Reducing Alert Fatigue Through AI Ranking: A Deployed Public Health Data Monitoring System

Ananya Joshi[1,2], **Nolan Gormley**[2], **Richa Gadgil**[2], **Catalina Vajiac**[2], **Tina Townes**[2], **Roni Rosenfeld**[2], **Bryan Wilder**[2]

[1]Johns Hopkins University
[2]Delphi Research Group, Carnegie Mellon University

Abstract

Public health experts need scalable methods to monitor large volumes of health data, like human-reported cases, hospitalizations, and deaths. These methods must identify individual data points that may indicate significant events, such as outbreaks or data quality issues. Experts then triage and analyze identified data points and report them so that they can prevent downstream errors in forecasting or policy. Still, traditional *alert-based* data monitoring systems, used for decades in public health practice, fail to identify relevant data for several reasons, including that these systems may not output real-time results from large data volumes, and when they do complete, they may return tens of thousands of unhelpful alerts.

We introduce a human-in-the-loop AI system for public health data monitoring that uses a *ranking-based* AI anomaly detection method. This system was developed through a multi-year interdisciplinary collaboration with participatory design from researchers, engineers, and public health data experts. From this process, we identified and designed around system goals, such as user control and efficiency. This system has since been *deployed* at a national public health organization under loads of up to 5 million data points daily. A three-month longitudinal deployment evaluation revealed a significant improvement in system goals, including a 54x increase in data reviewer efficiency and increased engagement compared to traditional alert-based methods. Finally, we discuss design considerations for managing uncertainty, including how experts interpret false positives and false negatives.

Introduction

Public health data monitoring systems are critical for detecting events like outbreaks and data quality issues (Fig. 1), but traditional alert-based systems struggle with modern, high-volume public health data streams (Shmueli and Burkom 2010). This change in data has been the result of significant investments in increasing the quality, speed, and volume of public health-related data (WHO 2022). Now, public health data is complex and heterogeneous, which unfortunately makes it incompatible with historical public health monitoring systems designed for smaller data volumes. In fact, identifying events like outbreaks and data quality issues paradoxically became more difficult after these investments as experts struggled to find *the data points that matter.*

Many traditional data monitoring systems use alert-based methods that rely on predefined thresholds to flag data anomalies corresponding to known types of data quality/outbreak events (Dong et al. 2022). However, these definitions do not adapt to modern data variance across a wide variety of data sources, can yield tens of thousands of alerts (far too many to inspect in real-time), and cannot catch *new, unexpected phenomena* that fall outside expected data quality and outbreak data patterns.

After re-implementing and studying the failure modes across multiple data monitoring systems, our interdisciplinary team of researchers, engineers, and data experts at the Delphi Group at Carnegie Mellon University, a national public health data curator , designed a novel monitoring system (Fig. 2) and interface (Figs. 4 & 5) that uses an unsupervised, AI-based anomaly ranking method to addresses the limitations of modern data monitoring. We used a participatory design process with methodologists, engineers, data reviewers, and public health stakeholders to determine which components of the system were well-suited for an AI approach. We also developed an evaluation to match the key performance indicators for our data experts, including a longitudinal study to measure the utility of the data monitoring system under deployed conditions. Our results demonstrate the 1) effectiveness of

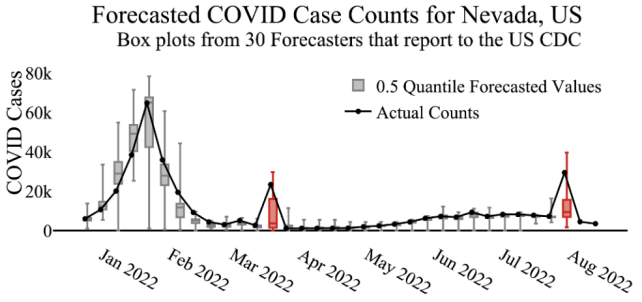

Figure 1: Data monitoring catches events like data quality changes in case counts, shown by the large spikes in actual counts when cases were trending down. Because these events were not caught, they incorrectly impacted the forecasted cases (red) from multiple institutional forecasts.

the ranking paradigm in practice, 2) efficiency of reviewers using the system, and 3) improved overall experience with the system.

