Bacteria, Enterics, Amoeba, and Mycotics Tracking: EDA Final Report

Sophia Chung, Nikki La, Rishabh Shah

December 12, 2024

1 Introduction

Enteric illnesses are a group of diseases that primarily affect the gastrointestinal system, resulting from the ingestion of contaminated food or water[**1**]. Common symptoms of enteric illnesses include nausea, vomiting, diarrhea, abdominal cramps, fever, chills, and loss of appetite. These symptoms can appear 30 minutes to 10 days after contact. Numerous agents, including bacteria (like *Salmonella* and *Escherichia coli*), viral agents (like Rotavirus and Norovirus), and parasite species (like *Giardia duodenalis* and *Cryptosporidium*) can cause enteric diseases[**2**]. Significant public health issues are also raised by additional illnesses that have been connected to fungus, waterborne, and foodborne sources. Aside from these sources, enteric illnesses can spread through person-to-person contact, animal contact, and contact with environmental contamination[**3**].

Effective prevention and management of enteric illnesses depend on data collection, analysis, and reporting. At the Centers for Disease Control and Prevention (CDC), surveillance systems such as the System for Enteric Disease Response, Investigation, and Coordination (SEDRIC) play a large role in understanding pathogen trends and serotype variations by streamlining outbreak investigations through integrating epidemiological, laboratory, and traceback data in real time[**4**]. This comprehensive approach enables investigators across the United States to collaborate effectively. Similarly, the National Antimicrobial Resistance Monitoring System for Enteric Bacteria (NARMS), also under the CDC, monitors trends in antimicrobial resistance in foodborne and other enteric bacteria[**5**]. Additionally, the National Outbreak Reporting System (NORS), collects reports of enteric disease outbreaks caused by bacterial, viral, parasitic, chemical, toxin, and unknown agents, as well as foodborne and waterborne outbreaks of non-enteric disease[**3**]. NORS gathers data on outbreaks transmitted through various routes, including water, food, animal contact, person-to-person transmission, and environmental contamination. This data include the date and location of outbreaks, the number of affected individuals and their symptoms, and the pathogens involved. Together, these systems provide a robust framework for understanding and responding to enteric illnesses.

The CDC's BEAM Dashboard offers real-time, automated analyses using data from systems like SEDRIC and NORS, providing critical insights into trends and patterns in enteric illnesses. Designed to support public health officials, researchers, and industry professionals, the dashboard enables timely decision making and targeted interventions. However, its reliance on interactive exploration can make it challenging for specific users to fully interpret the findings or extract actionable insights. This highlights the need for complementary approaches that present the data in a more accessible and concise manner, ensuring that it will reach a diverse and broader audience base.

This project aims to reframe the analyses presented in the BEAM Dashboard by conducting exploratory data analysis and applying key principles of data storytelling. We aim to present the data and insights through clear narratives and intuitive static visualizations, making them more accessible and impactful to improve public understanding and support public health education and decision-making efforts.

2 Defining Key Terms

Terminology used to report enteric illnesses, outbreaks, and isolates is often misconstrued and falsely conflated, often leading to confusion and misinterpretations of public health communication and data analysis. To clarify, the term **outbreak** refers to the incidence of two or more cases of similar illness originating from a common source[**6**]. For example, in the case where two or more people fall ill from the same contaminated food or drink, the incidence would be reported as a foodborne disease outbreak.

The term **pathogens** refers to any organism that can cause disease, including some types of bacteria such as *Salmonella* and some strains of *E. coli*[**5**]. The term **isolate** refers to a group of the same type of bacteria[**6**]. Isolates can originate from the environment, food, animals, and other sources. Whole genome sequencing is a method that can be used in laboratories to ascertain information regarding such bacteria and how they genetically relate to other bacteria. Importantly, it is possible to report isolates and singular cases of illnesses that do not necessarily result in enteric illness outbreaks. By monitoring individual cases, public health officials can identify potential emerging outbreaks before they become widespread.

The term **serotype** refers to groups in a single species of microorganisms. Salmonella, a previously mentioned isolate, has many serotypes such as Newport, Heidlberg, or [5],12:i: -. Some serotypes are found in specific species of animals or locations, while others can be found in numerous animals throughout the world. Some serotypes can cause serious illnesses, while others may cause milder illnesses. Although Salmonella serotypes may look similar under the microscope, there are structural differences that allow scientists to further classify them into serotypes. Further classifying isolates as serotypes can help scientists better understand illnesses caused by strains and track how these strains evolve over time[**7**].

3 Motivation

The purpose of the CDC's BEAM Dashboard is to provide up-to-date and interactive analyses using data collected through SEDRIC and NORS. Interactive visualizations such as the BEAM Dashboard can be useful due to the many variables and observations examined in the analyses and the fluidity of the data as time passes. The CDC created the BEAM Dashboard to provide the public, academia, industry, health officials, and regulatory agencies with real-time, automated insights to help prevent future enteric illnesses. However, the dashboard's reliance on interactive exploration can make it difficult for laypeople to interpret and extract meaningful insights. Although the BEAM Dashboard excels in delivering real-time, automated analyses, there is an opportunity to present the data in a more accessible and story-driven manner, tailored to non-expert audiences.

In this project, we are accessing and conducting exploratory data analysis on the data analyzed in the BEAM Dashboard ourselves. We are interested in digging deeper into the data and reframing the analyses using principles of data storytelling. By presenting clear narratives and static and intuitive visualizations, we aim to make the data and its implications for public health more interpretable for broader audiences. This approach allows us to uncover trends, patterns, and caveats within the data while contextualizing their relevance with public health. By bridging the gap between technical data and accessible communication, we hope this exploratory data analysis can contribute to public health education and awareness, empowering individuals and communities to better understand and respond to enteric illness risks.

4 Data Cleaning and Preparation

To conduct this analysis, we made several assumptions to guide our exploration of the data. The data from the BEAM Dashboard begin in 2018 and end with quarter 3 of 2024.

The following Python libraries were employed in this project for various purposes:

1. Pandas: Used for loading, cleaning, transforming, and analyzing tabular data.
2. NumPy: Utilized for numerical computations and handling missing data efficiently.
3. Matplotlib and MPL Toolkits: Applied for creating static, publication-quality visualizations and advanced plotting.

4. Seaborn: A statistical data visualization library used to simplify the creation of aesthetically pleasing and informative plots.

5. Pillow (PIL): Imported for any image processing tasks, likely used in handling or displaying visual content related to the analysis.

6. GeoPandas: Used for geospatial data manipulation and mapping as the analysis would involve geographical trends or mappings.

7. us: Provides easy access to standardized information about U.S. states and territories, including names, abbreviations, FIPS codes, and other metadata.

8. Warnings Module: Enabled the suppression of unnecessary warnings for a cleaner coding and debugging experience.

Data preparation involved several essential steps to make the primary BEAM Dashboard dataset suitable for analysis and visualization. First, the data were loaded from a CSV file and previewed. An initial preview of the data allowed for an understanding of its structure, including columns like year, month, state, source, pathogen, and serotype/species. Next, we dropped unnecessary columns with excessive missing values and removed columns irrelevant to the current analysis, such as outbreak associated isolates, new multistate outbreaks, and percent of isolates with clinically important antimicrobial resistance. Missing data were handled strategically, primarily through deletion, ensuring the integrity of subsequent analyses. The original data contained rows for Washington, D.C., but we chose not to include those rows due to the district not possessing statehood. Columns were renamed for clarity, and their order was adjusted to enhance interpretability in the visualizations. Transformation steps, such as converting months to categorical data or aggregating counts by year, were performed to prepare for visual storytelling. The data were filtered to focus on relevant subsets, such as specific pathogens, states, or time frames (year, quarter, month). Aggregations such as sum or mean were used to simplify trends. These steps ensured a clean and structured dataset, enabling the creation of accessible and impactful static visualizations to convey meaningful insights about enteric illnesses.

Geospatial data preprocessing involved meticulous adjustments to ensure accurate and visually clear mapping. Using geopandas and matplotlib, the process began by filtering out unincorporated U.S. territories, such as Guam, Puerto Rico, and American Samoa, which were unnecessary for the analysis. The data's coordinate reference system (CRS) was transformed to "ESRI:102003" to align the mapping projection with best practices for U.S. geospatial data visualization. Additionally, separate preprocessing steps were implemented to handle Hawaii and Alaska, isolating them to adjust their geometries for better placement and scale. These steps collectively enhanced the geographic representation and readability of the final visualizations.

Preliminary exploratory analyses of the original BEAM Dashboard dataset revealed that the majority of illnesses were associated with Salmonella, prompting a decision to investigate further into Salmonella serotypes and their connection to foodborne illnesses. A careful review of the Salmonella-derived foodborne illnesses dataset revealed overlapping year ranges, including 2011–2015, 2016–2020, 2012–2016, and 2017–2021. Notably, the 2012–2016 and 2017–2021 rows duplicated some data already included in the broader 2011–2015 and 2016–2020 ranges. To ensure consistency and avoid redundancy, these duplicated subsets were removed, retaining only the 2011–2015 and 2016–2020 rows, as they contained more comprehensive data. These preprocessing decisions eliminated unnecessary overlaps and enabled a clearer focus on Salmonella-related insights within the chosen time frames.

