot with data streams. Notably, *contextualization* for data across geographical tiers is controlled by legend items the reviewer can toggle to see data streams in the same tier that share the same regional parent (i.e., a sibling stream), as well as parent streams and child streams with a 95 % CI. To standardize the triaging process, they create a record corresponding to the type of event (e.g., a provider issue), its severity (low, medium, or high), and if the point identified was the source of the event (yes/no).

Notably, towards situational awareness, reviewers also used an interface to record *meta-events* that combine multiple events from individual data points into informative, higher-level phenomena. These meta-events are based on hypotheses across events that reviewers want to investigate, and are very informative to Delphi's stakeholders if there are a large number of pressing events. There were 3 modifications to the basic system that were developed and evaluated (M1, M2, M3):

M1: Data Point Filter Given the choice to focus on the temporal dimension of the data as the primary interaction segment, we need to design how users will access other data segments. The number of possible filtering combinations in this setting, especially considering the desire for multiple filters (including exclusions) per category of data provider, indicator, and geographic region. Given this, we use this step to validate if the performance simple filtering strategy is appropriate for data review.

M2: Situational Awareness Panels Our approach supports situational awareness via two displays of event scores segmented and aggregated across geography and indicator. This is only possible because event scores can be compared across these dimensions wheras raw values cannot due to spatial heterogeneity of public health data. Scores $\mathbf{s}(x)$ across different spatial tiers $e \in \mathcal{E}$ where $\mathcal{E}$ includes county, state, and nation) are aggregated across indicators $i \in I$, and time (T:T-7). For the map, the choropleth color value c for each county



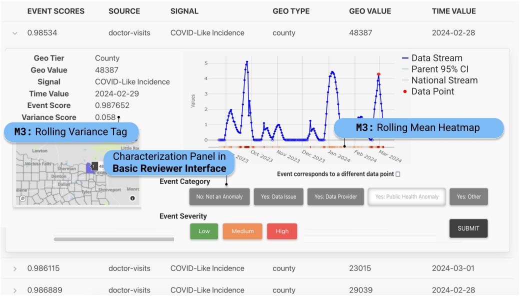
Indicators	Avg Event Score
COVID-19 Hospital Admissions from Claims	0.463
Smoothed COVID-19 Hospitalizations	0.463
Anosmia, Dysgeusia, Ageusia: Google Symptoms	0.320

Indicators
Source
Indicator

Geographic Regions
Geo Type
Geo Value

Selected Filters
Exclude Signal Source: Google Symptoms
Exclude Hampshire County, MA
CLEAR FILTERS

Figure 4: The initial displays support situational awareness and help reviewers get a sense of where data events may be found **(M2)**. This can help them guide their review via the data level filters (**M1**). **M3** captures the impact of data evolution on event scores for each data row, and is shown on Fig 5.



EVENT SCORES	SOURCE	SIGNAL	GEO TYPE	GEO VALUE	TIME VALUE
0.98534	doctor-visits	COVID-Like Incidence	county	48387	2024-02-28
0.986115	doctor-visits	COVID-Like Incidence	county	23015	2024-03-01
0.986889	doctor-visits	COVID-Like Incidence	county	29039	2024-02-28

Geo Tier	County
Geo Value	48387
Signal	COVID-Like Incidence
Time Value	2024-02-29
Event Score	0.987652
Variance Score	0.058

Event corresponds to a different data point □

Event Category
No: Not an Anomaly
Yes: Data Issue
Yes: Data Provider
Yes: Public Health Anomaly
Yes: Other

Event Severity
Low
Medium
High
SUBMIT

Figure 5: For mechanism **M3**, we added a tag with the rolling variance and a 1D heat-map of the rolling mean event scores so that reviewers have an intuition for how event severity changes over data revisions and time.

(on a scale of 0 to 1) is calculated using $\forall r \in \mathcal{R}, \quad c(\mathrm{r}) = \frac{\sum_{e \in \mathcal{E}} \frac{\sum_{i \in I} \log(\bar{s}(x_{i,e(\mathrm{r}),T-7:T})+1)(/log(\mathrm{w})}{|I|}}{|\mathcal{E}|}$

, where e(r) is the region that subregion r belongs to at tier e. The log scores make the more extreme events appear more clearly on the map. The indicator display scores are calculated using $\frac{\sum_{r \in \mathcal{R}} (s(x_{i,r})(T-7:T))}{|\mathcal{R}|}$.