Background

We used a participatory design approach to develop a monitoring system that balances the capabilities of AI with relevant constraints, priorities, and responsibilities. Over two years, we conducted interviews and design sessions with 5 public health departments and Delphi members to synthesize the design requirements for the system.

Stakeholder Needs

The Delphi group is responsible to its public health stakeholders, including public health decision makers, modelers, and journalists. Since 2020, these stakeholders have wanted to access data events relevant to their regions, but saw it as a 'needle in the *data* haystack problem'.

Public health stakeholders were generally concerned about missing important events, which they thought Delphi could identify given the volume of data we had access to. Additionally, they felt like a fully black-box AI approach might overlook less populous regions, especially since smaller populations have historically been deprioritized because the underlying alerting methods penalized regions with small counts (e.g., it can be more difficult to quickly find an outbreak in a rural county vs. a city using only statistics). Additionally, there needed to be humans-in-the-loop to provide relevant context for missing and rapidly changing data. In fact, all outputs needed to be 'human-verified' for stakeholders to invest resources and act upon.

The following parties (number involved) were then tasked with meeting stakeholder requirements for monitoring:

Team Makeup

- *Data Reviewers (3):* They perform the daily data monitoring. They need interfaces that are responsive, provide necessary context, and require minimal onboarding (Ooge, Stiglic, and Verbert 2022).
- *Engineers (3):* They build out the functional monitoring system to serve and visualize large volumes of data quickly.

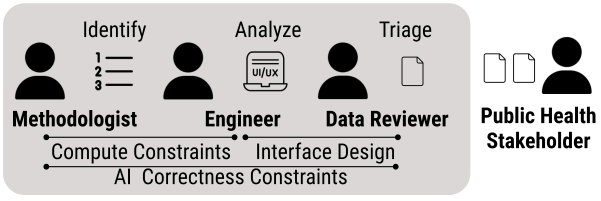

Figure 2: Data reviewers triage notable events from public health data for stakeholders interested in data quality issues or outbreaks. Our system is powered by an AI ranking method that identifies relevant events and formalizes how different stakeholders interact with the AI method.

- *Methodologist (3):* Few monitoring methods are appropriate for public health data because it is noisy, nonstationary, and has dynamic correlation structures. Existing methods generally have statistical problems, which should be avoided in the new design.

Over the next 18 months, this group met at least biweekly to refine the deployed AI model, visualization interface, and data pipeline. This iterative process ensured that the AI system would remain aligned with stakeholder concerns, daily reviewer workflows, engineering constraints, and statistical correctness over time.

Our initial interviews showed us that public health data reviewers perform three actions when monitoring data:

- *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.
- *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 recently triaged data points.

Reviewers found these three functions difficult under existing data monitoring paradigms.

Challenges in Deployed Data Monitoring

Challenges in public health data monitoring closely resemble those in other highly contextual domains (e.g., environmental, financial, and agricultural) because the status quo can change rapidly and decision-making requires 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. These approaches would not work in public health; each individual data point represents an aggregate population, and the consequences can be dire if missed.

Still, we initially tested several existing monitoring system designs, including using z-score variants with thresholds (Coletta and Zhou 2019) and batching alerts (Sarikaya et al. 2018; for Disease Control and Prevention 2023). From our experiences, we identified 3 inadequacies in these systems with statistical properties like modern public health data, including high-volume, low-quality, nonstationarity, noise, missing values, and inconsistencies over time.

Stagnant to Data Changes

Alerting systems at scale on modern data exhibit well-documented issues, including statistical inaccuracies (Joshi et al. 2024; Shmueli and Burkom 2010) and overwhelming alert volumes (Hurt-Mullen and Coberly 2005). Our observations led us to define the following fallacies:

Threshold Fallacy While anomalous data points in a single data stream can be identified using standard methods, there is no standard way to do so across millions of data streams. Thus, the range of high priority and low priority "alerts" are treated the same, even though reviewers should be most focused on high priority alerts. Further, 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, and thus diminishing engagement with the system.