5 BEAM Dashboard Analysis

5.1 Time Series Analysis

Figure 1 reveals interesting trends over the time period from 2018 to 2024. In 2018 and 2019, the total number of isolates was relatively high, exceeding 16,000 each year. However, in 2020 there was a noticeable drop to around 14,500 isolates. This could potentially be related to changes in data collection or reporting during the COVID-19 pandemic that year. This drop could also reflect a limited spread of the specific pathogens that cause enteric illnesses due to reduced contact between humans during the year.

After the 2020 decline, the number of isolates rebounded in 2021 and has continued rising steadily through 2023, reaching over 17,000. As of the end of September 2024, the total pathogen isolates for the 2024 year show a slight decrease compared to 2023, but the total isolates are still projected to remain high compared to the early years of data. This information could be valuable for public health officials and researchers tracking trends in the prevalence of pathogen isolates nationwide.

Figure 2 displays the number of isolates categorized by four source types: stool, urine, blood, and other. This breakdown of source types provides a more nuanced view of the data compared to the total isolate counts shown in Figure 1. Stool samples consistently account for the largest proportion of isolates throughout each year, a finding supported by research and current understandings of enteric illnesses[**8**]. Thus, primary protection and prevention of enteric infections include improving water and sanitation to decrease transmission of enteric pathogens via improved personal and community hygiene.

Aside from stool, the contributions from other source types are also noteworthy. Urine isolates make up the second-largest component, likely stemming from urinary tract infections caused by pathogens like E. coli[**9**]. Blood isolates, while smaller in number, still represent a significant source of data that could be indicative of more serious systemic infections.

Examining year-over-year changes, there are similar pattern to the total isolate counts discussed previously. There is a dip in 2020, potentially due to pandemic-related disruptions, followed by a rebound and continued growth in subsequent years. Interestingly, the relative proportions between source types remain fairly consistent, suggesting that underlying epidemiological trends may not have shifted dramatically despite the fluctuations in total numbers.

It is important to note once again that the 2024 data is incomplete, including data up to the end of September 2024. Caution should be exercised when drawing conclusions about the 2024 trends until the full-year data becomes available.

This breakdown by source type provides a more comprehensive understanding of the pathogen landscape, highlighting the contributions of other infection sources. This information can help public health officials fine-tune their surveillance, prevention, and response strategies to address the diverse origins of these pathogens.

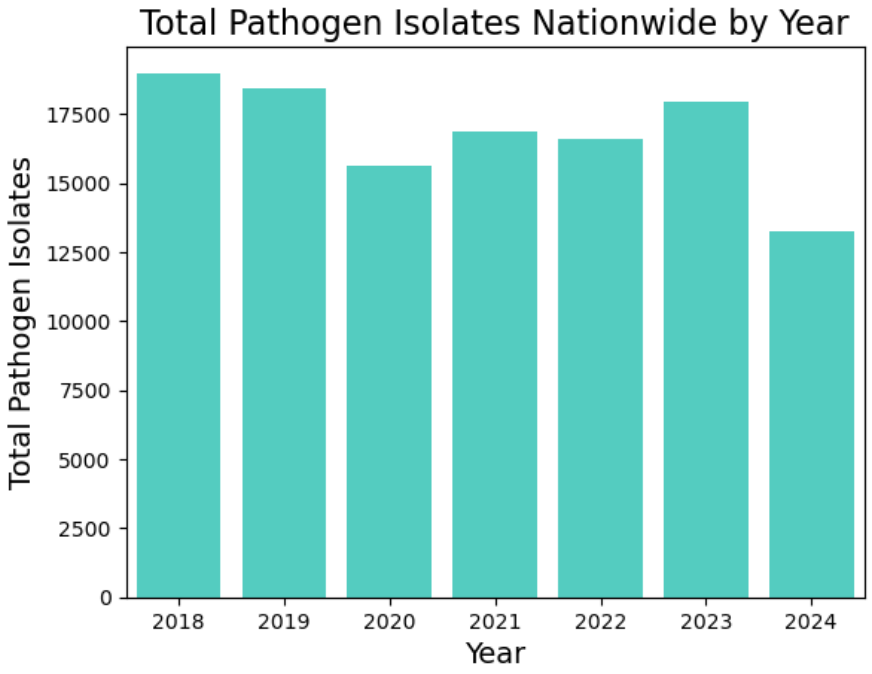

Figure 1: This is a bar graph of the total pathogen isolates throughout the United States by year. The x-axis shows each year of data (January 2018 to the end of September 2024). The y-axis shows the total pathogen isolates. There are slight fluctuations in the data from year to year and a notable dip in pathogen isolates during the year 2020.

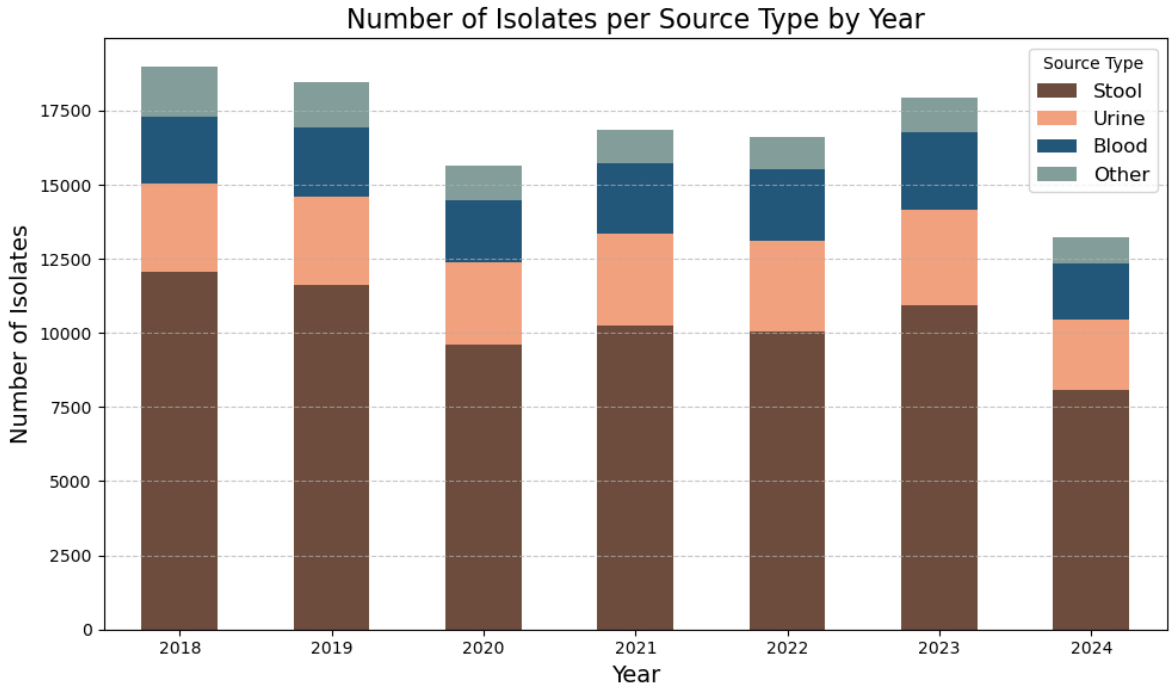

Figure 2: The stacked bar chart shows the number of isolates per source type from 2018 to 2024. The x-axis represents the years, while the y-axis shows the total number of isolates. The isolates are categorized into four source types: stool, urine, blood, and other. Stool isolates consistently account for the majority of the total isolates each year, with smaller contributions from urine, blood, and other sources. It is important to note that the data for 2024 is incomplete, which likely explains its lower total compared to previous years.

Figure 3 clearly depicts a cyclical pattern in the total pathogen isolates over the course of a year. The numbers start off relatively low in the winter months (December to February), with the lowest point typically seen in February. From there, the isolate counts begin to rise steadily, reaching a distinct peak in the summer months of July and August. Figure 4 displays the data broken down by quarter, showing how the numbers fluctuate throughout each year. Consistent with the Figure 3, the isolate counts consistently peak in the 3rd quarter (Q3) of each year, typically reaching the highest levels in July and August. This aligns with the understanding that warmer summer months tend to facilitate the spread of many foodborne and enteric pathogens.

Conversely, Figure 3's isolate counts regularly dip to their lowest points in the first quarter (Q1) of each year, reflecting Figure 3's nadir in February. This winter decline likely reflects the impact of colder temperatures, changes in human behaviors and food consumption patterns, and other seasonal factors that make the environment less conducive to pathogen transmission during those months.

Looking at year-over-year trends in Figure 4, there are some notable variations, such as the sharper Q3 peak in 2020 compared to other years. However, the overall cyclical pattern of Q3 highs and Q1 lows persists across the entire date range shown.

