Development and Assessment of Tools for Classification of Foodborne Pathogen Outbreak Cases

Timothy Hedspeth[1], **Anthony Girard**[1], **Yutong Li**[1]

[1]Department of Biostatistics, Brown University

Abstract

Background The Healthy People 2030 [1] sets data-driven national objectives aimed to prevent foodborne illness to improve health over the next decade. Though many previous studies focused on outbreak detection and prediction, well-defined definitions of an foodborne outbreak and more easily interpretable model are lacking.

Method Using the NCBI foodborne pathogen detection data of two bacterial pathogens: *salmonella* and *E. coli* from 2013 to 2022, we first defined outbreak based on the number of the same SNP cluster in a given year/month (more than 8 for *salmonella* and more than 10 for *E. coli*). Then, we constructed a risk score model to determine the foodborne outbreak with integer score assigned to variables including SNP cluster, AMR genotype, isolation source and region. In addition we build tree model to investigate classification of an outbreak based on time, location and source.

Result For *Salmonella* risk score model, three classes of SNP cluster and isolation source of Production and Food will increase the risk of a case being an outbreak by scores ranged from 1 to 3. A *Salmonella* illness case with region of USA, general will decrease the risk by a score of 2. For *E. coli* risk model, one SNP cluster, AMR.genotype and region of Northeast will increase the risk to be an outbreak by a score of 2, 2 and 3 respectively. A *E. coli* illness case with isolation source of human or poultry or region of USA, general will decrease the risk by a score of 1, 3 and 7 respectively.

Keywords: *Foodborne Pathogens; Outbreaks; Classification tools; NCBI database*

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 foodborne diseases yearly, a small subset leads to hospitalization, and an even smaller subset ends up contributing to death [12]. 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 [15]. 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 [14]. The often reported illnesses and occurrence of foodborne outbreaks have made the study the foodborne pathogens a necessity, namely outbreak detection and prediction are needed to help identify trends to inform actions could be taken to prevent future outbreaks.

Previous papers have adopted different approaches to study outbreak detection such as using time-series analysis to model the seasonality of outbreaks and whole-genome sequencing analysis to determine outbreak detection based on genomics information of foodborne pathogens [7, 10]. Though those study results provided insights for foodborne outbreak investigation and proposed method to assess seasonality in broad spectrum of illness to develop reliable outbreak forecasts [10], the model interpretation and construction might be complicated to audiences without sufficient background in biology or statistics. Therefore, in this study, we built a risk score model following a method similar to one described in the paper by Baik et al., (2020). To our knowledge a risk score model has not been developed to assess the risk that a case of a foodborne illness is part of a wider outbreak. A risk score model assigns an integer risk score to each potential determinant of a foodborne outbreak and the interpretation of the risk score relies on a simple tabulation resulting that yields a risk. The relative simplicity of the score makes it interpretable to a broad audience, a generalization that can't be made for other statistical models. In addition we utilized classification trees to predict if cases were considered to be part of an outbreak, as the results of these models are a simple tree that resembles a flowchart that when effectively presented can help classify if a case that is part of an outbreak with relative ease. Both methods are used so that the commonality and differences in the resulting models can be assessed.

This study aimed to build easily interpretable models to aid in the detection of foodborne outbreaks and investigate classification methods of reported illness cases to be an outbreak. 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 [1].

2. Literature Review

Newell et al., (2010) notes that the bacterial foodborne pathogens *Salmonella* and *E. coli* are the top reported foodborne agents. *Salmonella* has been paid considerable attention, this bacteria colonizes a broad list of hosts and livestock animals such as pigs, cattle and poultry [9]. *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 [9]. Contaminated food sources will be processed and put in the market. To relieve the burden of bacterial pathogen infections and often outbreaks, foodborne illness surveillance plays a critical role in limiting the impact of disease [9]. 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 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.

The objective of our work is to help in the surveillance and potentially predict outbreaks of foodborne illness, which is a challenge given the rarity of these outbreaks, as noted by the CDC. One paper looks at *Salmonella Typhimurium* that have 4 sets of genomes, for selected human isolates from sporadic cases dating 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 which is a DNA sequence that resembles a gene but mutated into an inactive form over the course of evolution [8]. The RF algorithm was able to rank genetic features by their importance. The algorithm found 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 [18].