Even the most sophisticated adaptive mechanisms for anomaly threshold setting (Burkom 2017) that try to separate high-priority alerts require constant retuning that is infeasible at a national scale.

Additionally, these alerts often tell reviewers about phenomena they were already aware of and miss high priority unknown phenomena, like a new pathogen outbreak.

Prescriptive 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 statistical 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 of anomalies that go undetected.

Once again, this is important because data reviewers cannot determine if the alerts they were inspecting were not interesting because it was a calm day or because the alerting thresholds were miscalculated. These fallacies caused whole-system issues for reviewers as follows.

Limited Awareness

Individual data point alerts still need to be inspected in the *context* of the larger public health data available (Burkom 2017) to triage, or classify, data points as events and analyze their severity. For example, to identify if there were truly data quality issues in Nevada in March 2022 (Fig. 1), a reviewer may check the surrounding areas or some of the state's counties' data. Instead of this manual process, some systems approach contextualization by fusing data across visual interfaces like dashboards or combined alerts (Deng et al. 2023; Cao et al. 2017; Vajiac et al. 2022). However, the statistical properties of public health data (Reinhart et al. 2021) introduce constrained relationships that are also highly variable (e.g., shifting lead times across data sources and complex correlation structures (Lakdawala and Joshi 2023)). So, static correlation-based techniques also often fail in this context (Fan, Han, and Liu 2014; Hilda, Srimathi, and Bonthu 2016). Anecdotally, existing alerting systems were designed to operate only at a local scale (Chen, Zeng, and Yan 2010). At that scale, *all* data could still be analyzed manually, so these systems were never built to support general data or situational awareness.

Poor Human Experience and Review Burden

Matching external findings (Coletta and Zhou 2019), our reimplementations of existing systems produced tens of thousands of alerts. This was far beyond what experts could review and led to review fatigue and disengagement. A strawman solution was to batch real-time alerts by location, but, for example, alerts for cases and deaths cannot be naively combined because an anomalous rise in deaths usually comes after an anomalous rise in cases . Another approach, handling alerts visually using drill-down tools (Preim and Lawonn 2020; Chen et al. 2010; Maciejewski et al. 2009), offers flexibility but demands high visual literacy and extensive manual effort, which can delay response during real-time crises.

Approach

The need for improved monitoring *systems* has been widely recognized by public health practitioners (Hopkins et al. 2017), and this goes beyond improving methods for data identification in isolation (Buckeridge 2007). In fact, our interdisciplinary systems design approach is inspired by the

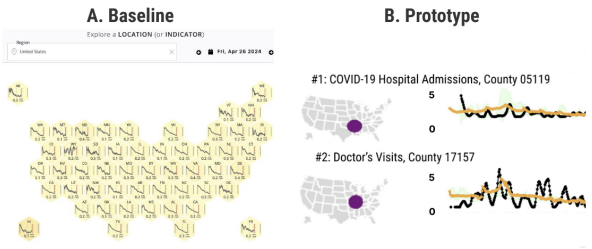

Figure 3: A. *Baseline* exploratory dashboard which displays raw data values B. *Prototype* for identifying ranked data streams *only*.

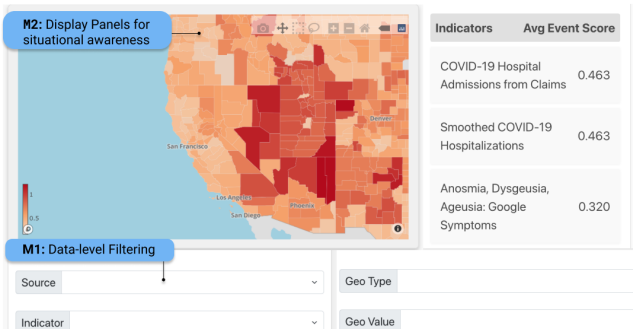

Figure 4: These displays help reviewers determine where anomalous events may be concentrated **(M2)** so that they can filter to the correct data slice (**M1**).

following quote:

> "The current disconnect among algorithm developers, implementers, and users has ... foster[ed] distrust in statistical monitoring and in biosurveillance itself" (Shmueli and Burkom 2010).

