Data Monitoring for Large Scale Public Health Data

Ananya Joshi

Carnegie Mellon University, Pittsburgh, PA
aajoshi@andrew.cmu.edu

Abstract

Modern public health data contains information about changes in disease dynamics that can have significant downstream benefits, if these phenomena can be identified. However, systemic data quality issues hamper automated analysis of these vast data , and there is now far too much data (e.g. 3-4 million data points/day) for public health data experts to inspect manually as they may have done in the past. This interdisciplinary thesis addresses practical questions about large-scale data monitoring that impact public health data users and are also reflected in the larger public health community. This work has been deployed for over a year and a half at the Delphi Research Group at Carnegie Mellon University, a national public health data curator, where data reviewers have been able to detect approximately 200 significant outbreaks, data issues, or changes in disease dynamic from 15 million new data points weekly.

Motivation

Public health data holds significant potential for mitigating disease spread, but is still limited due to data quality considerations (Kraemer et al. 2021). One particular concern is identifying unexpected changes in data, which could either be due to notable fluctuations in public health dynamics or data quality issues. However, recent changes in public health data characteristics, namely large volumes of nonstationary, noisy, incomplete, revised, and seasonal data, (Reinhart, Brooks et al. 2021) make it challenging to detect unexpected data points using existing biosurveillance approaches.

These dynamics impact the Delphi Group at Carnegie Mellon University, which publishes makes vast amounts of modern public health data daily. Since the start of the pandemic, when Delphi's data volume increased by **1000x**, multiple data users have wanted Delphi to identify these unexpected data points, for example, so that they can be removed from modeling tasks.

To address this problem, I initially worked on a team with engineers to apply existing outlier and outbreak detection methods in a data monitoring system, where we would review the flagged, unexpected data points. However, these existing methods, developed nearly two decades ago (Shmueli and Burkom 2010), returned too many alerts for us to inspect, with large numbers of false positives/negatives and constant parameter tuning. Digging deeper into the matter, I found that the statistical characteristics of modern public health data were incompatible with the underlying assumptions of many existing methods. This insight inspired my thesis research question: "How can data reviewers monitor modern public health data streams?"

Current Progress

The [..] disconnect among algorithm developers, implementers, and users has [...] foster[ed] distrust in [...] biosurveillance – (Shmueli and Burkom 2010).

Identifying unexpected public health data is an interdisciplinary problem with statistical, computational, and public health data reviewer constraints. To that end, my doctoral work introduces a human-in-the-loop framework for human data reviewers monitoring modern public health data in real-time that is supported by Delphi's engineering team has been used for over a year by public health data reviewers.

First, I worked on a method to identify univariate outliers in this data (Joshi et al. 2023) meeting engineering and public health constraints by using simple, scalable models. Here, I built on the expertise of researchers at Delphi in data forecasting (Reinhart, Brooks et al. 2021) to address the statistical properties of public health data explicitly, like weekday effects and data lag, as shown in Steps 2 and 3 of our FlaSH method in Fig. 1. I also designed an interactive survey that went through many iterations of participatory design where several humans familiar with public health data ranked outlier data points that we used as ground truth. Because the proposed method met constraints and outperformed previous outlier detection methods (including recent deep learning baselines), it was initially deployed in January 2023, and we began to get real-time feedback from data reviewers.

However, we noticed that when we scaled this method across Delphi's millions of data streams, there were thousands of maximally-tied outliers that reviewers could not prioritize from. To help them distinguish the most important outliers from these ties, I developed a method that uses extreme value theory and the hierarchical structure of public health data to identify the most outlying points based on context (Joshi et al. 2024). In another offline survey evaluation, this algorithm performed the best across preregistered