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

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 [13]. With advances in statistical methodologies it makes sense to leverage these techniques to help predict an outbreak. In the report it is noted that 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. Shah et. al. 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.

Using data from the CDC Akil and Ahmad sought to predict *salmonella* infections in Mississippi, as southern parts of the United States are more vulnerable to foodborne illness [4]. Akil and Ahmad used lab confirmed E. coli and Salmonella cases from 2002 to 2012 for selected states. Their data was grouped by the year and district, and adjusted the cases to 100,000 of the population. Akil and Ahmad 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].

In addition, risk score models for assessment of a disease using risk factors is also a helpful and more interpretable approach for risk prediction. Previous studies developed risk score models to evaluate the risk of disease on aspects of health such cardiovascular events risk and active tuberculosis [5, 11]. From the paper by Baik et al., (2020) constructing a risk score to evaluate TB risk, the risk factors to include in the model will be associated with the TB risk (excluding variables measured for reasons other than TB risk)[5]. The final risk score was then fitted to a logistic regression taking the TB status as the dependent variable [5]. Baik et. al.'s model showed good performance, despite the simplicity of the risk score, reasonable predictive accuracy was observed. This method shows that a risk score model considering risk factors such as time, location and biological factors related to foodborne illness is a feasible approach to predict outbreaks.

This initial review of literature containing methods and limitations for handling outbreak detection of foodborne illness informed us of analyses that would be beneficial to the goal of the Healthy People 2030 Foodborne Illness Reduction Committee. These references were taken into consideration as we developed our models to help public health, as the methods found regarding foodborne illness are rather complex in nature, and further motivated our goal of developing easily accessible tools for public health officials.

3. Methods

3.1. Pre-processing

We extracted data from the NCBI pathogen detection website [2] regarding 2 different foodborne pathogens, *Salmonella* and *E. Coli*, from the years 2013 to 2022. We believed selecting this range of years would allow us to develop a relevant outbreak model for classifying future cases, as we expect the trends in coming years to be more similar to this range of time compared to others (e.g. 1990-2000). 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 *Salmonella* data set contains 93,763 observations and *E. Coli* has 34,805 observations.

In the initial analysis we noted some preprocessing and cleaning was required. In our first R script for data cleaning, we subset the data to observations of the two illnesses of interest that

were within the United States. Then, we noted that the *Location* variable does not have a uniform format (136 levels in Salmonella data and 291 levels in E. coli data). We cleaned the location to have a standard two letters state abbreviation which resulted in 56 levels for both data. We further grouped the states into 5 regions: Midwest, Northeast, South, West and USA, General (Unclarified location from the data). No missing values were found for the *Location* variable.

After running the first 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 source group cleaning R script 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 date cleaning R script. 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.

Given that a goal of our analysis was the prediction of outbreaks we felt it was necessary to create an indicator of whether a case was part of an outbreak. The definition of an outbreak is unclear as we recall from the literature review, and the **Outbreak** variable that was available through the NCBI data base was mostly missing as seen in table 1 below. We decided that it was preferable to create our own variable for this propose. We created two functions that classifies outbreaks, the first of which classifies an outbreak when there are more than a given number of cases from the same **SNP.cluster**, that were collected in the same month and year. Given that **SNP.clsuter** is used to define this outbreak we reduce the levels of this factor after an outbreak has been classified. In another instance we consider cases that are very similarly related to other isolates from the same isolate in the same region of the US, though it is not used for the primary analysis, and rather could be used in future work.

In the exploration of the data we noticed that many of the variables had levels that were unique to observations, in order to make these variables more informative we set an arbitrary threshold that the top 7 levels in terms of frequency are kept while the other levels are assigned to be in an other category. If the highest of 7 levels in terms of frequency is not greater than 10 then we decide that there is likely not enough information in this variable, and subsequently drop these variables from our analysis. We concatenated this process to be a function that would allow us to easily clean many variables in our analysis.