Instead of the existing designs, Delphi's data reviewers reverted to *manual review* aided by data exploration tools for monitoring (Fig. 3A), which we refer to as the **Baseline**. As expected, identifying anomalies through this type of visual inspection was inefficient and error-prone (Joshi et al. 2023; Sibolla, Coetzee, and Van Zyl 2018; Hurt-Mullen and Coberly 2005), and a new approach was needed.

From what we learned, reviewer investment in data monitoring was directly tied to how important they thought the day's events were. Thus, reviewers wanted a system that they could use with minimal training at their convenience and would:

- Score data points based on their anomalousness, or the largest deviation from their contextualized expectations.
- Help them decide whether further inspection is warranted.
- Align with how they naturally assess data.

This led us to an AI anomaly ranking method developed explicitly for this setting that produces event scores per data point (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 on. An initial evaluation of the ranking **Prototype** was deployed for over 30 weeks, where we displayed interactive line plots in a static HTML file for the top-k data streams (Fig. 3B).

Interaction Modalities

Based on the AI method, we designed the following interaction modalities (Fig. 2). First, the AI method incorporates 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., incorporating multivariate forecasting approaches to account for signal lag or correlation, interpolating missing data ...). For maintenance, modifications to the method typically occurred 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 reflects the AI method's outputs. Together, the AI method, interaction modalities, and interface constitute the system design for our modern monitoring system. To evaluate the effectiveness of this system design beyond just the *prototype*, the group focused on strategies to enhance data awareness and user engagement *over time*. In our approach, we made 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, a need well documented in the literature (Carroll et al. 2014; Janes, Sillitti, and Succi 2013).

Experimental Design

The basic revised monitoring system takes the AI ranked list of data streams (Figs. 4 and 5) and slices 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 to assist with triage. 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, reviewers create a record corresponding to the type of event (e.g., a provider issue), its severity (low, medium, or high), and if the data point identified was the source of the event (yes/no).

Reviewers also used an interface to record *meta-events* that combine multiple events from individual data points into informative, higher-level phenomena (e.g., an outbreak across the East Coast of the United States could be a meta event that came from analyzing many highly-ranked East Coast states). These meta-events are based on hypotheses across events that reviewers want to investigate, and are very informative to Delphi's stakeholders. In our design, we developed and tested 3 modifications relevant to reviewer key performance indicators (KPIs): (M1, M2, M3):

M1: Data Point Filter

Given the choice to focus on the temporal dimension of the data as the primary data slice, we need to design how users will access other data slices (Fig. 4). The number of possible filtering combinations would be tens of thousands, especially considering the desire for multiple filters (including exclusions) per category of data provider, data source, and geographic region. Given this, we use this step to validate if the performance of a simple filtering strategy is appropriate for data review.

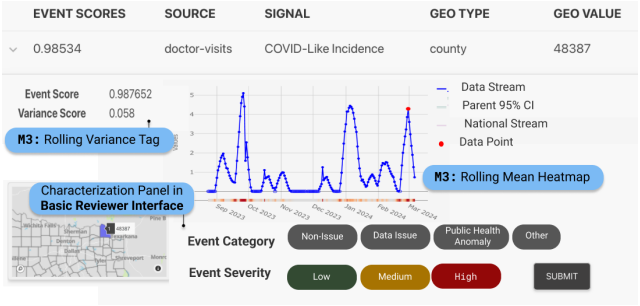

Figure 5: For **M3**, we display the rolling variance and a 1D heat-map of the rolling mean scores so reviewers can understand how anomalies change over revised data and time.

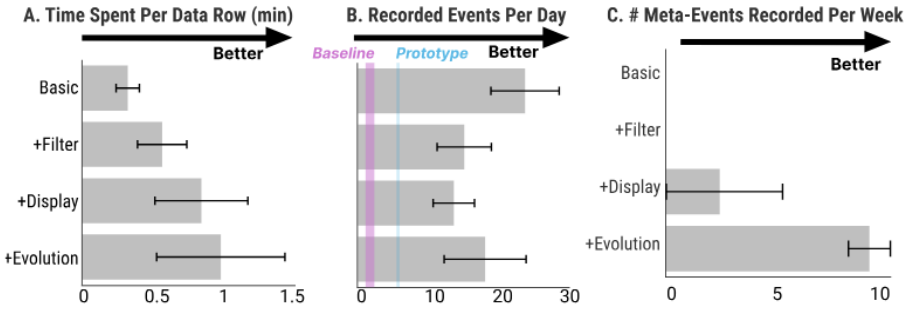

Figure 6: **A.** Reviewer engagement, displayed here with 95% CI bars, increased with each added modification. **B.** Reviewers recorded significantly more events on average than the prior baselines. **C.** More meta-events were identified after data awareness displays were added.