After the summer peak, pathogen isolate numbers consistently decline through the fall, hitting another low point in December before the cycle repeats. This winter dip may be tied to colder weather, less time spent outdoors, and changes in dietary habits that occur during the holiday season. This cycle aligns with scientific understanding regarding the seasonality of many enteric illnesses and foodborne pathogen outbreaks[**10**]. Warmer temperatures, increased outdoor activities, and changes in food consumption patterns during the summer months can create environments more conducive to the spread of certain pathogens. Factors such as humidity, rainfall, and changes in human behaviors likely all play a role in driving these seasonal trends. Illnesses from salmonella, the pathogen responsible for the majority of enteric illnesses analyzed in this report, are more common in the summer due to perfect growing conditions created by warmer weather and unrefrigerated foods.

Infectious enteric illness outbreaks analyzed here are particularly difficult to track and control given the complex networks found in food systems today. In epidemiological and infectious disease research,

scientists often note similar season co-occurrences and patterns. However, these phenomena are often underresearched due to limitations in complexity and methodology[**11**].

This granular, monthly and quarter-by-quarter view provides valuable insights that can help public health officials anticipate and prepare for seasonal surges in pathogen activity. By understanding these predictable fluctuations, they can optimize surveillance, prevention, and response efforts to minimize the impact of enteric illnesses throughout the year. The data can also inform educational campaigns to raise awareness among the public about seasonal pathogen risks and recommended mitigation strategies.

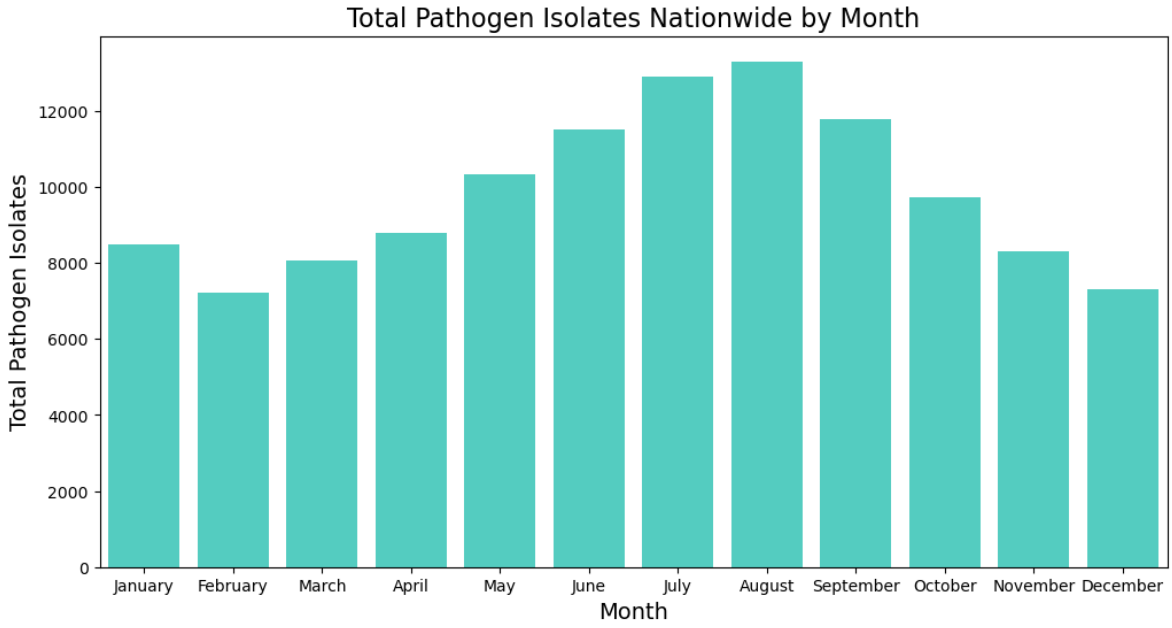

Figure 3: This is a bar graph of the total pathogen isolates throughout the United States by month. The x-axis shows each month of data. The y-axis shows the total pathogen isolates. Throughout the years, pathogen isolate counts peak in July and August and dip in February and December. The data highlights clear seasonal trends, which may be influenced by environmental factors such as temperature and humidity or seasonal human behaviors.

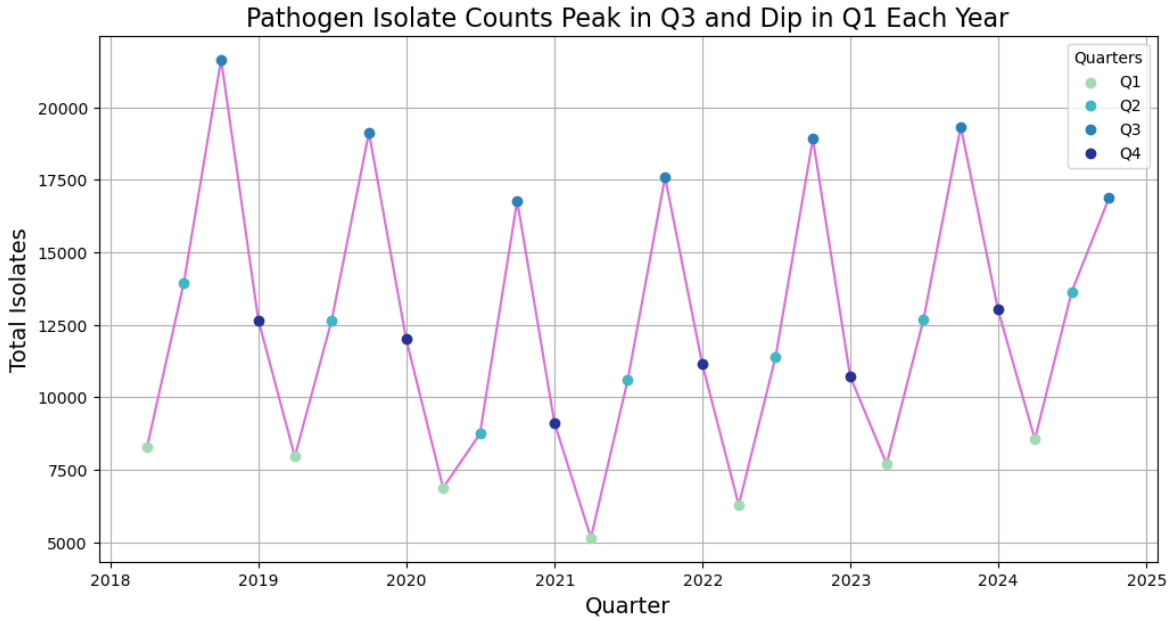

Figure 4: This is a line graph of the total pathogen isolates throughout the United States by quarter (Q1 indicating January 1st to March 31st, Q2 indicating April 1 to June 30, and so on). The x-axis shows each quarter of each year 2018-2024. The y-axis shows the total pathogen isolates. Throughout each year, pathogen isolate counts peak in Q3 and dip in Q1 each year.

5.2 Time-Series Analysis of States with Highest Populations

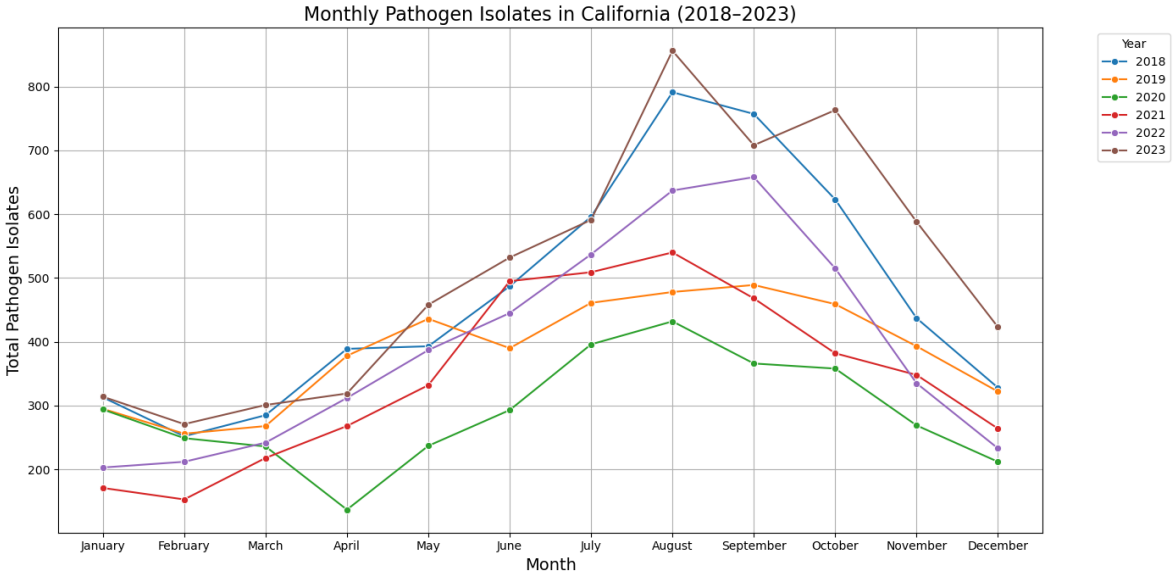

Figure 5: This line graph shows the monthly total of pathogen isolates in California from 2018 to 2023. The x-axis represents the months of the year. The y-axis shows the total pathogen isolates. Pathogen isolates consistently peak during the summer months (particularly in August) and decrease sharply towards the end of the year. Each year follows a similar seasonal trend, with 2023 showing a notably higher peak compared to previous years.