M3: Visualizing Data Evolution Finally, as data is revised over time, the historical event scores also change with the additional data availability and presence of new events. Capturing this evolution of event scores across historical revisions communicates the uncertainty [Carroll *et al.*, 2014] of the event scores over time to reviewers. Our approach for capturing the evolution event scores is inspired by the design of industrial stagger charts [Grove, 2015]. It involves calculating the rolling mean and standard deviation over time, across data revisions, via Welford's online methods [Welford, 1962]: $\bar{x}_T = \frac{\bar{x}_{T-1} \times (T-1) + x_T}{T}$; $v_T = \frac{v_{T-1} - (\bar{x}_{T-1} - x_T) \times (\bar{x}_T - x_T)}{T}$. We include a 1D heat map under the interactive time series plot to give reviewers the context of the event history and provide the average variance score across time. These help the reviewer understand the volatility in event scores over time, as shown in Figure 5.

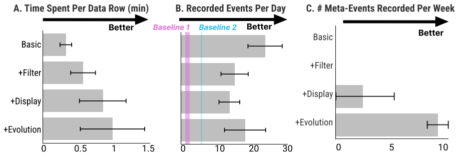


Figure 6: **A.** Reviewer engagement, displayed here with 95% CI bars, increased with each added modification. Baselines had no comparable metrics. **B.** Reviewers recorded significantly more events on average than the prior baselines. **C.** More meta-events were identified after situational awareness displays were added.

8 Evaluation

Evaluating public health monitoring systems generally remains a challenge [Hyllestad *et al.*, 2021]. In particular, longitudinal studies across large volumes of changing data remain under-discussed, despite their importance for adoption. As supported by Delphi's stakeholders, longitudinal studies are especially important for this data streaming evaluation because the daily reviewing load and the number of daily events vary. Typically, monitoring systems that initially perform well degrade over time, reducing reviewer trust and utility of these systems. Given this gap, our longitudinal, sequential evaluation uses metrics corresponding to key performance indicators (KPIs) data reviewers are evaluated against in meeting their intial goals, which we *preregistered* on OSF before any evaluation began.

Data Reviewer KPIs: Reflecting the initial 'gold standard' goal of correctly analyzing high-priority events effeciently, we used the following types of metrics:

- *Efficiency metrics:* (how quickly) time per row, number of events recorded per session
- *Efficacy metrics:* (how well) number of events that were later revised, number of *meta-events* recorded.
- *Output metrics:* filter use, % of times that the method identified data point was not the source of the event, and the distribution of the reviewer characterizations.

To **contextualize** these numbers, we also include excerpts from a reflection that expert data reviewers published in a blog. Each evaluation phase took the standard public health timeline of 3-4 weeks [Biggerstaff *et al.*, 2018].

Efficiency metrics quantified a) how long reviewers interacted with each data row and b) the number of recorded events per day. As Fig. 6A. shows, reviewers generally spent more time per row after each modification, particularly our choice of filters and adding display panels for situational awareness. This suggests that these modifications allowed reviewers to analyze the data deeply: *"[the system] allow[s] me to devote more of my time and efforts to assessing points of interest."*

Reviewers also recorded far more events on average than with the prior baselines (Fig. 6B.); reviewers were **54x** faster on average than while using the exploratory system in Baseline 1, and **5x** faster than the Baseline 2 when recording events/minute. These metrics are contextualized by the reviewer: *"[With the prior approaches], I was spending a good amount of time scrolling, manually sorting, documenting, and*

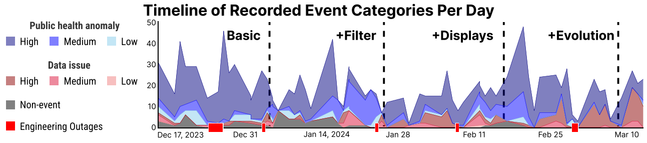


Figure 7: Recorded events per session (up to 49) are far greater than the 1-2 events reviewers would detect per session using Baseline 1. After filters were added, fewer rows were marked as a Non-event, suggesting that reviewers knew how to exclude data that was not interesting without complex synthesis strategies.

searching for specific [event] reports I wanted to examine rather than focusing solely on identifying, marking, and analyzing [events].”