M2: Data Awareness Panels

Our approach supports data awareness via two displays with event scores sliced and aggregated across geography and data source (Fig. 4). This is only possible because event scores from the AI method can be compared across these dimensions, whereas raw values cannot due to spatial heterogeneity of the data. Scores across different spatial tiers are aggregated across data sources, and the last week [1]. For the map, the choropleth color value for each county (on a scale of 0 to 1) is calculated using an average of anomaly scores from each data source at each regional tier to highlight regions with the most *anomalies*.

M3: Visualizing Data Evolution

One nuance of public health data is that it is revised over time. This means that historical counts are subject to change in real time. Thus, historical event scores also change with the additional data availability and new events. Capturing this evolution of event scores across historical revisions communicates the uncertainty 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, or changes to historical data in real time, via Welford's online methods (Welford 1962). 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 Fig. 5.

[1]With $\mathbf{s}(x)$ as event scores from the AI method, the spatial tier $e \in \mathcal{E}$ where $\mathcal{E}$ includes county, state, and nation), data sources $i \in I$, time (T:T-7), and c as the choropleth score, we have: $\forall r \in \mathcal{R}, \quad c(r) = \frac{\sum_{e \in \mathcal{E}} \frac{\sum_{i \in I} \log(\bar{s}(x_{i,e(r),T-7:T})+1)/(log(2))}{|I|}}{|\mathcal{E}|}$, 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 data source display scores are calculated using $\frac{\sum_{r \in \mathcal{R}} (\bar{s}(x_{i,r,T-7:T}))}{|\mathcal{R}|}$.

Evaluation

Evaluating public health monitoring systems is a challenge (Hyllestad et al. 2021). In particular, longitudinal studies across large volumes of changing data remain underdiscussed, despite their importance for adoption. In fact, as supported by Delphi's stakeholders, longitudinal studies are especially important because the daily reviewing load and the number of daily events vary. In fact, monitoring systems that initially perform well typically degrade in practice as the status quo changes, reducing reviewer trust and the utility of these systems. Given this gap, our longitudinal, sequential evaluation uses metrics corresponding to key performance indicators (KPIs) that data reviewers are evaluated against. For our evaluation, we *preregistered* our experimental design with GitHub commits on OSF before any evaluation began.

For data reviewer KPIs, we considered three categories of metrics:

- *Efficiency metrics:* (*speed*) time per row, number of events recorded per session
- *Efficacy metrics:* (*quality*) number of events that were later revised, number of *meta-events* recorded.
- *Output metrics:* resulting analysis and triage from reviewers.

Each evaluation phase took the standard public health timeline of 3-4 weeks (Biggerstaff et al. 2018). No additional system training was given.

Efficiency Metrics

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 after adding filters and display panels for data awareness. This suggests that these modifications allowed reviewers to analyze the correct data slices. One reviewer later documented in a blog post, *"[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* , and **5x** faster than the *Prototype* 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 searching for specific [event] reports I wanted to examine rather than focusing solely on identifying, marking, and analyzing [events]."*

Efficacy Metrics

Reviewers identify high-quality events using the triaging system. If reviewers made a mistake and wanted to correct a recorded event, they could easily update the record. In the past, this function was frequently used as there were multiple informative external sources of outbreaks that reviewers contextualized against. Still, while the *Prototype* 's responses had at least 3 edits across a similar experimentation timeline, there were no edits using the new system. 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* and only 2 were recorded for *Prototype*, as shown in Fig. 6C.

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.