Figure 5 highlights the monthly trends in pathogen isolates in California from 2018 to 2023, illustrating a clear and consistent seasonal pattern. The isolate counts start relatively low during the winter months, particularly in December and February, reflecting the impact of colder temperatures, which

inhibit pathogen growth and transmission. During this time, individuals are more likely to prepare and consume food in controlled environments, reducing the risk of contamination. As spring progresses and temperatures rise, isolate counts begin to increase steadily, reaching a pronounced peak in August. This summer peak aligns with California's warmest months, during which outdoor activities such as picnics and barbecues are more common. These gatherings often involve less stringent food safety practices, such as improper storage or handling of perishable items, creating favorable conditions for pathogen transmission.

Interestingly, the year 2023 exhibits a notably higher peak compared to previous years. This increase could be attributed to several factors, such as heightened pathogen prevalence, improved surveillance systems, or increased human interactions following the pandemic-related restrictions of earlier years. After the peak in August, isolate counts decline steadily through the fall, returning to low levels by December. This cyclical trend underscores the seasonal nature of pathogen activity in California, driven by a combination of environmental conditions and behavioral patterns.

Public health officials can leverage these insights to implement targeted interventions during the summer months, such as educational campaigns emphasizing proper food handling and storage, increased monitoring of high-risk foods, and enhanced pathogen surveillance. These measures could significantly reduce the burden of enteric illnesses during the peak season in California.

Figure 6 reveals a similar seasonal pattern for Texas, where pathogen isolate counts begin at their lowest levels during the winter months, particularly in December. Despite Texas's relatively mild winters, the cooler temperatures during this period still contribute to a reduction in pathogen transmission. The spring months mark the beginning of a steady increase in isolate counts, peaking in late summer, typically in August and September. This late-summer peak reflects Texas's prolonged warm season, characterized by high temperatures and humidity, which create ideal conditions for pathogen survival. Social factors, such as the prevalence of outdoor events, fairs, and festivals during these months, further amplify the risk of foodborne pathogen exposure.

One notable anomaly in the data is the year 2019, which exhibits a less pronounced peak compared to other years. This deviation may be linked to environmental factors, such as hurricanes disrupting food systems and healthcare access, or inconsistencies in public health reporting during that period. Following the late-summer peak, the isolate counts decline steadily through the fall, returning to low levels by December.

This consistent seasonal cycle emphasizes the importance of directing public health resources toward the summer months, when pathogen activity is at its highest in Texas. Educational initiatives, stricter food safety regulations, and improved surveillance during this period could help mitigate the risks associated with enteric illnesses. Additionally, examining anomalies like 2019 can provide valuable insights into how external factors influence pathogen trends and highlight areas for improvement in data collection and analysis.

Figure 7 provides an analysis of monthly pathogen isolates in Florida from 2018 to 2023, revealing a slightly different seasonal pattern compared to California and Texas. In Florida, isolate counts peak earlier in the year, typically in June and July. This earlier peak can be attributed to Florida's subtropical climate, where high temperatures and humidity dominate the early summer months, creating optimal conditions for pathogen proliferation. Additionally, Florida experiences a significant influx of tourists during the summer, leading to increased food handling and preparation in high-traffic areas, which raises the likelihood of contamination and transmission.

The years 2018 and 2020 stand out with notably higher peaks, suggesting periods of heightened pathogen activity or improvements in surveillance and reporting systems during those years. These peaks may also reflect environmental factors, such as unusually high rainfall or extended periods of warm weather, which further promote pathogen growth. After the early summer peak, isolate counts decline steadily through the fall, reaching their lowest levels in December. This decline is driven by cooler temperatures, reduced humidity, and changes in human behaviors, such as a shift toward indoor dining and home-prepared meals during the fall and winter months.

Florida's earlier seasonal peak highlights the need for proactive public health strategies tailored to the state's unique climatic and demographic conditions. Initiatives such as targeted food safety campaigns, increased inspections of high-risk food establishments, and educational outreach to tourists and residents could help mitigate the risks associated with pathogen transmission during the early summer months. These efforts are especially critical in a state where tourism plays a significant role in the economy and increases the potential for widespread exposure.

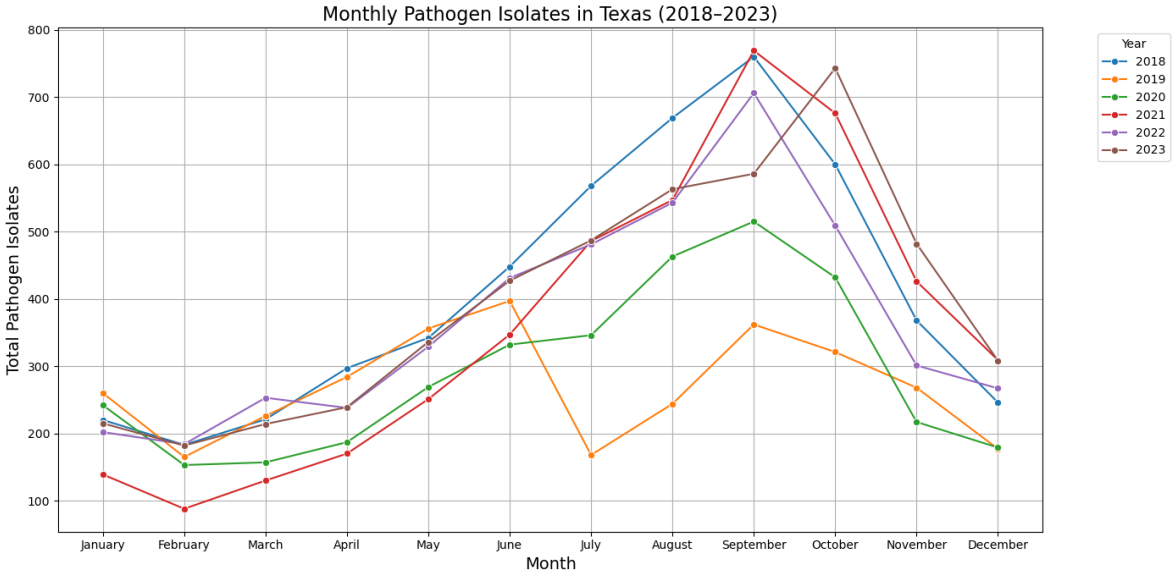

Figure 6: This line graph depicts the monthly total of pathogen isolates in Texas from 2018 to 2023. The x-axis represents the months of the year. The y-axis shows the total pathogen isolates. Pathogen isolates in Texas generally peak around late summer to early fall, particularly in August and September, before declining significantly toward the end of the year. However, the year 2019 deviates from this pattern, with lower peaks and a more gradual rise compared to other years.

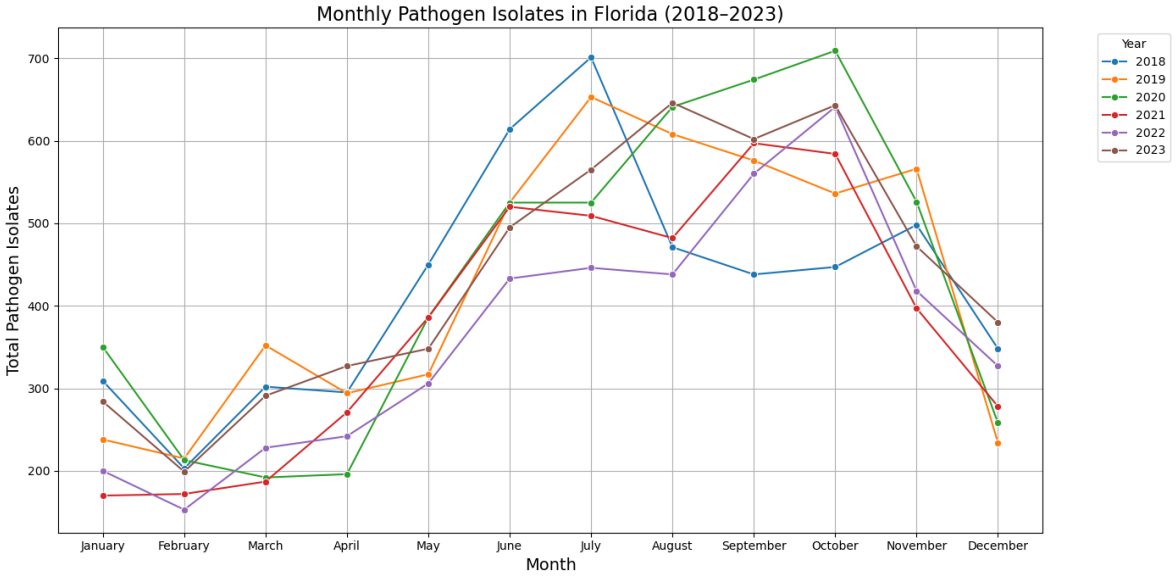

Figure 7: The line graph shows the monthly total of pathogen isolates in Florida from 2018 to 2023. The x-axis represents the months of the year. The y-axis shows the total pathogen isolates. Pathogen isolates in Florida peak during the summer months, particularly in June and July, before declining steadily through the fall and winter. The years 2018 and 2020 exhibit sharper and higher peaks compared to the other years (2019, 2021, 2022, 2023). Despite these variations, the overall seasonal trend of an increased pathogen count in the summer remains fairly consistent.

