Drug Resistance Updates 78 (2025) 101174

Contents lists available at ScienceDirect

Drug Resistance Updates

journal homepage: www.elsevier.com/locate/drup

AI-driven visualization tool for analyzing data and predicting drug-resistant outbreaks

ARTICLE INFO

Keywords:
Outbreak visualization
Pathogen and serotype
Generative AI
CDC dataset

ABSTRACT

A tool was developed to identify potential disease outbreaks using pathogen and serotype data. By analyzing isolate numbers and comparing them to a two-year average, the tool highlights anomalies suggestive of outbreaks. When applied to Salmonella data, it revealed potential outbreaks related to specific serotypes.

The BEAM Dashboard, standing for Bacteria, Enterics, Amoeba, and Mycotics is a dynamic platform designed to retrieve and illustrate data from the System for Enteric Disease Response, Investigation, and Coordination (SEDRIC) operated by the Centers for Disease Control and Prevention (CDC) (Katz et al., 2024). The dashboard delivers up-to-date information on the patterns of pathogens and details of serotypes to aid in the prevention of diseases caused by food and animal exposure. At present, the dashboard concentrates on data related to Salmonella, Shiga toxin-producing E. coli (STEC), Shigella, and Campylobacter bacteria, along with antimicrobial resistance and outbreaks occurring in multiple states. It offers up-to-date information on the patterns of pathogens and specifics of serotypes, serving as a valuable resource for initiatives aimed at preventing diseases caused by food and animal, and environment interactions (CDC, 2024; EFSA, 2024).

The CDC has released a dataset on pathogens and serotypes, available online. This paper presents a tool for researchers to study pathogen behaviors and identify serotype outbreaks. The dataset includes 'Number of isolates' and monthly '% Change' against a two-year average baseline.

'Number of isolates' refers to the count of pathogen samples isolated for study. Each 'isolate' originates from a host, the environment, or a disease. An isolate is akin to a germ family derived from one original germ. The 'Number of isolates' is the count of these germ families, crucial for disease study and treatment development. For instance, the NCBI database has over 300,000 pathogen isolates aiding in studying drug resistance.

Tracking the 'Number of isolates' over time can reveal pathogen and serotype patterns. It represents the count of isolated pathogens in a given period. Observing this number can help understand the prevalence and spread of these pathogens or serotypes.

In a comprehensive study spanning multiple countries, researchers gathered over 400 isolates of the invasive pathogen, Pseudomonas aeruginosa, from a variety of locations (Nasrin et al., 2022). The objective was to ascertain the distribution of serotypes, identify flagellin types, and assess patterns of antibiotic susceptibility. Their study underscored the significance of tracking the number of isolates, which can provide valuable insights into the epidemiology of a pathogen and inform strategies for vaccine development.

The '% Change' compared with 2-year average as baseline can help identify anomalies and their duration. It represents the rate at which the 'Number of isolates' increases or decreases for a particular pathogen or serotype from one time period to another. A significant increase (% change) could indicate an outbreak or surge of a particular pathogen or serotype, while a significant decrease could indicate a successful containment or decrease in the prevalence of the pathogen or serotype. The duration of these changes can also provide insights into the longevity and impact of these anomalies. An outbreak is defined as an increase in events, such as infections or number of organisms above the baseline rate during a specified period of time (Sood et al., 2016).

So, both these metrics, 'Number of isolates' and '% Change' against the baseline of 'Past two years average', provide valuable insights but from different perspectives. They can be used together to get a comprehensive understanding of the situation for detecting outbreaks (Moritz et al., 2023; Du et al., 2023).

The proposed tool allows users to create two kinds of a graph: 'Number of isolates' and '% Change' against the baseline of 'Past two years average'.

This paper presents two outbreak cases caused by the Salmonella pathogen, specifically focusing on two serotypes: Sundsvall and Africana. Download the CDC dataset (CDC, 2024) and rename it to `data.csv`. The final Python code, pasero.py, created by generative AI (Copilot) and available through the PyPI package pasero, can be accessed at https://github.com/y-takefuji/pasero (GitHub, 2024). Copilot is a code assistant that leverages OpenAI's GPT model to generate functional code from natural language prompts.

According to the [CDC site](https://www.cdc.gov/salmonella/outbreaks.html), a salmonella outbreak in 2023 was reported due to Cantaloupes. The outbreak resulted in 407 illnesses, 158 hospitalizations, 6 deaths, and affected 44 states. The bold line in Figs. 1–1 indicates that a peak in the change of Salmonella Sundsvall was observed around November 2023. Early symptoms appeared that could potentially be utilized for predicting future outbreaks.

