Spatial and Temporal trends of Food Borne Pathogens

Anthony Girard, Timothy Hedspeth, Yutong Li

October 23^{rd} 2022

1 Introduction

Have you ever woken up in the morning and saw a news headline, like "E Coli. Outbreak at Chipotle leads to national recall"? Chances are even if you don't remember it this has happened to you to some extent in the past. News like this is easy to shrug this off, but in the United States we note that every year there are an estimated 37 million cases of food borne illness [3]. Though there is a huge burden of food borne diseases yearly, a small subset leads to hospitalization, and an even smaller subset ends up contributing to death [11]. Approximately 1 in 6 Americans get sick with a foodborne illness, and 3000 associated deaths occur annually in the US. Unfortunately, progress on reducing the incidence of foodborne illness has stalled in recent years [14]. Globally, foodborne illness is a problem that has been on the rise, as the global supply chain has increased the network of food dispersion throughout the world [13]. The goal of the Healthy People 2030 Foodborne Illness Reduction Committee that we share is to further limit the impact and incidence of foodborne illness in the US. For our project we approach this goal hoping to inform methods for outbreak detection and analysis.

2 Literature Review

Foodborne illness surveillance plays a critical role in limiting the impact of disease. Early detection of an outbreak can allow public health officials and other entities to respond in time to reduce the burden of disease. The World Health Organization (WHO) defines a disease outbreak as "the occurrence of cases of disease in excess of what would normally be expected in a defined community, geographical area or season" [16]. Whereas by the CDC's definition, a foodborne disease outbreak occurs when two or more individuals experience a similar illness through the consumption of a common food [3]. This definition is quite strict and a challenge to evaluate given the structure of our data, leading us to favor the WHO definition for the purposes of our analysis. According to the CDC, 95% of cases are sporadic, not belonging to an outbreak for which specific sources of cases are not determined. This means an outbreak was not identified to attribute the cases to, not that no outbreak exists, and only further highlights the potential for improvement in automatic surveillance of foodborne illness. The benefits of improving outbreak detection are clear, and this review seeks to explore methods for outbreak detection, emphasizing foodborne illness, found in the literature.

Outbreak detection largely relies on longitudinal, or time series data, due to the temporality of the definition. In addition, inclusion of spatial data is extremely important. Cases owing to the same outbreak are expected to occur similarly in terms of temporal and spatial distance. Other factors are often included in analysis of time series data for foodborne illness outbreaks, and we will discuss those we have access to in the exploratory data analysis section. For example, the CDC provides longitudinal data, which details the incidence of food borne pathogens by disease linking strains that are similar via a phylogenetic tree, which details the genetic similarity among the strains [4]. We will be using the FoodNet database from the NCBI for this project, which contains spatial and temporal data on foodborne illness incidence, allowing for our intended analysis regarding outbreak detection.

Below we include a discussion of publications found in the area of outbreak detection and time series analysis of foodborne illness.

Looking at a subset of 7 of the common foodborne illnesses and two protozoa researchers have started to create time series data regarding the incidence per 1,000,000 individuals. The analysis conducted utilizes synchronization analysis, for the 9 infections across 11 regions, and the authors performed Negative binomial regression with a log link and harmonic regression where they accounted for seasonal changes [13]. The authors found that there was high variability in the magnitude of infections across states and across time. The observation of sporadic outbreaks is most observable when looking at the state level. Infections peak at different times of the year. This paper demonstrates the importance of controlling for seasonality and synchronization between outbreaks of different diseases, and it includes some methods for modeling both.

Hidden markov models (HMM) have been used to improve Salmonella outbreak detection [16]. In this paper, two approaches are taken to assess whether incidence of cases are at an endemic baseline level or if cases can be considered part of an outbreak, where incidence of Salmonella is greater than the region's endemic baseline. The paper gives an overview of the Farrington-Noufaily (FN) algorithm and HMM approaches to outbreak detection, both with cited past use for this problem. Time series data was extracted from the SurvNet database for counties in Germany. Models were fit in a supervised and unsupervised manner. For the supervised models, outbreak labels are needed. Another paper employing HMM found it highly advantageous to SVM approaches for outbreak detection [15]. It mentions the benefit of multiple streams of data in syndromic surveillance systems, though this would require an effort to create mobile applications where users report cases themselves as part of an automatic surveillance system.

The objective of our work is to help in the surveillance and potentially predict outbreaks of foodborne illness, which is a rather rare occurrence as 95% of the cases of foodborne illness are sporadic. One paper looks at Salmonella Typhimurium that have 4 sets of genomes, for selected human isolates from sporadic cases from 1949-2014, the data was run through PCA and subsequently run through a Random Forest algorithm to predict sources for Salmonella genomes [18]. The authors were able to find pseudo gene differences for pseudogenes. The RF algorithm was able to rank genetic features by their importance. The algorithm was able to find 10 core mutations and the 2 most important features were SNPs that are related to cell regulation. This method shows that we can learn genotypes that can be used to predict zoonotic illnesses.