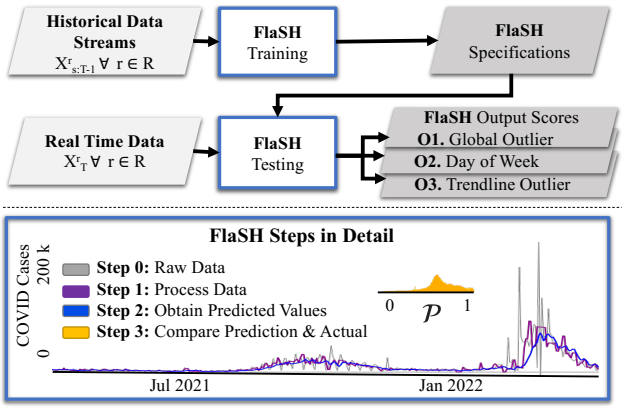

Figure 1: Forecasting advances in computational epidemiology enabled our univariate outlier detection method to be sensitive to some statistical properties of public health data. Image Reference (Joshi et al. 2023)

metrics. Then, I worked with the Delphi engineering team to deploy this method at scale, and data reviewers identified notable outliers **9.1x** faster than the prior approach.

Finally, we built a new interface for reviewers to analyze the ranked outliers. In a 3-month, longitudinal evaluation of the interface, reviewers showed significant improvements, including a total average **53**x efficiency increase over the manual approach. Experts also detected several types of patterns over the notable outliers. In total, this led to reviewers identifying approximately 200 significant outbreaks, data issues, or changes in disease dynamic from 15 million new data points weekly.

Large-scale systems like these are team efforts, and I grateful for the guidance of my advisors and collaborators, especially Bryan Wilder, Roni Rosenfeld, Nolan Gormley, Tina Townes, Richa Gadgil, Katie Mazaitis, Luke Neureiter.

Proposed Work

The following ideas that are important to different domains:

1. **Methodological:** Advances in foundational time series and language models can improve algorithmic performance and automatically extract essential insights from user interactions and annotations. However, these approaches primarily have not yet been tested on epidemiological data or public health contexts (e.g., the limitations on multi-variate foundational models), which is concerning as public health data exhibits high nonstationarity and rapid context changes that may be incompatible with these methods (Su et al. 2024). Determining if the potential improvement in forecasting/summarizing capabilities of these approaches is appropriate, given computational cost and variance in performance, is a topic of interest for stakeholders.
2. **Computational:** Public health data is also being updated more frequently (Simonsen et al. 2016). Based on these trends, we anticipate stakeholders may want updated outlier rankings on the order of hours instead of days (the prior update frequency) and may be willing to make tradeoffs between time and accuracy. Identifying guidelines for these tradeoffs would help practicioners with different timeliness constraints.
3. **Public Health:** Data users often see sequences of data points that are jointly concerning (v.s. individual outliers). Instead of inspecting each outlier in an anomalous sequence separately, ranking anomalous sequences that may be abrupt, gradual, incremental, or recurrent (Gomes et al. 2017) could support reviewers in identifying patterns over the events quickly.

We hope that bridging these perspectives from multiple communities will enable shared support for strengthening public health infrastructure in a modern data era.

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

Gomes, H. M.; Bifet, A.; Read, J.; et al. 2017. Adaptive random forests for evolving data stream classification. *Machine Learning*.

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

Joshi, A.; Rosenfeld, R.; Wilder, B.; et al. 2024. Outlier Ranking for Large-Scale Public Health Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Kraemer, M. U.; Scarpino, S. V.; Marivate, V.; et al. 2021. Data curation during a pandemic and lessons learned from COVID-19. *Nature Computational Science*, 1(1): 9–10.

Reinhart, A.; Brooks, L.; et al. 2021. An open repository of real-time COVID-19 indicators. *Proceedings of the National Academy of Sciences*, 118(51): e2111452118.

Shmueli, G.; and Burkom, H. 2010. Statistical challenges early outbreak detection in biosurveillance. *Technometrics*.

Simonsen, L.; Gog, J. R.; Olson, D.; and Viboud, C. 2016. Infectious disease surveillance in the big data era. *The Journal of infectious diseases*.

Su, C. J. X., J.; Jiang; et al. 2024. Large language models for forecasting and anomaly detection: A systematic literature review. *arXiv preprint arXiv:2402.10350*.