5.3 Geospatial Analysis

State populations are a critical factor in interpreting geospatial analyses, especially when data are normalized on a per capita basis. While per capita normalization adjusts for differences in state

population sizes, allowing for meaningful comparisons, several caveats must be considered to ensure accurate interpretation of the data.

Impact of Small Populations: States with smaller populations, such as Wyoming or Vermont, are particularly sensitive to small changes in raw counts. Even minor increases in reported cases can result in disproportionately high per capita values, exaggerating the apparent burden of illnesses or pathogen isolates relative to larger states.

Urban-Rural Divide: The distribution of urban and rural populations within states also adds complexity. States with densely populated urban centers, such as Texas or California, often exhibit higher per capita values due to better access to healthcare facilities and higher testing rates. In contrast, rural regions within these states may underreport cases due to limited healthcare accessibility, leading to uneven representation across the state.

Demographic Factors: Demographics play a significant role in influencing reported values. For instance, Florida's aging population naturally has higher healthcare utilization rates, which may result in increased testing and reporting of illnesses. This can inflate per capita values even if the overall disease burden is similar to other states with younger populations.

Healthcare Infrastructure and Reporting: The robustness of healthcare infrastructure and public health funding across states significantly impacts the quality and completeness of data. Larger states or those with well-funded healthcare systems tend to have more comprehensive data collection processes, while underfunded or smaller states may underreport cases, introducing biases into geospatial analyses.

Contextualizing Results: While per capita normalization provides a level playing field for comparing states, it is essential to interpret these analyses with caution. Differences in population size, demographics, urbanization, and healthcare infrastructure can introduce biases that affect the observed trends. Recognizing these limitations ensures that geospatial findings are contextualized within broader social and structural realities, leading to more accurate and meaningful interpretations.

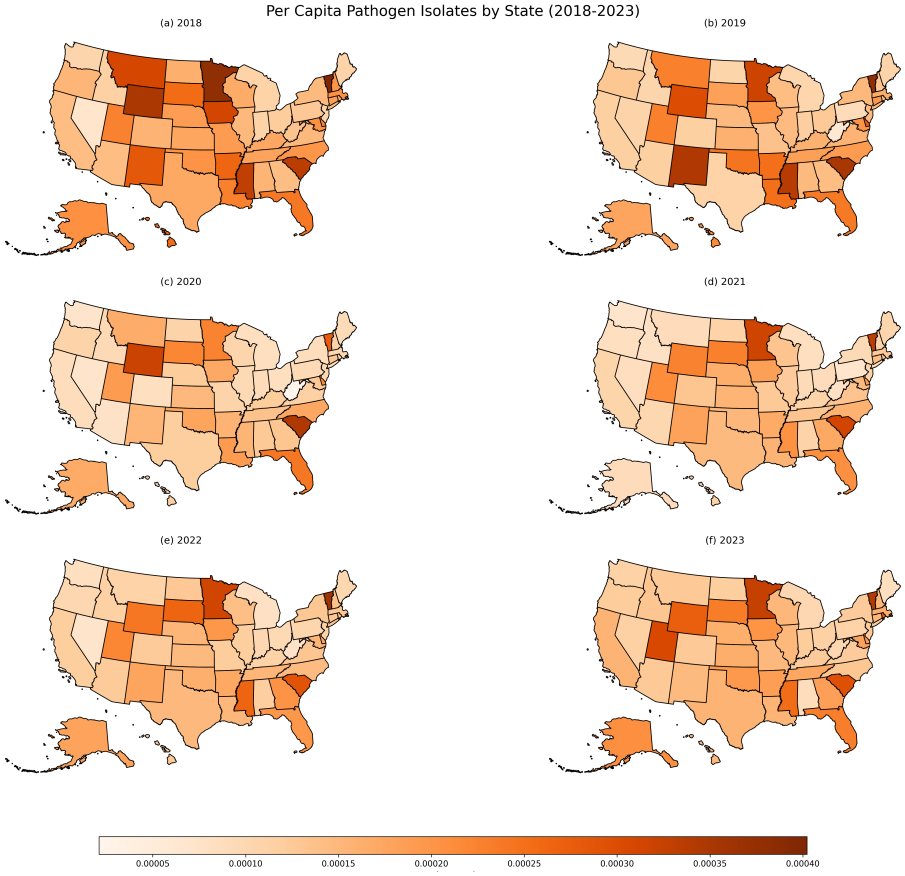

Figure 8: The geospatial maps illustrate the distribution of per capita pathogen isolates across U.S. states from 2018 to 2023. Higher isolate counts per capita are shown in darker shades, with states like Minnesota and Wisconsin consistently standing out. A sharp decrease in isolate counts is observed in 2020, likely due to reduced reporting caused by a shift of public health resources during the COVID-19 pandemic. By 2022, the values recover as testing levels return to normal. The maps show consistent geographic trends, with sparsely populated states generally exhibiting lower isolate counts per capita.

Figure 8 provides a geospatial analysis of per capita pathogen isolates across U.S. states from 2018 to 2023, offering a clear view of how pathogen activity varies relative to state populations. By normalizing pathogen counts by population, this metric allows for meaningful comparisons between states, highlighting regions where the burden of enteric illnesses is disproportionately high. The visualization uses a gradient of shading, with darker tones indicating higher per capita isolate counts and lighter tones representing lower values.

In 2018 and 2019, states like Minnesota and Wisconsin consistently display darker shading, suggesting higher per capita pathogen counts. These elevated values may reflect robust public health systems in these states, which enable more comprehensive testing and reporting of pathogen isolates. Conversely, states with lighter shading, such as those in the Southwest or sparsely populated states like Wyoming, may indicate lower pathogen burdens or potential underreporting due to limited healthcare resources and infrastructure.

A significant change occurs in 2020, where per capita isolate counts decline markedly across most states, resulting in a lighter overall shading on the map. This decrease aligns with the onset of the COVID-19 pandemic, which disrupted routine public health operations and redirected resources to pandemic management. Additionally, pandemic-related behavioral changes, such as reduced travel and fewer large gatherings, likely contributed to lower exposure and transmission of enteric pathogens. However, the decline in isolate counts may also reflect limitations in testing and reporting capacity during this period, masking the true burden of illness in some states.

By 2021, a gradual recovery is evident, as the darker shades reappear in states like Minnesota and Wisconsin, signaling a return to more typical pathogen reporting levels. This recovery underscores the

resilience of public health systems in these states, which adapted to pandemic challenges and restored their capacity for routine surveillance. Nevertheless, some states continue to exhibit lighter shading, highlighting disparities in healthcare infrastructure and the uneven pace of recovery across the country.

In 2022 and 2023, the maps show stabilization, with shading patterns resembling those observed in pre-pandemic years. States with consistently higher per capita isolate counts, such as those in the Upper Midwest, reaffirm the strength of their public health reporting systems. These states' ability to consistently capture and report pathogen data reflects their investment in healthcare infrastructure and public health initiatives. Meanwhile, states with lighter shading may still face challenges in achieving comprehensive surveillance and reporting, underscoring the need for targeted improvements in resource allocation and infrastructure.

The year-by-year trends in Figure 8 highlight the interplay between public health funding, healthcare accessibility, and population density in shaping pathogen reporting. States with darker shading are often those with strong surveillance systems and a commitment to public health priorities, while lighter-shaded states may represent areas where underfunding or limited infrastructure constrains data collection. This disparity underscores the importance of equitable resource distribution to ensure all states can accurately capture and respond to pathogen-related illnesses.

By examining per capita isolate counts, public health officials can identify states that may require additional support or interventions to enhance their surveillance and response capabilities. This analysis also serves as a reminder of the importance of investing in robust public health infrastructure to address not only routine illnesses but also unforeseen challenges, such as those posed by a global pandemic. These geospatial insights are critical for shaping data-driven strategies to improve health outcomes and reduce disparities in pathogen management across the United States.