Output Metrics

Reviewer-outputted events, as shown in Fig. 7, were a mix of data quality and public health issues. Before filters were added to help reviewers access the appropriate data slice, there were many more points marked as non-events. However, after reviewers became more familiar with filters, they could exclude data that would generate high event scores but were not contextually important or meaningful, like unmaintained data sources. Reviewers used filters on average 2.75 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. This validates the simple segmentation strategy instead of the fusion-based norm in data visualization.

Additionally, Fig. 7 shows the variance in the number of events reviewers documented, validating the ranking strategy displayed as a list is helpful because the number of rows reviewers will process (k) is *unknown and unknowable*. After all, 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.

Discussion

Contextualizing the efficacy of triaged events was important to public health experts and statisticians (Burkom 2017). Specifically, they requested guidelines on how to understand false positives and negatives.

False Positives

From the reviewer's perspective, false positives occur when a data reviewer misclassifies an event. This may happen if a reviewer is biased towards the AI ranking algorithm and doesn't thoroughly analyze or triage the data. However, such occurrences are unlikely since reviewers need to record events and we conducted a quality assurance check which showed there were no instances of a single recorded event being updated.

From an algorithmic perspective, false positives happen when the event detection algorithm erroneously ranks non-event data highly or when the identified data point doesn't correspond to the event. Before our filtering, some false positives in rows were triaged as non-events by reviewers, but these have been limited since then, as seen in Fig. 7. Additionally, about 14% of events evaluated by reviewers were due to data points near the one identified, but not exactly matching the identified data point. Thus, both the event detection algorithm and the reviewer-in-the-loop approach were needed to identify the event correctly. This insight supports incorporating the evolving relationship between reviewer expectations and event detection algorithm output as part of the system, like we do in Fig 2.

False Negatives

From the reviewer's perspective, false negatives may occur when reviewers incorrectly classify data as non-events, which is uncommon, or when unreviewed data contains events, which is likely more common given the data scale. Still, reviewer capacity is limited, so not all data corresponding to events will be reviewed, and the accuracy of the presented ranking depends on the underlying event detection algorithm. Humans may also anchor their triage on the AI-produced ranking (Cho et al. 2017; Vasconcelos et al. 2023), so a reviewer may stop the investigation on a particular day if there are several uninteresting rows. However, the thought-intensive analysis and triaging process (Vasconcelos et al. 2023) may anecdotally reduce this anchoring effect.

Overall events that were incorrectly triaged (false positives) were far less common than events that were not shown to reviewers to triage (false negatives). Still, public health experts emphasize reducing false negatives, "for outbreak and event detection, practitioners prioritize timeliness and sensitivity over positive predictive value (Hopkins et al. 2017)". Doing so in a way that accounts for the human limitations of data reviewers motivates future work in this domain.

Lessons Learned and Guidelines

Our findings highlight the value of building human-centered AI systems tailored to domain-specific needs and constraints—in this case, a public health monitoring system built around an AI anomaly ranking method.

Our most important lesson was that conducting a deployed *longitudinal evaluation* should be a part of standard public health system evaluations. Real-world factors such as data

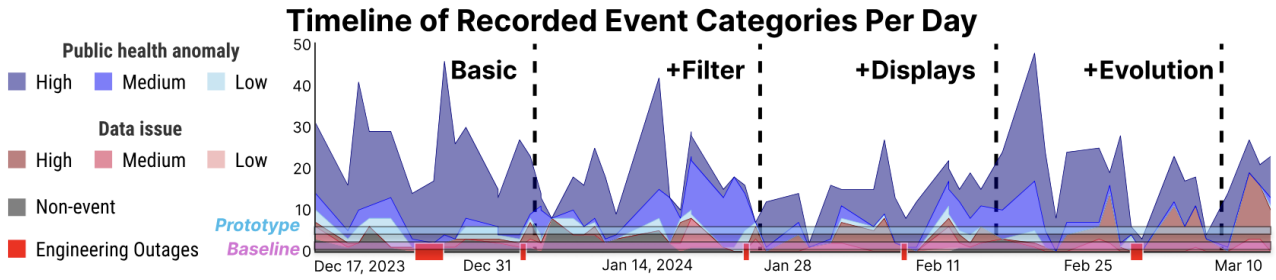

Figure 7: Recorded events per session (up to 49) are far greater than the 1-2 events reviewers would detect per session using *Baseline* . After filters were added, fewer rows were marked as non-events, suggesting that reviewers knew how to exclude data that was not interesting without complex synthesis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