In a survey of foodborne illnesses in China authors look at classifying real outbreaks of food borne illness as a classification problem [17]. The authors use a multitude of ML algorithms to fit to the data for the purposes of classification, namely SVM, RF, gradient boosting, logistic regression for determining if there was a true outbreak. The authors found that boosting had great precision and that all of the models generally performed well for their data.

Isolating the source and understanding how a disease spreads has long been an issue of foodborne outbreaks with records of people attempting to understand and prevent it dating back to the 1800s [12]. With advances in statistical methodologies it makes sense to leverage these techniques to help predict an outbreak. The current process for predicting an outbreak is through the use of abnormal lab results and an uptick in cases, though the authors of this paper sought to use google search data and machine learning algorithms to predict outbreaks. They develop an algorithm to predict the restaurants that are at highest risk for transmitting the illness, the model was validated with an application to real restaurants in Las Vegas and Chicago and they were able to find that over half the restaurants flagged by the algorithm were in fact unsafe which improves the accuracy of inspections.

In China there is a long-term project to support the National Foodborne Disease Outbreak Surveillance System [6]. This system is multifaceted with systems focused on monitoring/reporting, outbreak management, and a molecular traceback network to help inform their food safety. This group looks at data crawler methods to use NLP from websites to extract information regarding websites to create a map to isolate the potential sources of the food borne outbreaks. In order to develop a model that can predict pathogens the authors designed an extreme gradient boosting model. This model takes

into account the spatial, temporal and symptom features to predict the pathogen that is involved in an outbreak, while also taking into account the food source. This can be used to quickly identify a pathogen and can help support an outbreak prediction. The authors also developed a real time outbreak prediction tool, as most of there is a lot of false labeling of events that are not actually outbreaks (only 20% correct). This is costly as these events must be investigated. This model incorporates information from the food and the individual. Their model finds that information regarding individuals' health status has an important role in predicting outbreaks. The authors also looked at retrospective data to assess the risk of foodborne illness in the medium and long term for different spatial reasons. This requires the use of historical data to predict the outbreaks. More specifically the data can be used. The authors use a multigraph structured long short term memory network to predict the spatiotemporal risk of disease as it accounts for the dependence between temporal and spatial dependence. Suggest that it could be beneficial to use search engine data, which we do not have.

Due to the fact that this problem is related to spatial and temporal data, it requires non-trivial ways of considering how to model and predict the outcomes [7]. There are a number of ways to address this but the authors of this paper developed the multistep spatial-temporal based on encoder-decoder structure and module. Their model decomposes a city or region into subregions and creates an undirected graph. The data is then loaded corresponding to the region and the number of cases in a given time. The approach uses a multigraph fusion module to take into account the spatial correlations, complexity and importance all into one in the model, and can even consider things such as holidays and seasonality. Look at data by month to help determine when an outbreak is coming. Given the complexity of the data presented the authors utilized the China National Center for FoodSafety Risk Assessment, which consists of foodborne disease records from hospitals, though they were only able to focus on a small subset of the information, 3 spatial dependencies in Beijing. The authors compared the results of this method with other methodologies that are applicable such as simple historical averages, autoregressive analysis, Autoregressive integrated moving average, LSTM, and spatial-temporal graph convolutional neural networks. The authors concede that the historical average models fail to account for the complex relationships in the data. Autoregressive models use time series analysis with a linear combination of values at previous time steps. ARIMA uses autoregressive terms with moving average value, with the data preprocessed. LTSM are more frequently used in NLP problems, and learn sequence dependence upon the structure of the data. While Spatial temporal graph convolutional neural networks are an extension of convolutional neural networks. Their proposed model out performed the other methods that they mention.

Using data from the CDC authors sought to predict salmonella infections in Mississippi, as southern parts of the United States are more vulnerable to food borne illness [5]. The authors use lab confirmed E. Coli and Salmonella cases from 2002 to 2012 for selected states. The authors grouped the data by the year and district, and adjusted the cases to 100,000 of the population. The authors used regression analysis to analyze the impact of the socioeconomic factors on these patterns. But their main objective was to apply a Neural Network to the number of cases in a district, they use a General Regression Neural Network. Their neural networks predict based on the demographics of the district. The authors have worked on integrating machine learning with existing techniques to help with the inputting of foodborne diseases and pathogens [6]. The use of surveillance have had positive impacts on the

For the bacterial pathogens of food-borne illness, Salmonella, E. coli, Shigella and Campylobacter jejuni are the top reported agents. Salmonella has been paid considerable attention which colonizes a broad list of hosts and livestock animals such as pigs, cattle and poultry [10]. Therefore, food-borne outbreaks of Salmonella are the most frequently reported. E. coli is another often heard bacteria which is a successful gut colonizer and one problem with the E. coli family is its ability to constantly mutate types which has not been characterized before [10]. Campylobacter is the most frequently reported cause of bacterial food poisoning in the EU and one feature of C. jejuni is it cannot grow outside the host [10]. But it responds quickly to the environmental changes which cause antimicrobial

problems. Those contaminated food sources will be processed and put in the market. To relieve the burden of bacterial pathogen infections and often outbreak of food-borne illness, there were studies applied data mining, machine learning and deep learning methods to recognize the data patterns in food-borne disease outbreaks and risk prediction based on the large size of reported cases. One study used a deep learning method of implementing artificial neural networks to discover patterns in food-borne disease. They used data from the CDC database of food-borne disease outbreaks [8]. For data preprocessing, the study replaced empty values in named illness. Hospitalization and fatalities by means of particular attributes. The data from CDC has ambiguity and this part was removed to make a prediction. The data was separated into 80% training set and 20% test set. Model performance evaluated using confusion matrix.