We found it of interest to impute the missing information for the day of the **Collection.date**, as this is largely missing, (table 1). While the **Creation.date** is fully observed, it was noted that **Creation.date** was a natural candidate to be used for imputing collection date. Though the distribution of the differences for complete cases raised concern (very high variability) and thus we could not utilize the day of the week to define an outbreak even though it was of interest.

3.2. Exploratory Analysis

The NCBI collects data from surveillance and research efforts that are currently ongoing [2]. The data contains a multitude of sources for foodborne illnesses such as food, patients, production facilities, etc [2]. After data is submitted it is clustered to related pathogens, allowing researchers to look for strains of pathogens that are closely related [2]. Given that the data is being uploaded from multiple sources with what we presume to be free text fields in some columns as in the Isolation source variable there is a lot of heterogeneity in reporting and in the data's quality. The quality of some responses 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
Strain	0.01	0.07
Serovar	0.22	0.87
Host.disease	0.99	0.90
Isolation.source	0.11	0.15
Lat.Lon	0.98	0.81
Source.type	1.00	1.00
SNP.cluster	0.06	0.47
Min.same	0.06	0.53
Min.diff	0.17	0.84
Assembly	0.11	0.07
Outbreak	1.00	1.00
day	0.85	0.67
month	0.69	0.54

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* and *Source_ type*. Given the extreme missingness for the aforementioned 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, and in most cases is optional [2]. The *Strain*,*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 Salmonella, but not nearly to this extent. This could be due to issues or delays in processing the E. coli pathogens and its nature of constant mutation. 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.

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:

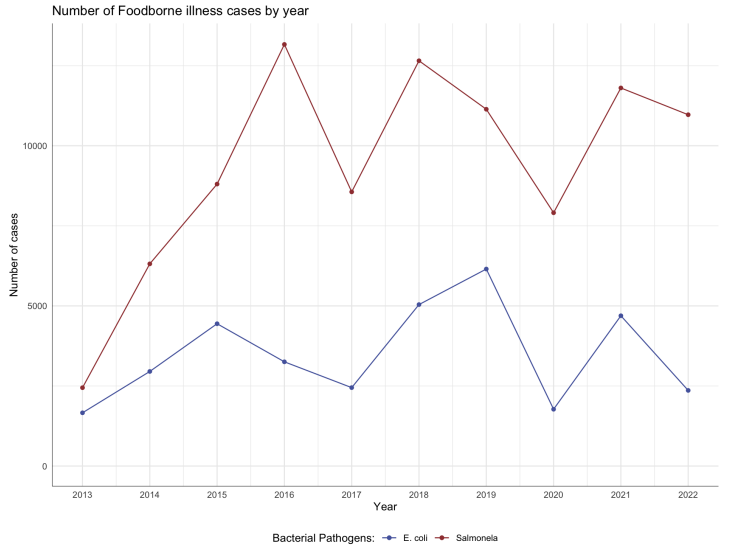

Figure 1: Number of cases per year, all illnesses

From Figure 1 we note that there are more cases of Salmonella than E. coli in any given year. We observe that for both illnesses the number of reported cases decreased 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 a 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.