outages, changing reporting patterns, and evolving reviewer needs cannot be captured through static benchmarks alone. For instance, several unplanned data outages provided natural stress tests for the system and revealed the resilience of the human-in-the-loop approach that would not have surfaced in conventional testing settings. These insights were only possible through long-term deployment and ongoing engagement with users.

Based on this experience, we advocate for human-centered AI research to adopt more practical, longitudinal evaluation frameworks—especially in high-stakes or data-rich domains where static testing is insufficient.

Future Work

Future work will continue to address evolving stakeholder needs from an interdisciplinary perspective, including:

- Reducing the human cost of system maintenance through a data-reviewer initiated 'push' mechanism for system updates.
- Enhancing meta-event detection by identifying temporal subsequences (Shmueli and Burkom 2010).
- Automating event summaries to generate structured, stakeholder-facing narratives (Coletta and Zhou 2019).

These directions are well-motivated both by feedback from Delphi's data reviewers and priorities identified in public health surveillance research more broadly (Burkom 2017).

Conclusion

We present a deployed AI-based monitoring system for public health data that replaces traditional threshold-based alerting with a ranking-based anomaly detection approach. This system, designed in close collaboration with researchers, engineers, and public health data reviewers, reflects a paradigm shift in how large-scale population data can be monitored. Unlike previous methods, our approach supports triage, contextualization, and sustained engagement with modern data. In particular, the ranking paradigm helps reviewers prioritize attention across millions of data streams, despite nonstationarity and noise. Reviewer evaluation shows that the system improves both engagement and accuracy, enabling users to detect relevant events up to **54× faster** than with prior deployed systems. This work illustrates how human-centered AI can lead to lasting adoption and opens new directions for real-time data monitoring systems in healthcare and beyond.

Ethical Statement

This research was conducted in accordance with the principles embodied in the Declaration of Helsinki and in accordance with local statutory requirements. All participants consented to the study and could exit the study at any time. Approval was granted from Carnegie Mellon University IRB.

Acknowledgments

This work was supported by the Centers for Disease Control & Prevention as part of a cooperative agreement funded solely by CDC/HHS under federal award identification number U01IP001121, "Delphi Influenza Forecasting Center of Excellence"; and by CDC funded contract number 75D30123C15907, "Digital Public Health Surveillance for the 21st Century". This material is also based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1745016 and DGE2140739. The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by, CDC/HHS, National Science Foundation, or the U.S. Government.

References

Biggerstaff, M.; Johansson, M.; Alper, D.; Brooks, L. C.; Chakraborty, P.; Farrow, D. C.; Hyun, S.; Kandula, S.; McGowan, C.; Ramakrishnan, N.; et al. 2018. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. *Epidemics*, 24: 26–33.

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

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

Cao, N.; Lin, C.; Zhu, Q.; Lin, Y.-R.; Teng, X.; and Wen, X. 2017. Voila: Visual anomaly detection and monitoring

with streaming spatiotemporal data. *IEEE transactions on visualization and computer graphics*, 24(1): 23–33.

Carroll, L. N.; Au, A. P.; Detwiler, L. T.; Fu, T.-c.; Painter, I. S.; and Abernethy, N. F. 2014. Visualization and analytics tools for infectious disease epidemiology: a systematic review. *Journal of biomedical informatics*, 51: 287–298.

Chen, H.; Zeng, D.; and Yan, P. 2010. *Infectious disease informatics: syndromic surveillance for public health and biodefense*, volume 21. Springer.

Chen, H.; Zeng, D.; Yan, P.; Chen, H.; Zeng, D.; and Yan, P. 2010. Data visualization, information dissemination, and alerting. *Infectious Disease Informatics: Syndromic Surveillance for Public Health and BioDefense*, 73–87.