This initial review of literature containing methods and limitations for handling outbreak detection and seasonality of foodborne illness informs us of analyses that would be beneficial to the goal of the Healthy People 2030 Foodborne Illness Reduction Committee.

3 Data Tools

In this analysis we consider longitudinal data from 2013-2022 regarding 3 food borne illnesses, Salmonella, E.coli, and Campylobacter extracted from the NCBI pathogen detection website [1]. We have 3 different data sets each of which contain 25 variables for the illnesses individually. Of the 3 illnesses the Salmonella data set is the largest with 93,763 observations, Campylobacter is the second largest with 44,949 observations and E. coli is the smallest with 34,805 observations. In order to increase the efficiency of the exploratory analysis we load these data sets separately so computation can be done on individual data sets reducing the potential for bottle necks that could come with the aggregation of these data sets.

The data was extracted from NCBI [1] and in our extraction process we conditioned on the *Collection.date* variable, guaranteeing that the data used in our analysis was collected between the years 2013-2022. The objective of this analysis, at least at baseline, is to focus on food borne illness in the United States and historical trends that could help inform potential surveillance methods. Aside from conditioning on *Collection.date* we imposed no other restriction besides the illness of interest.

In the initial analysis we noted some preprocessing was required. As noted in the objectives in the **Introduction** we are interested in examining data regarding the United States, the first R scripts (*illnessname*_LocationClean.R) are used to subset the three illnesses to only include only locations that are in the United States and clean for location levels using the **%like%** command, as the extracted data contained information from many countries, which is dropped from the data before we proceed with our analysis. Then, we noted that the *Location* variable does not have a uniform format (136 levels in the Salmonella data, 291 levels in the E. coli data and 74 levels in Campylobacter data). We cleaned the location to have a standard two letters state abbreviation and region names which resulted in 56 levels in Salmonella and E. coli data, and 53 levels in the Campylobacter data. No missing values were found for the *Location* variable. After running this script we noted an extreme level of heterogeneity($>$ 6,000 levels in the Salmonella data) in the *Isolation.Source* variable, which required preprocessing prior to conducting an exploratory analysis. The second R script (source_group.R) uses the **tolower()** command to make the levels all lowercase and therefore easier to work with. With the data in a more manageable format we used the **%like%** command to find specified levels and place them into more general groups for our analysis, e.g. grouping levels that contain apple into a produce category. We placed the source categories into 8 groups, including a group for missing and other sources. Missing values of source accounted for 62% of total cases. We grouped sources that did not match our initial search into this category, but this accounted for less than 1% of observations, none of which had a frequency over 50 ($<$0.02%). After the data is extracted from the second R script, we noted that the *Collection.date* variable, the primary time metric of interest, had different reporting practices, which we correct for in our third R script (Clean_dates.R). This script uses the **tidyvserse**, **anytime** and **lubridate** packages to create data sets that account for the different date

formats (full date, year and month, just year) and decompose *Collection.date* into 3 new variables *year*, *month*, and *day*, so that it is easier to work with the date information. The R scripts used for this analysis and the data sets used in our analysis are publicly available on *OurGithub*. The code for the exploratory analysis, can be found under Final_project_EDA_markdown.RMD or EDA.R. As we further refine our methods for analysis we will update this section with other relevant packages and their purposes.

4 Exploratory Data Analysis

4.1 Missing Data

The NCBI collects data from surveillance and research efforts that are currently ongoing [1]. The data looks at a multitude of sources for these illnesses such as food, patients, production facilities, etc [1]. After data is submitted it is clustered to related pathogens, allowing for people to look for strains of pathogens that are closely related [1]. Given that the data is being uploaded from multiple sources with what we presume to be free text fields in some columns eg. Isolation source there is a lot of heterogeneity in reporting and in data quality. Which is a major limitation of this data, as the high variability in these columns makes them very challenging to clean, requiring a substantial amount of time that could be used analyzing the data.