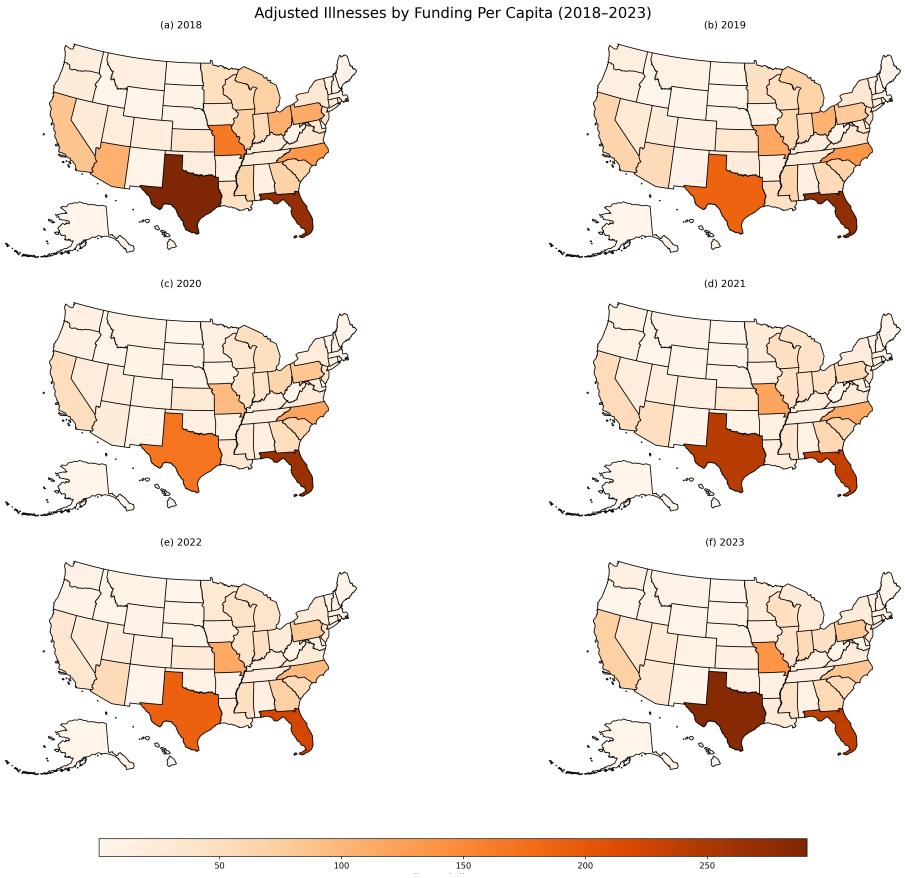

Figure 9: The geospatial maps depict adjusted illnesses by funding per capita across U.S. states from 2018 to 2023. Darker shades on the maps indicate states with higher adjusted illnesses per capita. States such as Texas and Florida consistently exhibit higher values, while others show varying patterns over time. A decline is noticeable in 2020, likely caused by the disruption in healthcare operations during the COVID-19 pandemic. By 2022 and 2023, the values stabilize, reflecting a recovery in healthcare reporting and data collection.

Figure 9 presents the geospatial distribution of adjusted illnesses across U.S. states from 2018 to 2023. Adjusted illnesses are calculated as the pathogen count for a given year divided by the corresponding public health funding for that year, providing a measure of illness burden relative to the resources allocated to each state. This adjustment offers valuable insights into how efficiently states utilize their funding to manage and mitigate enteric illnesses.

In 2018 and 2019, states like Texas and Florida consistently exhibit higher adjusted illness counts, as shown by the darker shades on the maps. These trends suggest a significant burden of illness relative to the funding levels in these states, indicating either insufficient resources to meet public health needs or inefficiencies in their utilization. Meanwhile, states with lower adjusted illness counts, represented by lighter shades, such as those in the Midwest, may benefit from more effective resource allocation or inherently lower pathogen prevalence.

The year 2020 reveals a dramatic reduction in adjusted illness counts nationwide, with most states appearing in lighter shades. This decline reflects the widespread impact of the COVID-19 pandemic, which diverted resources and attention away from routine illness surveillance and reporting. Additionally, pandemic-related behavioral changes, such as reduced travel and public gatherings, likely contributed to a temporary decrease in pathogen transmission, further influencing the adjusted illness counts.

In 2021, adjusted illness counts begin to rebound, with darker shades returning to states like Texas and Florida. This recovery highlights the gradual restoration of public health activities and funding allocations as the immediate pandemic pressures subsided. However, disparities persist, as some states exhibit slower recoveries, reflecting continued challenges in resource availability or systemic

inefficiencies.

By 2022 and 2023, the maps stabilize, closely resembling the patterns observed in pre-pandemic years. States like Texas and Florida remain prominent with higher adjusted illness counts, while others show more moderate values. These trends underscore the ongoing disparities in public health funding and resource effectiveness across the United States. The stability of the data in these years also suggests that public health systems have largely adapted to post-pandemic realities, resuming normal surveillance and reporting activities.

Adjusted illnesses provide a critical lens for understanding public health efficiency. States with consistently higher adjusted counts may need to reassess their funding strategies and operational priorities to improve resource utilization. Conversely, states with lower adjusted illness counts demonstrate more effective use of their funding, potentially serving as models for best practices in resource allocation and disease management. This metric also highlights the importance of equitable funding distribution to ensure all states can adequately address pathogen-related illnesses and reduce health disparities.

The geospatial analysis of adjusted illnesses provides actionable insights for policymakers and public health officials. Identifying regions with disproportionately high adjusted illness counts allows for targeted interventions, such as increased funding, enhanced healthcare infrastructure, or improved operational efficiencies. By leveraging this data, stakeholders can work toward a more equitable and effective public health system capable of addressing enteric illness burdens across the United States.

6 Foodborne Illnesses

Since the overwhelming majority of pathogen isolates analyzed in the BEAM Dashboard are Salmonella serotypes, Figure 10 provides important context on the relative contribution of different food sources to the burden of Salmonella-related illnesses.

Figure 10 clearly indicates chicken as the primary source of Salmonella-linked foodborne illnesses, accounting for over 2,700 cases during the 2011-2020 timeframe. This is substantially higher than the other meat sources examined, with pork coming in second at around 1,700 cases, followed by beef and turkey.

There are several scientific and social factors that likely contribute to chicken being the dominant source of Salmonella infections. First, Salmonella bacteria are commonly found in the gastrointestinal tracts of chickens and other poultry. Cross-contamination during processing and improper handling/cooking can lead to Salmonella transmission to consumers. Second, chicken is a widely consumed meat in the United States, both in terms of total volume and frequency. The high demand and production of chicken products increases the potential for Salmonella exposure compared to other meats. Third, chicken is often perceived as a quick, easy-to-prepare protein, but improper cooking or cross-contamination in home kitchens can fail to kill Salmonella and lead to illness. Lastly, given the regulatory scrutiny poultry production and processing industries have historically faced, chicken-associated Salmonella outbreaks may be more readily detected and reported than those linked to other meat types, contributing to the higher case numbers observed.

Salmonella infections can pose significant health risks, particularly for vulnerable populations. Though anyone can contract a Salmonella illness, certain individuals are at a greater risk of developing serious complications. High-risk groups include children under the age of 5, older adults, and individuals with compromised immune systems due to underlying medical conditions such as diabetes, liver or kidney disease, cancer, or cancer treatments.

It is important to note that the reported cases of Salmonella infections likely represent only a fraction of the actual incidence. Epidemiological studies suggest that for every laboratory-confirmed Salmonella case, there are approximately 30 additional unreported illnesses[**12**]. Many individuals who experience food poisoning symptoms do not seek medical attention or submit samples for testing, leading to an underestimation of the true burden of Salmonella-related illnesses in the community.

The discrepancy between confirmed cases and the estimated number of actual illnesses highlights the critical need for improved surveillance, public health education, and comprehensive reporting mechanisms to better understand and address the full scope of Salmonella-related public health challenges. Continued efforts are needed to further enhance food safety in this sector. Understanding these source-specific patterns is crucial for targeted interventions to reduce the public health burden of Salmonella.

The data analyzed in Figure 11 reveal clear patterns in the distribution of Salmonella serotypes across beef, chicken, pork, and turkey. For chicken, beef, and turkey, the top Salmonella serotypes display a distinct gamma distribution, with a few dominant serotypes accounting for the majority of illnesses and a longer tail of less prevalent strains. This gamma distribution is particularly pronounced for chicken, in which the top Salmonella serotype, Saintpaul, is responsible for over 60 illnesses, dwarfing the next most common serotypes. A similar skewed distribution is seen for turkey, with the Hadar serotype accounting for the largest share. Interestingly, the pork data does not follow this gamma pattern, with a more even distribution across the top Salmonella serotypes. This finding suggests that the epidemiology of Salmonella in pork-related illnesses may be more complex, potentially driven by a higher diversity of serotypes.

These source-specific Salmonella serotype profiles provide valuable insights that can guide targeted prevention and mitigation strategies. By understanding the key strains associated with each food type, public health officials and the food industry can focus surveillance, intervention, and education efforts on the most relevant and impactful Salmonella threats.

Integrating this granular view shown in Figure 11 with the broader Salmonella burden trends shown in Figure 10 offers a comprehensive picture of the pathogen landscape, underscoring the importance of understanding serotype-specific patterns in foodborne illness outbreaks for targeted prevention and control measures. This multi-layered understanding is essential for developing effective, evidence-based policies and programs to reduce the significant public health impact of Salmonella-related foodborne illnesses.