Cho, I.; Wesslen, R.; Karduni, A.; Santhanam, S.; Shaikh, S.; and Dou, W. 2017. The anchoring effect in decision-making with visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 116–126. IEEE.

Coletta, M.; and Zhou, H. 2019. What Can You Really Do with 35,000 Statistical Alerts a Week Anyways? *Online Journal of Public Health Informatics*, 11(1).

Deng, Z.; Weng, D.; Liu, S.; Tian, Y.; Xu, M.; and Wu, Y. 2023. A survey of urban visual analytics: Advances and future directions. *Computational Visual Media*, 9(1): 3–39.

Dong, E.; Ratcliff, J.; Goyea, T. D.; Katz, A.; Lau, R.; Ng, T. K.; Garcia, B.; Bolt, E.; Prata, S.; Zhang, D.; et al. 2022. 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.

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

for Disease Control, C.; and Prevention. 2023. National Syndromic Surveillance Program (NSSP) New Users. https://www.cdc.gov/nssp/new-users.html.

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

Hilda, J. J.; Srimathi, C.; and Bonthu, B. 2016. 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.

Hopkins, R. S.; Tong, C. C.; Burkom, H. S.; Akkina, J. E.; Berezowski, J.; Shigematsu, M.; Finley, P. D.; Painter, I.; Gamache, R.; Vilas, V. J. D. R.; et al. 2017. A practitioner-driven research agenda for syndromic surveillance. *Public Health Reports*, 132(1_suppl): 116S–126S.

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

Hyllestad, S.; Amato, E.; Nygård, K.; Vold, L.; and Aavitsland, P. 2021. The effectiveness of syndromic surveillance for the early detection of waterborne outbreaks: a systematic review. *BMC Infectious Diseases*, 21: 1–12.

Janes, A.; Sillitti, A.; and Succi, G. 2013. Effective dashboard design. *Cutter IT Journal*, 26(1): 17–24.

Joshi, A.; Mazaitis, K.; Rosenfeld, R.; and Wilder, B. 2023. Computationally assisted quality control for public health data streams. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 6004–6012.

Joshi, A.; Townes, T.; Gormley, N.; Neureiter, L.; Rosenfeld, R.; and Wilder, B. 2024. Outlier Ranking in Large-Scale Public Health Streams. *Proceedings of the AAAI Conference*.

Kesavan, S. P.; Fujiwara, T.; Li, J. K.; Ross, C.; Mubarak, M.; Carothers, C. D.; Ross, R. B.; and Ma, K.-L. 2020. A visual analytics framework for reviewing streaming performance data. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, 206–215. IEEE.

Lakdawala, T.; and Joshi, A. 2023. Identifying changing variant behavior during a pandemic: An exploratory analysis. https://delphi.cmu.edu/blog. Accessed: 2025-09-01.

Maciejewski, R.; Rudolph, S.; Hafen, R.; Abusalah, A.; Yakout, M.; Ouzzani, M.; Cleveland, W. S.; Grannis, S. J.; and Ebert, D. S. 2009. A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, 16(2): 205–220.

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

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

Reinhart, A.; Brooks, L.; Jahja, M.; Rumack, A.; Tang, J.; Agrawal, S.; Al Saeed, W.; Arnold, T.; Basu, A.; Bien, J.; et al. 2021. An open repository of real-time COVID-19 indicators. *Proceedings of the National Academy of Sciences*, 118(51): e2111452118.

Sarikaya, A.; Correll, M.; Bartram, L.; Tory, M.; and Fisher, D. 2018. What do we talk about when we talk about dashboards? *IEEE transactions on visualization and computer graphics*, 25(1): 682–692.

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

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

Vajiac, C.; Chau, D. H.; Olligschlaeger, A.; Mackenzie, R.; Nair, P.; Lee, M.-C.; Li, Y.; Park, N.; Rabbany, R.; and Faloutsos, C. 2022. 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.

Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M. S.; and Krishna, R. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–38.

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

WHO. 2022. WHO Hub for Pandemic and Epidemic Intelligence. {https://pandemichub.who.int/publications/m/item/the-who-hub-for-pandemic-and-epidemic-intelligence-strategy-paper}. Accessed: 2023-06-05.