Table 1: Percent missing for each illness dataset

	Salmonella	E. Coli	Campylobacter
Serovar	0.22	0.87	1.00
Host.disease	0.99	0.90	1.00
Isolation.source	0.11	0.15	0.04
Lat.Lon	0.98	0.81	1.00
Source.type	1.00	1.00	1.00
SNP.cluster	0.06	0.47	0.15
Min.same	0.06	0.53	0.16
Min.diff	0.17	0.84	0.37
Assembly	0.11	0.07	0.01
Outbreak	1.00	1.00	1.00
day	0.85	0.67	0.98
month	0.69	0.54	0.84

We note from table 1 that there are some variables that are missing in great quantities across all of the data sets, such as *Serovar*, *Host_ disease*, *Lat.Lon*, *Source_ type*, *Computed.types*, and **outbreak**. Given the extreme missingness for the previously mentioned variables in most of our illness types we decide to remove them from contention for our analysis. On the contrary we have no missingness in the *organism.group*, *Isolate*, *Create.date*, *Collection.date*, *Location*, *Biosample*, and the variables we created in the preprocessing *Isolation.source.category* and *year*. Though we observe that there are extremes there are variables with differing levels of missingness, which is likely due to the fact that a lot of the data is reported by the researcher submitting to NCBI [1]. The *Strain.name*,*Isolation.Source*, and *Isolation.type* are missing in very small quantity which could be due to an oversight by the researcher when submitting their data. It appears that *SNP.cluster* is missing for almost half of the E.coli data, and has some missingness in the other illnesses, but not nearly to this extent. This could be due to issues or delays in processing the E. coli pathogens. It is also apparent that the month and

year variable that we created for this analysis have a lot of missingness, which can be explained by the issues in *Collection.date* that we observed. Generally a lot of variables are dependent upon reporting practices of the people submitting to the pathogen detection database, which, in our opinion, can explain why some of these variables have a lot of missing data.

4.2 Variables of interest

Prior to proceeding to further analysis we will discuss the variables we will use. Recall that the goal of this project is to look at potential outbreaks of food borne illness and explore their seasonality. This means that our main interest lays in the *Collection.Date*, the date the sample was collected, and *Location*, the location where the sample was collected [2]. Given that we want to observe if there are temporal or spatial trends regarding the illnesses the aforementioned variables are of the utmost importance. On top of this we consider information regarding, *Isolation.Source* and *Isolation.type* which give where the sample was derived from and if the setting was clinical or environmental [2]. Also it is of interest to look for similarities between strains, to accomplish this we can look at the *SNP.Cluster*, which gives related clusters to the isolate, or the *Min.Same* and *Min.Diff* variables which look at the closest Isolates (in terms of SNP) for Isolates of the same or different types respectively [2]. Since there is approximately half of the data missing in the *SNP.Cluster* we will assess relative closeness with the *Min.Same* variable. The other variables included in the data extracted are not explored here as they are not prevalent to our analysis, but descriptions can be found on the NCBI website [2].

4.3 Trends over Time and Location

Given that we are interested in outbreaks we find it most pertinent to first understand where and when cases are popping up. This will, hopefully, afford us the opportunity to see historical trends in the data and if cases are more concentrated in some regions of the United States, which will help inform our decisions regarding modeling in future work. With this in mind we first look to figure 1 below:

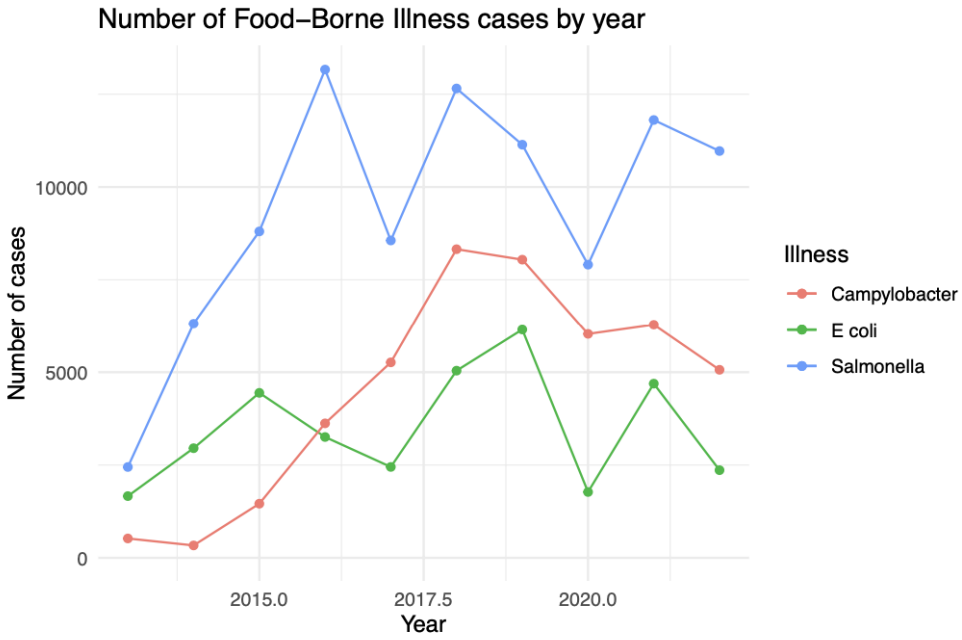

Figure 1: Number of cases per year, all illnesses

From Figure 1 we note that there are more cases of Salmonella than any other food borne illnesses in any given year. The number of cases of campylobacter is the smallest at onset, but the number of campylobacter cases surpasses the number of E.coli cases in 2016 and the number of cases remains consistently higher than the number of E. coli cases for remaining years. We observe that for all of the illnesses that the number of cases decrease in 2020, which we note is the year the Covid-19 pandemic started, though we cannot confirm a causal relationship with the provided data we can assume that this is due to less reporting or less going out to eat and getting exposure to food that could make people ill, though this is conjecture and we cannot further comment on this. We next want to see if we can notice any trends in seasonality, which we explore in figure 2 below.