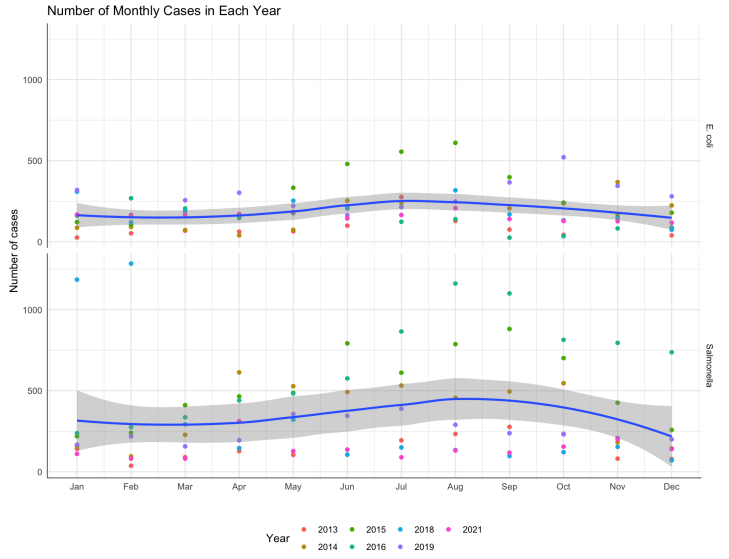

Figure 2: Number of cases per month, by illness type

We further expand upon our analysis of number of cases over time by looking at the number of cases per month for each illness. There is no month level data observed for any of the illnesses in 2022, and there is no month level data for the year 2020. Looking at figure 2, it can be observed from the smoothed regression line fit to the data indicates 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. We note that as previously mentioned there is a high level of missingness in the month of the collection date, even though we do observe some trends over time with this subset

of the data. We next sought to assess biological trends, over time for the included pathogens.

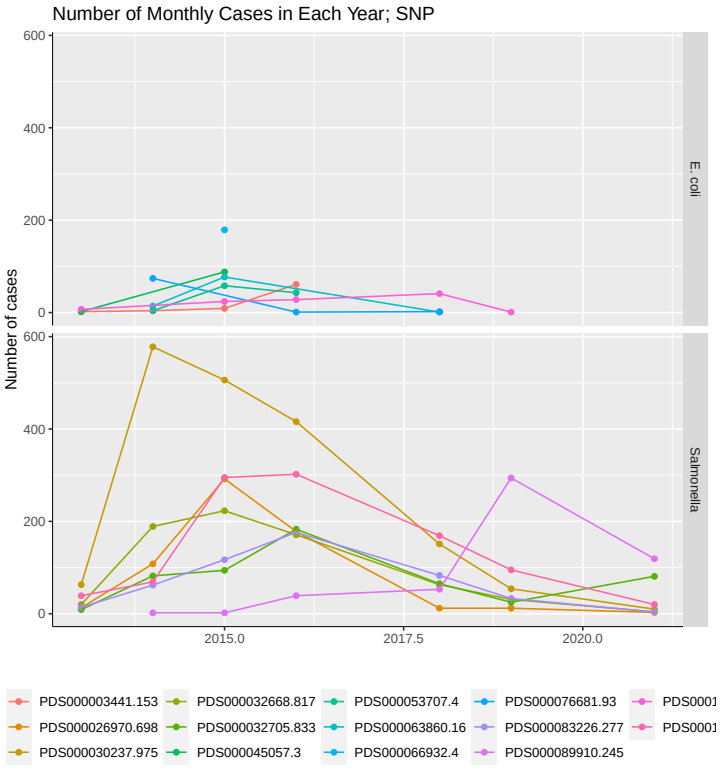

Figure 3: Number of cases from each of the top 7 SNP clusters over our time frame

Figure 3 above shows us the trends in SNP cluster prevalence over the years of the study. We can see that the pathogen E. coli's top SNP clusters generally do not have a high frequency or persist for long periods of time. Where as the SNP clusters in Salmonella generally have a pretty high frequency at the start of our study and their prevalence drops over time. It appears that generally over time the prevalence of SNP clusters wane for both pathogens, likely due to evolution and mutations though we do not have the biological training to make concrete claims, rather this comes from our intuition based on the trends we observed in the data. We finally move to exploring the trends in cases with regard to their location.

From table 2 below 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 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 the two 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.

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

year	USA, General	Midwest	Northeast	South	West
Salmonella					
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

3.3. Modeling

As we recall the goal of our analysis is to hone in on and better understand foodborne pathogen outbreaks. The definition of an outbreak is variable between different public health organizations [16, 3]. This variability led to define our own outbreak criteria based on the data. For the purposes of this analysis, an outbreak was considered to be more than 8 for Salmonella and more than 10 for E. coli from the same **SNP.cluster** in a given year/month. This outbreak definition allows for us to consider genetically similar cases popping up across the country as being part of an outbreak. We felt taking into account this genetic information would best capture an outbreak, as a high number of cases in a region may be genetically different, which we would consider to be more spontaneous in nature.