Salmonella Africana found in Cucumbers caused 449 illnesses (38 new Africana, 215 Braenderup new), 125 hospitalizations (11 new Africana, 50 Braenderup new), and 0 deaths. The outbreak spread across 31 states and the District of Columbia (1 new Africana, 26 Braenderup).

https://doi.org/10.1016/j.drup.2024.101174
Received 16 July 2024; Received in revised form 18 November 2024; Accepted 18 November 2024
Available online 19 November 2024

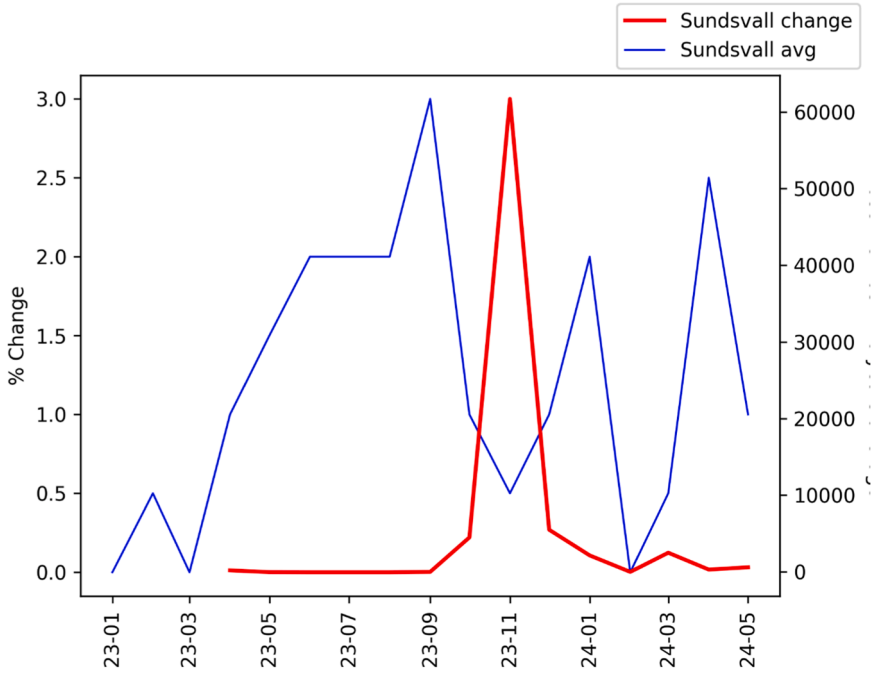

Fig. 1–1. Salmonella Sundsvall outbreak detection around November 2023.

Similarly, `pasero.py` can generate a graph with the bold line in Fig. 1–2 showing a steep surge in the percentage change from March 2024 to present.

This study used Python-based data analysis and the BEAM Dashboard to identify potential outbreaks from pathogen and serotype trends. The "% Change" metric, calculated from a two-year average, highlighted anomalies for further investigation. This tool could form an early warning system, but its effectiveness depends on the quality of the underlying data and should be used alongside other epidemiological and clinical data.

The current analysis was limited to certain pathogens and serotypes. A broader range of microorganisms and geographical information could provide a more comprehensive view of disease trends and help identify regional outbreaks. Future research should integrate advanced statistical methods and machine learning to enhance detection capabilities. Real-time surveillance systems with human validation could improve this approach's effectiveness.

Correlations between outbreaks and specific food products or environmental factors could inform targeted prevention strategies. While promising, this preliminary exploration of data analytics for outbreak detection requires further research to become a reliable public health surveillance tool. Addressing limitations and exploring enhancements

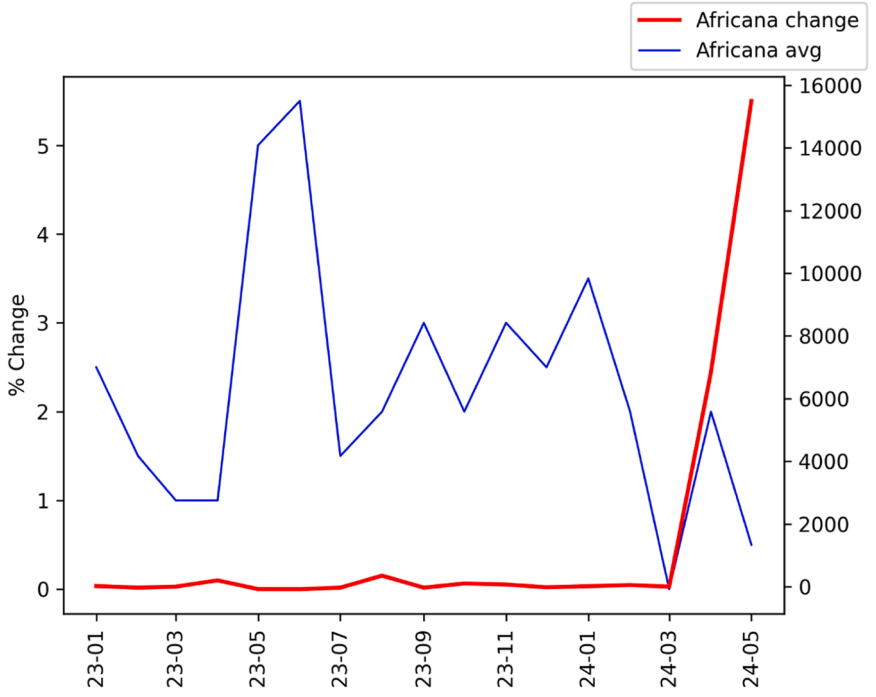

Fig. 1–2. Salmonella Africana detection from March 2024.

can contribute to developing effective early warning systems for infectious diseases.