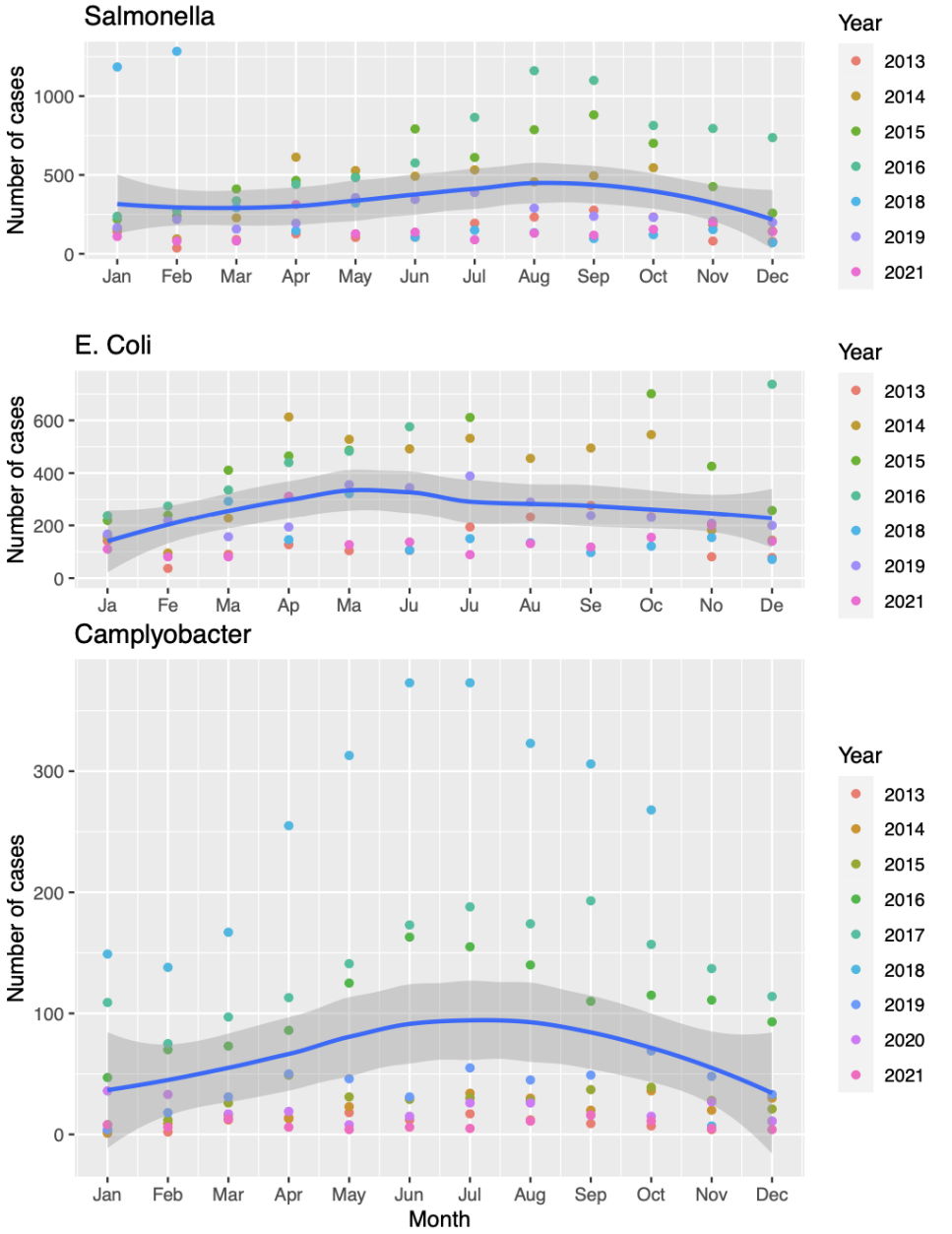

Figure 2: Number of cases per month, by Illness type

We further expand upon our analysis of number of cases over by looking at the number of cases per month for each given illness. For the illnesses, there are no month level data that is observed for any of the illnesses in 2022, and there is no month level data for the year 2020 for the Salmonella and E. coli data. Looking at figure 2, it can be observed from the smoothed regression line fit to the data that there are some potential trends in the seasonality. Notably the number of Salmonella cases looks to increase in early summer before hitting a peak around august and September before falling in the fall months. E. coli on the other hand seems to start increasing in the winter months and peak around May/June before generally starting to decrease. Out of all the illnesses Campylobacter shows the most distinct trend in seasonality, with the number of cases increasing from the winter and spring to a peak in the summer around July and starting to decrease in the later summer/fall into the winter. We note that as mentioned in the EDA there is a high level of missingness in the month of the collection date, which limits our analysis, even though we do observe some trends over time with this subset of the data. We next want to look to spatial data, but given the grunality needed to look at data on a state level we decide to look at data by regions of the United States.