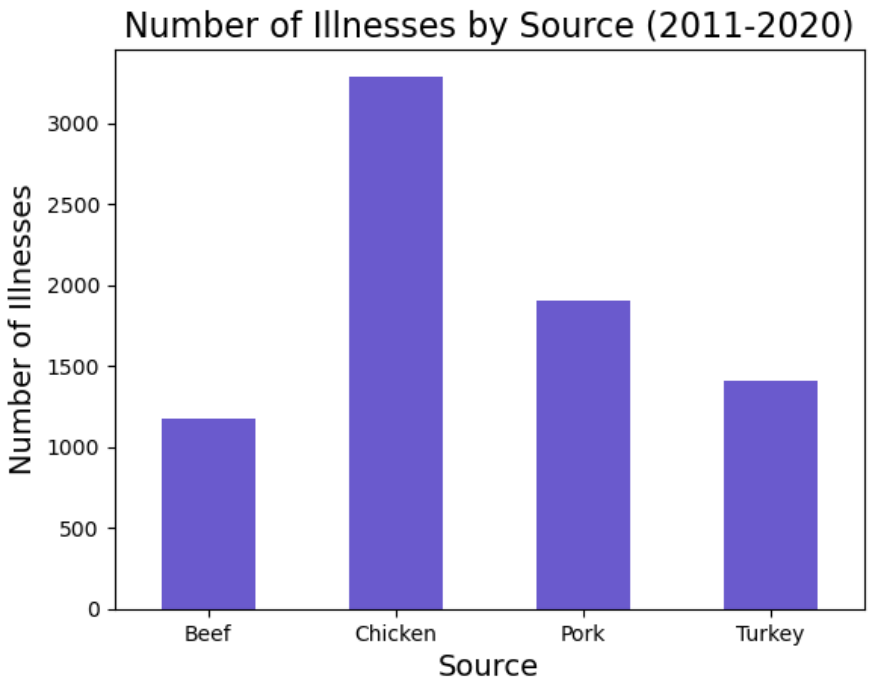

Figure 10: This is a bar graph of the number of reported foodborne illnesses from Salmonella, broken up by source. The x-axis shows the meat sources of foodborne illnesses: beef, chicken, pork, and turkey. The y-axis shows the number of foodborne illnesses. The majority of foodborne illnesses from Salmonella originate from chicken.

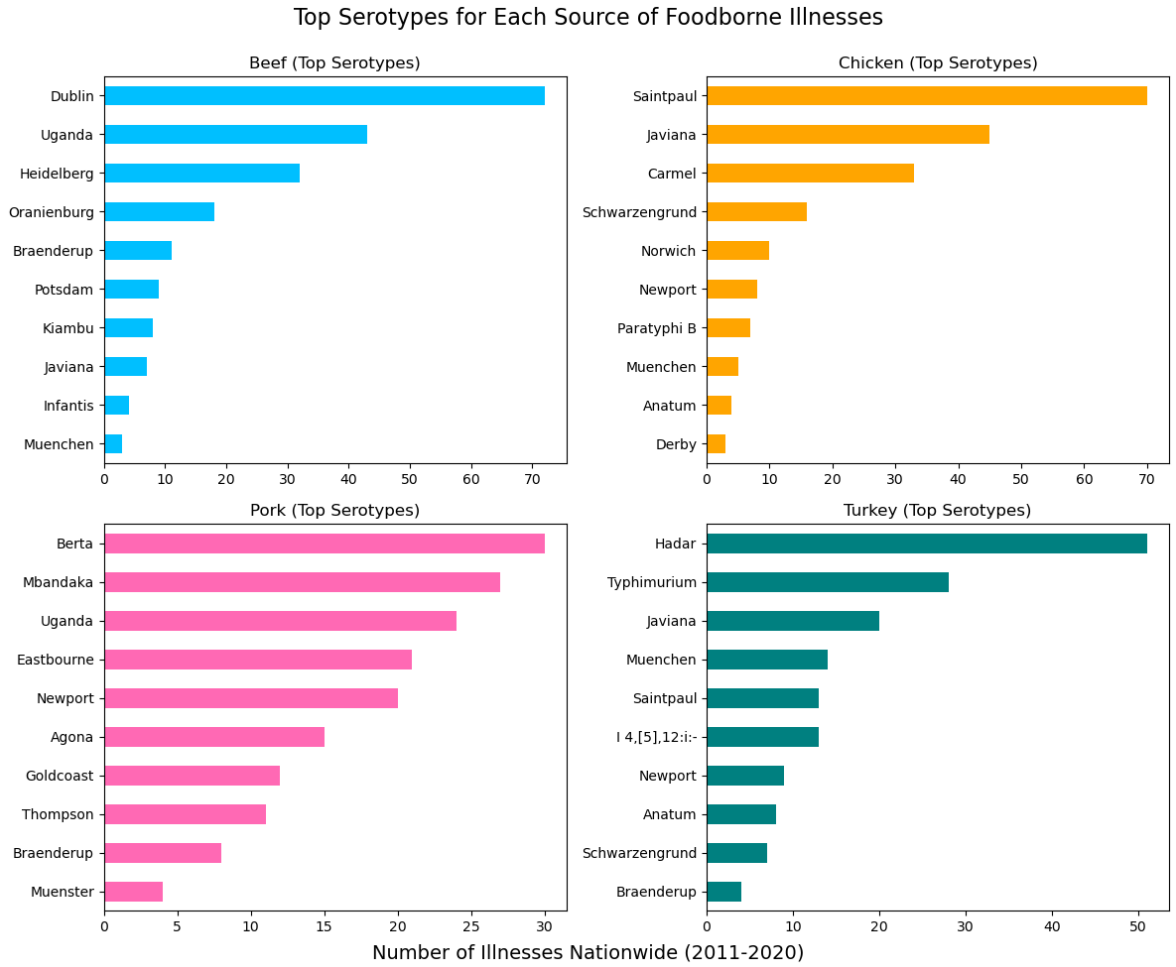

Figure 11: This is a set of 4 horizontal bar graphs that show the top serotypes for each source (beef, chicken, pork, turkey) of foodborne illnesses. The x-axis shows the number of illnesses nationwide from 2011 to 2020. The y-axis shows the top 10 serotypes of Salmonella identified in the meat sources. Each chart highlights the variability in the prevalence of specific serotypes across different meat sources, with distinct trends observed for beef, chicken, pork, and turkey. For example, Dublin is most prevalent in beef, while Saintpaul is the leading serotype in chicken, Berta in pork, and Hadar in turkey.

7 Limitations

Analyzing data from the CDC BEAM Dashboard comes with limitations due to the dependency of reporting rates on public health funding per state. Changes in reporting procedures and resource availability over time have significantly influenced the quantity and quality of reported foodborne disease outbreaks[**13**]. For example, variations in funding and priorities have led to fluctuations in the number of reported outbreaks, with a substantial increase in annual outbreaks after 1998, coinciding with changes in reporting standards. The ability to identify causes of outbreaks has improved over time, likely reflecting advances in diagnostic technologies and laboratory capacities, which are also tied to funding. This key limitation highlights how social, economic, and regulatory factors can influence the completeness and comparability of the dataset, complicating long-term trend analyses and potentially introducing biases in understanding the epidemiology of foodborne illnesses.

The analysis of foodborne outbreak reporting highlights significant variability across states, influenced by factors such as state capacity, public health funding, and surveillance practices. Reporting through the NORS system varies widely, with high-reporting states documenting up to four times as many outbreaks per capita compared to low-reporting states[**14**]. This disparity extends to outbreak types, with low-reporting states focusing more on larger outbreaks caused by reportable pathogens

like Salmonella or STEC O157 and less on attributing settings or food vehicles. Regional factors, such as the higher prevalence of fish toxin outbreaks in coastal areas, further contribute to differences in reporting. Importantly, states with higher per capita funding demonstrate stronger surveillance capacity, reporting more outbreaks across multiple pathogen groups.

Structural limitations and resource constraints significantly impact state-level outbreak detection and reporting. Differences in surveillance systems, laboratory capacities, and investigation priorities affect the consistency and completeness of outbreak data[**14**]. For instance, jurisdictions may prioritize outbreaks linked to severe pathogens while deprioritizing norovirus outbreaks or those unlikely to yield actionable insights. Additionally, variations in the timeliness and completeness of reporting, coupled with a lack of standardized metrics across states, limit the ability to assess improvements over time. While investments in public health funding improve surveillance capacity, their impact varies based on local preparedness and structural factors. This emphasizes the need for targeted funding, cross-jurisdictional collaboration, and updated national surveys to enhance outbreak detection and response capabilities, ensuring that public health efforts are both efficient and equitable.

8 Future Work

To address limitations identified in this analysis, future work should focus on developing standardized metrics for foodborne outbreak reporting across all states. These metrics would improve comparability of outbreak data and improve long-term trend analyses. Additionally, incorporating socioeconomic, demographic, and healthcare access variables into surveillance systems would provide a more comprehensive understanding of disparities in public health capacities and enhance the equity of outbreak detection efforts.

Advancing the analytical capabilities of tools like the CDC BEAM Dashboard is another critical area for future work. Enhancements could include features that allow for population-adjusted outbreak rates, improved data storytelling methods, and more user-friendly interfaces to ensure accessibility for a wider audience. Leveraging machine learning and predictive modeling could further aid in identifying outbreak hotspots in real time and improve the timeliness of response strategies.