CRediT authorship contribution statement

Yoshiyasu Takefuji completed this research and wrote the program and this article.

Funding

This research has no fund.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. : Methods

Generative AI is used to create Python code as a tool to achieve the desired outcome. The tool has two modes: one to illustrate raw data of 'Number of isolates' of pathogens and serotypes, and another to illustrate '% Change' against the baseline of 'Past two years average'. Multiple conversations are needed to communicate with Generative AI to finalize the code which is publicly available at https://github.com/y-takefuji/pasero (GitHub, 2024). The final code, pasero.py, designed for colorblind individuals, was generated using Microsoft's Copilot and generative AI. To begin, download the dataset from the CDC website (CDC, 2024) and rename it to data.csv. If you would prefer to use colored lines, such as blue and red, you can utilize pasero2.py, which was specifically developed for this purpose. You can inquire about color options with generative AI, or simply modify the color settings manually by changing color='red' or color='blue' for your desired color updates.

Initial query to AI: show Python code with data.csv to draw a graph with up to 8 black lines

using 4 line-style: solid, dotted, dash, and dashdot with 2 widths (1,2), total 8 unique black lines. Combining 'Year' and 'Month' variables indicates X-axis. Sort year-month data.

show unique 'Pathogen' values and user is allowed to select one by number.

show unique 'Serotype/Species/Subgroup' values and user is allowed to select

up to 4 items by number with separated by space.

Show two choices and user is allowed to select one by number either 'Number of isolates' for 1 line or '% Change' with 'Past two years average' for 2 lines as Y-axis values. If 'Number of isolates' is selected, Y-axis is entitled 'Number of isolates' else '% Change' on left Y-axis and 'Past two years average' on right Y-axis. sort data and show the graph with rotating 90 degrees in X-axis labels such as year+month. save csv file and drawn figure (png) as pathogen+serotype name.

Before running the command, ensure that Python is installed on your system. Execute the following command to run `pasero.py` or install PyPI pasero.

```
$ python pasero.py
OR
$ pip install pasero
$ pasero
```

The program will interactively present a list of pathogens:
Pathogens:

1. Campylobacter
2. STEC
3. Salmonella
4. Shigella
5. Vibrio

Select a pathogen by number:
For Salmonella, select `3`. The program will then display all serotypes. Enter `33` for Sundsvall. Next, select `2` to generate a graph of % Change with baseline of Past two years average as shown in Figs. 1–1.

1. Number of isolates
2. Past two years average and % Change

Data availability statement

The author has no permission to share data.

References

CDC. BEAM Dashboard – Top 30 Most Common Serotypes, released on June 14, 2024. ⟨https://data.cdc.gov/Foodborne-Waterborne-and-Related-Diseases/BEAM-Dashboard-Top-30-Most-Common-Serotypes/ch83-ush6/about_data⟩.

Du, Z., Shao, Z., Zhang, X., Chen, R., Chen, T., Bai, Y., Wang, L., Lau, E.H.Y., Cowling, B. J., 2023. Nowcasting and forecasting seasonal influenza epidemics - China, 2022-2023. China CDC Wkly. 5 (49), 1100–1106. https://doi.org/10.46234/ccdcw2023.206.

EFSA Panel on Biological Hazards (BIOHAZ), Koutsoumanis, K., Allende, A., et al. (2024). Persistence of microbiological hazards in food and feed production and processing environments. EFSA journal. European Food Safety Authority, 22(1), e8521. https://doi.org/10.2903/j.efsa.2024.8521.

GitHub. (2024). pasero.py and PyPI: pasero. https://github.com/y-takefuji/pasero.

Katz, T.S., Harhay, D.M., Schmidt, J.W., Wheeler, T.L., 2024. Identifying a list of Salmonella serotypes of concern to target for reducing risk of salmonellosis. Front. Microbiol. 15, 1307563. https://doi.org/10.3389/fmicb.2024.1307563.

Moritz, E.D., Ebrahim-Zadeh, S.D., Wittry, B., Holst, M.M., Daise, B., Zern, A., Taylor, T., Kramer, A., Brown, L.G., 2023. Foodborne Illness Outbreaks at Retail Food Establishments - National Environmental Assessment Reporting System, 25 State and Local Health Departments, 2017-2019. Morb. Mortal. Wkly. Report. Surveill. Summ. (Wash., D. C.: 2002) 72 (6), 1–11. https://doi.org/10.15585/mmwr.ss7206a1.

Nasrin, S., Hegerle, N., Sen, S., et al., 2022. Distribution of serotypes and antibiotic resistance of invasive Pseudomonas aeruginosa in a multi-country collection. BMC Microbiol 22, 13. https://doi.org/10.1186/s12866-021-02427-4.

Sood, G., Perl, T.M., 2016. Outbreaks in health care settings. Infect. Dis. Clin. North Am. 30 (3), 661–687. https://doi.org/10.1016/j.idc.2016.04.003.

Yoshiyasu Takefuji[1]
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan
E-mail address: takefuji@keio.jp.

[1] **ORCID:** 0000–0002-1826–742X