Table 2: Percent of each years total cases by region of US by Illness

year	USA, General	Midwest	Northeast	South	West
Salomella					
2013	0.13	0.19	0.13	0.32	0.24
2014	0.19	0.14	0.09	0.29	0.29
2015	0.31	0.22	0.1	0.23	0.14
2016	0.48	0.18	0.04	0.21	0.09
2017	0.08	0.27	0.07	0.45	0.13
2018	0.18	0.23	0.09	0.35	0.14
2019	0.03	0.31	0.11	0.41	0.14
2020	0.13	0.27	0.1	0.36	0.13
2021	0.27	0.23	0.08	0.31	0.11
2022	0.37	0.15	0.06	0.37	0.06
E. Coli					
2013	0.32	0.15	0.07	0.17	0.29
2014	0.27	0.23	0.07	0.3	0.13
2015	0.46	0.04	0.11	0.32	0.07
2016	0.3	0.08	0.11	0.16	0.34
2017	0.25	0.23	0.14	0.27	0.11
2018	0.25	0.25	0.12	0.23	0.15
2019	0.07	0.25	0.09	0.37	0.22
2020	0.14	0.25	0.13	0.35	0.12
2021	0.31	0.3	0.07	0.22	0.1
2022	0.57	0.12	0.11	0.17	0.03
Campylobacter					
2013	0.12	0.29	0.11	0.18	0.3
2014	0.52	0.16	0.02	0.04	0.26
2015	0.13	0.19	0.11	0.29	0.28
2016	0.29	0.19	0.09	0.24	0.2
2017	0.29	0.2	0.09	0.25	0.17
2018	0.34	0.16	0.07	0.33	0.11
2019	0.03	0.2	0.09	0.54	0.14
2020	0.02	0.16	0.09	0.61	0.12
2021	0.09	0.15	0.09	0.56	0.11
2022	0.14	0.15	0.1	0.52	0.09

From table 2 above we note that for all the illnesses that the percent of total cases by region is generally variable, over time. Though we can see that for Salmonella and Camylobacter that the Southern region of the United States generally has the highest percent of cases in any given year, especially in more recent years. On the contrary the Northeast generally has the lowest percentage of

cases across all of the illnesses we explore. In E.Coli it appears that in earlier years there is generally a higher percent of the cases in the Western Region of the United States, though in later years it appears that the percent of cases in the US lowers for the West, and starts to increase for the Southern United States. Thus it appears that when looking into outbreak prevention it may be beneficial to consider focusing on the South as they have higher rates of food borne illness. Though we do concede that these are raw percentages that fail to account for population density in these regions which could change our conclusions.

4.4 Sources of disease

We next want explore causes of disease, which requires us to look at the new variable *Isolation.Source.category* that we created in the preprocessing steps. We hope this will be indicative of large contributors to food borne illness that we can use in future modeling procedures.