On efficacy, reviewers identify high quality events using the triaging system. If reviewers make a mistake and wish to correct a recorded event, they can easily update the record. In the past, this was frequently used as there are multiple informative external sources of outbreaks that reviewers contextualize against. While historically, this led to edits (Baseline 2’s responses had at least 3 edits across a similar experimentation timeline), there were no edits during the duration of this experiment. More importantly, reviewers identified meta-events when they could investigate patterns in the events that suggested higher-level phenomena. For example, a reviewer identified the following meta-event: *“Several counties in Puerto Rico are repeatedly experiencing sudden upward trending, [respiratory illness] spikes, this month.”*. No such meta-events were recorded for Baseline 1 and only 2 were recorded for Baseline 2, as shown in Fig. 6.

Finally, reviewers also seem to have a positive experience with this system, sharing: *“the updated [triaging system] now enables me to [make meta-events] for exciting [events], trends and other issues of importance, and maintain these notes in an organized, searchable fashion within the platform.”* In a quality assurance check, these meta events were re-analyzed and corresponded with notable public health events.

In terms of output, the resulting documented events from the monitoring process, as shown in Fig. 7, were a mix of data quality and public health issues. Before filters were added to support the segmentation decision, there were many more points marked as a Non-event. However, after reviewers became more familiar with filters, they could exclude data that would generate high event scores but were not important or meaningful, like indicators that providers have stopped maintaining. Reviewers used filters on average 2.75 ± 0.43 times per day. Each filter can have up to 4 predicates (across signal, source, geo value, geo region), but reviewers only use an average of 1 predicate and only 1 value per predicate – usually across geography or signal provider, once again supporting the desire for simple segmentation strategies instead of the fusion based norm in visualization.

Because of the variance of the number of events detected in Fig. 7, online interfaces like the basic review interface are helpful because the number of rows reviewers will process (k) is *unknown and unknowable* because reviewers are not required to use the system in any way – their interest and engagement depend on their trust in the system and the number of important events that day.

8.1 Lessons Learned and Guidelines:

The findings from this study underline the importance in designing and building a Human-centered AI system that matches the needs and constraints of the domain, as described here with public health monitoring stakeholders and an AI anomaly ranking method. Based on the relevance of our evaluation strategy for practical adoption, we advocate more researchers in human-centered AI focus on longitudinal evaluation. For example, the data outages that sporadically occurred over months were unplanned, but are typical for data monitoring in practice and performance of the system would not have necessarily been tested on that setting. Because these results were measured under realistic conditions, Delphi’s data reviewers have continued to use this system and approach, which demonstrates promise for other practical applications. Our next steps aim to address evolving stakeholder pain points from an interdisciplinary lens:

- Generalizing analysis across more data dimensions (e.g., age subgroups when available) [Rolka *et al.*, 2007]
- Improved meta-event detection by identifying event subsequences [Shmueli and Burkom, 2010]
- Automating event summaries and generating narratives for stakeholders [Coletta and Zhou, 2019]

These next steps are well motivated given their importance to both Delphi’s data reviewers and the field more generally [Burkom, 2017].

9 Conclusion

This work contributes an AI-based public health data monitoring system, with implications for large-scale data monitoring generally. This approach is designed to address the limitations of traditional alert-based methods at scale by using a collaborative design process involving engineers, computer scientist, and data reviewers with an AI ranking method. Our resulting user evaluation reveals that this approach boosts user engagement and detection rates and allows reviewers to correctly identify events **54x** faster than the previous deployed system (Baseline 1) and 6x faster than the method alone (Baseline 2), as is relevant to the data reviewer’s goal while monitoring public health data.

References

Matthew Biggerstaff, Michael Johansson, David Alper, Logan C Brooks, Prithwish Chakraborty, David C Farrow, Sangwon Hyun, Sasikiran Kandula, Craig McGowan, Naren Ramakrishnan, et al. Results from the second year of a collaborative effort to forecast influenza seasons in the united states. *Epidemics*, 24:26–33, 2018.

Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54(3):1–33, 2021.

David L Buckeridge. Outbreak detection through automated surveillance: a review of the determinants of detection. *Journal of biomedical informatics*, 40(4):370–379, 2007.

Howard S Burkom. Evolution of public health surveillance: status and recommendations, 2017.

Nan Cao, Chaoguang Lin, Qiuhan Zhu, Yu-Ru Lin, Xian Teng, and Xidao Wen. Voila: Visual anomaly detection and monitoring with streaming spatiotemporal data. *IEEE transactions on visualization and computer graphics*, 24(1):23–33, 2017.

Lauren N Carroll, Alan P Au, Landon Todd Detwiler, Tsung-chieh Fu, Ian S Painter, and Neil F Abernethy. Visualization and analytics tools for infectious disease epidemiology: a systematic review. *Journal of biomedical informatics*, 51:287–298, 2014.

Hsinchun Chen, Daniel Zeng, and Ping Yan. *Infectious disease informatics: syndromic surveillance for public health and biodefense*, volume 21. Springer, 2010.

Hsinchun Chen, Daniel Zeng, Ping Yan, Hsinchun Chen, Daniel Zeng, and Ping Yan. Data visualization, information dissemination, and alerting. *Infectious Disease Informatics: Syndromic Surveillance for Public Health and BioDefense*, pages 73–87, 2010.

Michael Coletta and Hong Zhou. What can you really do with 35,000 statistical alerts a week anyways? *Online Journal of Public Health Informatics*, 11(1), 2019.

Zikun Deng, Di Weng, Shuhan Liu, Yuan Tian, Mingliang Xu, and Yingcai Wu. A survey of urban visual analytics: Advances and future directions. *Computational Visual Media*, 9(1):3–39, 2023.

Xueying Ding, Nikita Seleznev, Senthil Kumar, C Bayan Bruss, and Leman Akoglu. From explanation to action: An end-to-end human-in-the-loop framework for anomaly reasoning and management. *arXiv preprint arXiv:2304.03368*, 2023.

Ensheng Dong, Jeremy Ratcliff, Tamara D Goyea, Aaron Katz, Ryan Lau, Timothy K Ng, Beatrice Garcia, Evan Bolt, Sarah Prata, David Zhang, et al. The johns hopkins university center for systems science and engineering covid-19 dashboard: data collection process, challenges faced, and lessons learned. *The lancet infectious diseases*, 22(12):e370–e376, 2022.

Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014.

Centers for Disease Control and Prevention. National syndromic surveillance program (nssp) new users. https://www.cdc.gov/nssp/new-users.html, 2023.

Andrew S Grove. *High output management*. Vintage, 2015.

J Jabanjalin Hilda, C Srimathi, and Bhulakshmi Bonthu. A review on the development of big data analytics and effective data visualization techniques in the context of massive and multidimensional data. *Indian Journal of Science and Technology*, 9(27):1–13, 2016.

Richard S Hopkins, Catherine C Tong, Howard S Burkom, Judy E Akkina, John Berezowski, Mika Shigematsu, Patrick D Finley, Ian Painter, Roland Gamache, Victor J Del Rio Vilas, et al. A practitioner-driven research agenda for syndromic surveillance. *Public Health Reports*, 132(1_suppl):116S–126S, 2017.

Kathy J Hurt-Mullen and J Coberly. Syndromic surveillance on the epidemiologist's desktop: making sense of much data. *MMWR Morb Mortal Wkly Rep*, 54(Suppl):141–6, 2005.

Susanne Hyllestad, Ettore Amato, Karin Nygård, Line Vold, and Preben Aavitsland. The effectiveness of syndromic surveillance for the early detection of waterborne outbreaks: a systematic review. *BMC Infectious Diseases*, 21:1–12, 2021.

Andrea Janes, Alberto Sillitti, and Giancarlo Succi. Effective dashboard design. *Cutter IT Journal*, 26(1):17–24, 2013.

Ananya Joshi, Kathryn Mazaitis, Roni Rosenfeld, and Bryan Wilder. Computationally assisted quality control for public health data streams. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6004–6012, 2023.