Cross-jurisdictional collaborations should be prioritized to address disparities in surveillance capacity between regions. By improving partnerships and sharing resources, public health systems can better address differences in their ability to detect and respond to outbreaks. This is particularly important for under-resourced areas that may lack the infrastructure or workforce necessary to conduct thorough investigations. Investments in laboratory infrastructure, workforce training, and diagnostic technologies are also essential for improving outbreak detection. Improved laboratory capabilities would allow for faster and more precise disease identification, while workforce training ensures that health officials are ready to face new threats. Diagnostic advancements, such as rapid testing and molecular methods, can help detect outbreaks that might otherwise go unnoticed, including those caused by rare or less severe pathogens[**15**].

Additionally, integrating genomic data into outbreak analyses can significantly improve our understanding of how pathogens evolve and spread. Genomic sequencing enables researchers to track the transmission paths of infectious pathogens and detect mutations that could increase virulence or resistance to treatments. Genomic data offer new insights into epidemiologic and evolutionary dynamics, and sequencing pathogen samples is becoming increasingly routine, providing information critical to public health[**16**]. These data allow determination of the phylogeny of isolates, shedding light on potential transmission networks and routes of infection. By identifying these routes, researchers can estimate risk factors for disease transmission and develop targeted infection control strategies. This level of insight ensures that resources are directed where they are most needed to prevent and control outbreaks effectively.

9 Conclusion

The CDC's BEAM Dashboard delivers real-time, automated analyses using data from sources such as SEDRIC and NORS, offering valuable insights into patterns and trends in enteric illnesses. It is designed to aid public health officials, researchers, and industry professionals in making timely decisions and implementing targeted interventions. However, the dashboard's interactive format can make it

difficult for some users to interpret findings or derive actionable insights. Moreover, the dashboard's map displays do not account for population size or public health funding differences across states, essentially reflecting population density rather than adjusted outbreak rates. This limitation highlights the need for additional tools and effective data storytelling that present adjusted data in a clear and accessible format, helping to ensure that the information is useful to a wide and varied audience.

The ability to translate complex data into clear and impactful narratives is especially crucial when addressing public health challenges like enteric infections, which significantly affect nutrition, childhood development, and global health outcomes[**8**]. While innovations in medical therapies have successfully reduced mortality caused by dehydration, they have not addressed the broader morbidity associated with these infections. By integrating insights from the growing understanding of how enteric pathogens interact with human genetics and drive inflammation, effective data storytelling and exploratory data analysis can illuminate opportunities for innovative interventions. This underscores the importance of presenting data in ways that not only inform but also inspire actionable strategies to improve health outcomes worldwide.

References

[1] CDC, "About the National Antimicrobial Resistance Monitoring System (NARMS)," *National Antimicrobial Resistance Monitoring System for Enteric Bacteria (NARMS)*, Sep. 2024.

[2] CDC, "Enteric Disease Fact Sheet," 2014.

[3] CDC, "About Antimicrobial Resistance," *Antimicrobial Resistance*, Jun. 2024.

[4] CDC, "About National Outbreak Reporting System (NORS)," Jan. 2024.

[5] CDC, "Serotypes and the Importance of Serotyping Salmonella," *CDC Salmonella Atlas*, Sep. 2022.

[6] CDC, "BEAM Dashboard FAQs: Bacteria, Enterics, Ameba, and Mycotics Data," Sep. 2024.

[7] CDC, "SEDRIC: System for Enteric Disease Response, Investigation, and Coordination," *Foodborne Outbreaks*, Jun. 2024.

[8] W. A. Petri, M. Miller, H. J. Binder, M. M. Levine, R. Dillingham, and R. L. Guerrant, "Enteric infections, diarrhea, and their impact on function and development," *The Journal of Clinical Investigation*, vol. 118, no. 4, pp. 1277–1290, 2008.

[9] Cleveland Clinic, "E. coli Infection."

[10] R. B. Simpson, A. V. Kulinkina, and E. N. Naumova, "Investigating seasonal patterns in enteric infections: a systematic review of time series methods," *Epidemiology and Infection*, vol. 150, p. e50, 2022.

[11] R. B. Simpson, B. Zhou, and E. N. Naumova, "Seasonal synchronization of foodborne outbreaks in the United States, 1996–2017," *Scientific Reports*, vol. 10, no. 1, p. 17500, 2020.

[12] CDC, "Salmonella Prevention."

[13] T. F. Jones and J. Yackley, "Foodborne disease outbreaks in the United States: a historical overview," *Foodborne Pathogens and Disease*, vol. 15, no. 1, pp. 11–15, 2018.

[14] A. E. White, A. R. Tillman, C. Hedberg, B. B. Bruce, M. Batz, S. A. Seys, et al., "Foodborne Illness Outbreaks Reported to National Surveillance, United States, 2009–2018," *Emerging Infectious Diseases*, vol. 28, no. 6, pp. 1117–1127, 2022.

[15] E. Gerace, G. Mancuso, A. Midiri, S. Poidomani, S. Zummo, and C. Biondo, "Recent Advances in the Use of Molecular Methods for the Diagnosis of Bacterial Infections," *Pathogens*, vol. 11, no. 6, p. 663, Jun. 2022.

[16] C. J. Worby, M. Lipsitch, and W. P. Hanage, "Shared genomic variants: identification of transmission routes using pathogen deep-sequence data," *American Journal of Epidemiology*, vol. 186, no. 10, pp. 1209–1216, 2017.

A Appendix

A Appendix A: BEAM Dashboard Data

	year	month	state	source	pathogen	serotype_species	num_isolates
0	2024	1	AK	Stool	Campylobacter	jejuni	1
1	2024	1	AL	Stool	Campylobacter	coli	1
2	2024	1	AL	Stool	Campylobacter	jejuni	2
3	2024	1	AR	Stool	Campylobacter	jejuni	1
4	2024	1	CA	Stool	Campylobacter	jejuni	1

Figure 12: Snippet includes $n = 5$ for the beam dashboard csv file. Consists of the following columns: year, month, state, source, pathogen, serotype_species, and num_isolates.

B Appendix B: Merged Geospatial Data

	year	month	state	source	pathogen	serotype_species	num_isolates	geometry
0	2024	1	AK	Stool	Campylobacter	jejuni	1	MULTIPOLYGON (((-2778499.467 -1620603.792, -27...
1	2024	1	AL	Stool	Campylobacter	coli	1	POLYGON ((1023282.917 -582853.454, 1023269.168...
2	2024	1	AL	Stool	Campylobacter	jejuni	2	POLYGON ((1023282.917 -582853.454, 1023269.168...
3	2024	1	AR	Stool	Campylobacter	jejuni	1	POLYGON ((461603.731 -366288.124, 461727.959 -...
4	2024	1	CA	Stool	Campylobacter	jejuni	1	MULTIPOLYGON (((-1976277.366 663684.581, -1976...

Figure 13: Table ($n = 5$) of geospatial data for isolates, including the year, month, state, source, pathogen, serotype_species, num_isolates, and the associated geometry information after merging with shapefiles for mapping purposes.

C Appendix C: Illnesses by Source and Serotype

	No_of_illnesses	source	serotype
0	11	Beef	Braenderup
1	72	Beef	Dublin
2	81	Beef	Enteritidis
3	32	Beef	Heidelberg
4	4	Beef	Infantis

Figure 14: This table provides detailed data on the number of reported illnesses associated with specific food sources and serotypes. The No_of_illnesses column records the total cases linked to each serotype, while the source column identifies the food category responsible (e.g., Beef). The serotype column specifies the pathogen serotype responsible for the reported illnesses.

D Appendix D: Census Population Data (2018–2023)

	Geographic Region	2018	2019	2020	2021	2022	2023
0	United States	32,66,87,501	32,82,39,523	33,15,26,933	33,20,48,977	33,32,71,411	33,49,14,895
1	Alabama	48,87,681	49,03,185	50,31,864	50,50,380	50,73,903	51,08,468
2	Alaska	7,35,139	7,31,545	7,32,964	7,34,923	7,33,276	7,33,406
3	Arizona	71,58,024	72,78,717	71,86,683	72,72,487	73,65,684	74,31,344
4	Arkansas	30,09,733	30,17,804	30,14,348	30,28,443	30,46,404	30,67,732

Figure 15: Table ($n = 5$) summarizes the population data for various geographic regions in the United States from 2018 to 2023. The Geographic Region column lists the regions and states, including a summary row for the entire United States. Subsequent columns represent population counts for each year, formatted as integers with thousands separated by commas for clarity. The data shows annual population trends at both the national and state levels, providing a basis for longitudinal demographic analysis.

E Appendix E: State Funding Data

	location	year	funding	state_abbrev
0	Alabama	2018	57.23	AL
1	Alabama	2019	54.25	AL
2	Alabama	2020	46.81	AL
3	Alabama	2021	52.35	AL
4	Alabama	2022	49.48	AL

Figure 16: This table ($n = 5$) provides a comprehensive overview of annual funding allocations to U.S. states from 2018 to 2022. The location column identifies each state by name, complemented by the state_abbrev column, which lists the standardized two-letter abbreviations for ease of reference. The year column delineates the specific fiscal year for each funding entry, while the funding column quantifies the allocated amounts in millions of dollars, precise to two decimal places.