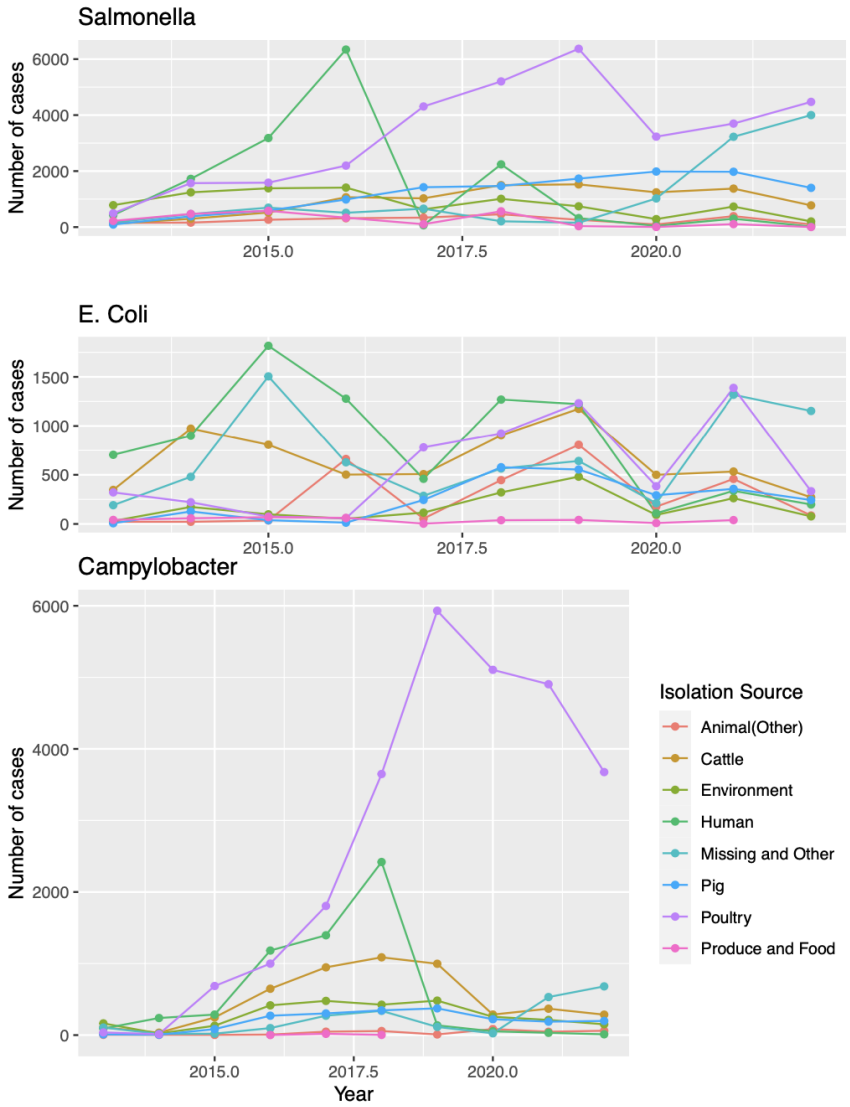

Figure 3: Number of cases by Isolation Source

We note that looking at figure 3, that the trends over time the Isolation source generally indicate

poultry is one of (if not the biggest) causes of the illnesses we have included. Though we do note that the number of Isolation from humans is quite large for all illnesses as well. Other sources such as produce and pig based sources seem to be relatively low across all 3 illness, which could indicate that in our modeling that we may want to focus on sources that have higher incidence for our illnesses over time as it may be easier to capture an outbreak or conduct surveillance in this way.

4.5 Closely Related

We recall from the NCBI data that *Min.same* tells us the closeness to other strains that were isolated from the same source (clinical or environmental) [2]. As we recall from our discussion with Subject matter expert Dr. Julian if isolates are within 7 SNPs they are closely related to each other [9]. Since we want to look at outbreaks we assume that these outbreaks would be when isolates that are close to others genetically are popping up in the same region that we could call this an outbreak. Given the level of missingness in Date information that we will consider the trends over month, though it would be more beneficial to look at trends by week. To begin with this analysis we first consider how closely related the reported pathogens are based on the threshold Dr. Julian lad out.

Table 3: Percentanges regarding Closely related SNPs by Illness

Minimum Same	Percent
Salmonella	
Missing	0.06
Closely related (within 7 SNPs)	0.67
Not closely related	0.27
Ecoli	
Missing	0.53
Closely related (within 7 SNPs)	0.29
Not closely related	0.18
Campylobacter	
Missing	0.16
Closely related (within 7 SNPs)	0.6
Not closely related	0.24

We note from table 3, that of the observations with the Min.same information collected that a high percent (of observed data) of the observations are closely related by our definition. Given that the data is longitudinal in nature over 9 years with many regions, and are within our threshold of highly related we impose some restrictions to find substantial outbreaks in our historical data. We a define a substantial outbreak as a month where over 100 cases were reported and the average *Min.same* was less than 3 SNPs in a given region.

Table 4: Average (Standard Deviation) Min.same for closely related strains when more than 100 cases occur in a region, Salmonella

Year	Month	Number of cases	Midwest	South	West	Northeast
2013	Oct	119	1.2 (2.81)			
2014	Jun	117	1.4 (2.15)			
2014	Oct	254	2.83 (5.53)			
2015	Aug	240		2.58 (4.78)		
2015	Jul	118		2.14 (5.17)		
2015	Jun	247		2.41 (5.2)		
2015	Nov	139	2.41 (4.48)			
2018	Feb	111	2.07 (5.92)			
2018	Mar	154	1.49 (5.44)			
2018	May	176			1.63 (4.83)	
2019	Jun	183				2.64 (3.89)
2021	Apr	219	0.85 (2.74)			
2021	Nov	113		2.56 (5.7)		

Table 5: Average (Standard Deviation) Min.same for closely related strains when more than 100 cases occur in a region, E. Coli

Year	Month	Number of cases	West	Midwest	South	USA
2013	Nov	130	1.71 (2.52)			
2014	Nov	117		1.03 (1.85)		
2015	Jun	310			1.54 (4.8)	
2015	Sep	133			2.28 (6.53)	
2018	Aug	114				0.83 (1.52)
2018	May	139				1.59 (1.65)
2019	Apr	198			1.78 (4.89)	

We note that the data regarding Salmonella, shown in table 4, has the most instances (13) of substantial outbreaks, while there are also some instances with E. coli (7), shown in table 5. We observe that when we look across the years and months for campylobacter none of these instances meet our threshold for inclusion as the largest outbreak in our data for this pathogen was September 2018 when there were 43 closely related cases of campylobacter in the Southern United States. Looking at table 4, we observe that for salmonella in 2015 the 3 summer months we had multiple cases of highly related strains in the Southern United States, which are likely linked, similarly we observe that in March and February of 2018 there is a large amount of closely related cases in the Midwest. No apparent trends pertaining to a certain regions stick out in the E. coli data.

5 Conclusions

The literature on foodborne illness supports outbreak surveillance as a meaningful approach to reducing incidence of foodborne illness in the US - the goal of Healthy People 2030 Foodborne Illness Reduction. Specifically, outbreak detection is a core component of public health monitoring of foodborne illness, and research in this field is ongoing. We plan on leveraging the extensive NCBI FoodNet database for analysis of spatial and temporal trends in outbreaks for Salmonella, as well as E.coli, and Campylobacter.