Ananya Joshi, Tina Townes, Nolan Gormley, Luke Neureiter, Roni Rosenfeld, and Bryan Wilder. Outlier ranking in large-scale public health streams. *Proceedings of the AAAI Conference*, 2024.

Suraj P Kesavan, Takanori Fujiwara, Jianping Kelvin Li, Caitlin Ross, Misbah Mubarak, Christopher D Carothers, Robert B Ross, and Kwan-Liu Ma. A visual analytics framework for reviewing streaming performance data. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pages 206–215. IEEE, 2020.

Tara Lakdawala and Ananya Joshi. Identifying changing variant behavior during a pandemic: An exploratory analysis. https://delphi.cmu.edu/blog, Dec 2023.

Doris Jung-Lin Lee, Himel Dev, Huizi Hu, Hazem Elmeleegy, and Aditya Parameswaran. Avoiding drill-down fallacies with vispilot: Assisted exploration of data subsets. In *Proceedings of the 24th International Conference on Intelligent User Interfaces,{IUI} 2019, Marina del Ray, CA, USA, March 17-20, 2019*, 2019.

Ross Maciejewski, Stephen Rudolph, Ryan Hafen, Ahmad Abusalah, Mohamed Yakout, Mourad Ouzzani, William S Cleveland, Shaun J Grannis, and David S Ebert. A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):205–220, 2009.

Brian Montambault, Camelia D Brumar, Michael Behrisch, and Remco Chang. Pixal: Anomaly reasoning with visual analytics. *arXiv preprint arXiv:2205.11004*, 2022.

Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. Explaining artificial intelligence with visual analytics in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(1):e1427, 2022.

Miguel Angel López Peña, Carlos Area Rua, and Sergio Segovia Lozoya. A "fast data" architecture: Dashboard for anomalous traffic analysis in data networks. In *2016 Eleventh International Conference on Digital Information Management (ICDIM)*, pages 37–42. IEEE, 2016.

Bernhard Preim and Kai Lawonn. A survey of visual analytics for public health. In *Computer Graphics Forum*, volume 39, pages 543–580. Wiley Online Library, 2020.

Alex Reinhart, Logan Brooks, Maria Jahja, Aaron Rumack, Jingjing Tang, Sumit Agrawal, Wael Al Saeed, Taylor Arnold, Amartya Basu, Jacob Bien, et al. An open repository of real-time covid-19 indicators. *Proceedings of the National Academy of Sciences*, 118(51):e2111452118, 2021.

Henry Rolka, Howard Burkom, Gregory F Cooper, Martin Kulldorff, David Madigan, and Weng-Keen Wong. Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: research needs. *Statistics in Medicine*, 26(8):1834–1856, 2007.

Alper Sarikaya, Michael Correll, Lyn Bartram, Melanie Tory, and Danyel Fisher. What do we talk about when we talk about dashboards? *IEEE transactions on visualization and computer graphics*, 25(1):682–692, 2018.

Galit Shmueli and Howard Burkom. Statistical challenges facing early outbreak detection in biosurveillance. *Technometrics*, 52(1):39–51, 2010.

Bolelang H Sibolla, Serena Coetzee, and Terence L Van Zyl. A framework for visual analytics of spatio-temporal sensor observations from data streams. *ISPRS International Journal of Geo-Information*, 7(12):475, 2018.

Xian Teng, Yu-Ru Lin, and Xidao Wen. Anomaly detection in dynamic networks using multi-view time-series hypersphere learning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 827–836, 2017.

Catalina Vajiac, Duen Horng Chau, Andreas Olligschlaeger, Rebecca Mackenzie, Pratheeksha Nair, Meng-Chieh Lee, Yifei Li, Namyong Park, Reihaneh Rabbany, and Christos Faloutsos. Trafficvis: visualizing organized activity and spatio-temporal patterns for detecting and labeling human trafficking. *IEEE transactions on visualization and computer graphics*, 29(1):53–62, 2022.

BP Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420, 1962.

WHO. Who hub for pandemic and epidemic intelligence. https://pandemichub.who.int/publications/m/item/the-who-hub-for-pandemic-and-epidemic-intelligence-strategy-paper, Dec. 2022. Accessed: 2023-06-05.