The data collected and cleaned in our initial work seems well suited for our analysis. Most importantly, the data includes information on the time and location of isolate collection, allowing for the construction of time series models, which were very prevalent in our exploration of outbreak detection in the literature. We plan on drawing from established approaches to this problem such as machine learning classification of outbreaks and hidden markov models. Additionally, we identified the source of disease and genetic similarity variables as important factors to investigate further. Our project is intended to be a practical analysis to inform future outbreak detection of foodborne illness in the US.

References

[1] Home - pathogen detection - ncbi.

[2] Ncbi - pathogen detection - ncbi.

[3] Table 2: CDC estimated annual number of episodes of illnesses caused by 31 pathogens transmitted commonly by food.

[4] The NCBI Pathogen Detection Project [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. 2016 May [cited 2022 OCT 23]. Available from: https://www.ncbi.nlm.nih.gov/pathogens/.

[5] Luma Akil and H Anwar Ahmad. salmonellainfections modelling in mississippi using neural network and geographical information system (gis). *BMJ Open*, 6(3), 2016.

[6] Yi Du and Yunchang Guo. Machine learning techniques and research framework in foodborne disease surveillance system. *Food Control*, 131:108448, 2022.

[7] Yi Du, Hanxue Wang, Wenjuan Cui, Hengshu Zhu, Yunchang Guo, Fayaz Ali Dharejo, and Yuanchun Zhou. Foodborne disease risk prediction using multigraph structural long short-term memory networks: Algorithm design and validation study. *JMIR Medical Informatics*, 9(8), 2021.

[8] Pranav Goyal, Dara Nanda Gopala Krishna, Divyansh Jain, and Megha Rathi. Foodborne disease outbreak prediction using deep learning. In Manoj Kumar Sharma, Vijaypal Singh Dhaka, Thinagaran Perumal, Nilanjan Dey, and João Manuel R. S. Tavares, editors, *Innovations in Computational Intelligence and Computer Vision*, pages 165–172, Singapore, 2021. Springer Singapore.

[9] Ernest Julian. Questions for dr. julian.

[10] Diane G. Newell, Marion Koopmans, Linda Verhoef, Erwin Duizer, Awa Aidara-Kane, Hein Sprong, Marieke Opsteegh, Merel Langelaar, John Threfall, Flemming Scheutz, Joke van der Giessen, and Hilde Kruse. Food-borne diseases — the challenges of 20years ago still persist while new ones continue to emerge. *International Journal of Food Microbiology*, 139:S3–S15, 2010. Future Challenges to Microbial Food Safety.

[11] Elaine Scallan, Robert M. Hoekstra, Frederick J. Angulo, Robert V. Tauxe, Marc-Alain Widdowson, Sharon L. Roy, Jeffery L. Jones, and Patricia M. Griffin. Foodborne illness acquired in the United States—major pathogens. *Emerging Infectious Diseases*, 17(1):7–15, 2011.

[12] Nigam Shah. Faculty opinions recommendation of machine-learned epidemiology: Real-time detection of foodborne illness at scale. *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature*, 2018.

[13] Ryan B. Simpson, Bingjie Zhou, and Elena N. Naumova. Seasonal synchronization of foodborne outbreaks in the united states, 1996–2017. *Scientific Reports*, 10(1), 2020.

[14] Danielle Tack, Logan Ray, Patricia Griffin, Paul Cieslak, John Dunn, Monique Duwell, Alison Muse, Sarah Lathrop, Rachel Jervis, Tamara Rissman, and et al. Preliminary incidence and trends of infections with pathogens transmitted commonly through food - foodborne diseases active surveillance network, 10 u.s. sites, 2016–2019, Apr 2020.

[15] Aydin Teyhouee, Sara McPhee-Knowles, Chryl Waldner, and Nathaniel Osgood. Prospective detection of foodborne illness outbreaks using machine learning approaches. In Dongwon Lee, Yu-Ru Lin, Nathaniel Osgood, and Robert Thomson, editors, *Social, Cultural, and Behavioral Modeling*, pages 302–308, Cham, 2017. Springer International Publishing.

[16] Benedikt Zacher and Irina Czogiel. Supervised learning using routine surveillance data improves outbreak detection of salmonella and campylobacter infections in germany. *PLOS ONE*, 17(5):1–14, 05 2022.

[17] Peng Zhang, Wenjuan Cui, Hanxue Wang, Yi Du, and Yuanchun Zhou. High-efficiency machine learning method for identifying foodborne disease outbreaks and confounding factors. *Foodborne Pathogens and Disease*, 18(8):590–598, 2021. PMID: 33902323.

[18] Shaokang Zhang, Shaoting Li, Weidong Gu, Henk den Bakker, Dave Boxrud, Angie Taylor, Chandler Roe, Elizabeth Driebe, David M. Engelthaler, Marc Allard, and et al. Zoonotic source attribution ofsalmonella entericaserotype typhimurium using genomic surveillance data, united states. *Emerging Infectious Diseases*, 25(1), 2019.