When considering what approaches to take for our modeling, we placed an emphasis on the utility of the models that we were selecting. To maximize said utility we selected risk score models and classification trees to predict if a case was part of an outbreak. Due to the large degree of missingness in most of the variables, it was decided that imputation was not a suitable approach, and thus a complete case analysis was utilized, thus removing variables with a large degree of missingness was imperative to maintain sample size. After removing variables that were highly missing the risk score model still had many predictors to estimate (>40) for our data sets. Due to the large number of potential covariates Lasso regularized logistic regression using 12 fold CV was implemented to aid in the selection of important coefficients. Given that our outbreaks are associated based on the month they were observed our cross validation split the data by month, to avoid the bias that would result from using cases from the same outbreak in the test and train splits. The coefficients of the resulting model were divided the by the median of the non-zero coefficients and rounded to the nearest integer value. Thus, we utilized a similar methodology to Baik et al., in the development our risk score models [5]. In order to assess this model a logistic regression was fit using just the score associated with each observation as the predictor of an outbreak. In addition to this Risk score model classification trees were fit using the rpart package in R to classify cases that were outbreaks. We fed the same candidate predictors used in the risk score models into these classification trees. The tree was pruned to the complexity parameter that minimized the relative error, which was assessed via a plot of the error vs complexity at a given number of splits (not

shown in this report, but are available on our github in the Models.and.results.for.final.report.pdf) to avoid over fitting. For both of our approaches we considered the utilization of interaction terms that were decided a priori based on our understanding of the data. Though these models became overly complex and confusing yielding parameters as large as $10 * 10^{10}$. Given that these models were intended for a general audience, simpler models with reasonable risk score coefficients were preferential, so the models without interactions were selected.

Prior to our model fitting we split our data into testing and training sets for both pathogens. The years we selected for training were 2013-2018, where as the years 2019 and 2021 (2020 and 2022 had no month level data) were selected for testing our models. Both of the models were assessed for their discrimination and calibration in the testing and training set. We discuss the performance of the models on the training data, but do not present the full table of results here, on the other hand all the training data results presented in the results (full results for training sets are available on our GitHub). All of the code for this analysis is available in *OurGitHub*.

4. Results

4.1. Modeling

As discussed in the methods, we attempted models that included and did not include interaction terms. The results of the models that included interactions were not feasible for the goals of creating an easily interpretable model for individuals that are not versed in statistics, though the code is available in our GitHub repository. The focus of our results will be on our models that did not include interaction terms. We first describe the models (4.2.1) and then assess their performance (4.2.2).

4.1.1. Resulting Models

In our salmonella data we had defined an outbreak to be 8 cases (same SNP.cluster in a given year and month). This resulted in 15% of cases being classified as part of an outbreak in the training data and 14% of cases in the testing data being part of an outbreak. Where as a threshold of 10 cases for E. coli resulted in 15% of the cases in the training data set being part of an outbreak and 9% of cases in the test set were part of an outbreak. Using the aforementioned methods we derived the following risk score model for Salmonella shown below in table 3.

Table 3: Score coefficents for Salmonella Risk model

	Score
SNP.clusterPDS000030237.975	3
SNP.clusterPDS000032668.817	1
SNP.clusterPDS000120941.3	1
Computed.typesantigen_formula=9:g,m:-,serotype=Enteritidis	1
Isolation.source.categoryProduce and Food	1
regionUSA, General	-2

We are able to see in table 3 that of the variables included in the risk score model the only one that has more than one level with non-zero scores is the SNP cluster. The SNP cluster shown in the first line has a score of 3, the highest of any in magnitude regardless of sign, indicating its the most important piece of information in determining if a case is part of an outbreak. Following this we note that when we do not know the exact region of the country that the case was identified in the score is penalized, indicating granular level regarding location is important in determining the risk. The other two SNP clusters, one of the computed types, and having a sample derived from produce or food, all increase the risk of the case being in an outbreak by 1 point. With an understanding of our risk score model for salmonella we next move to our tree based model.

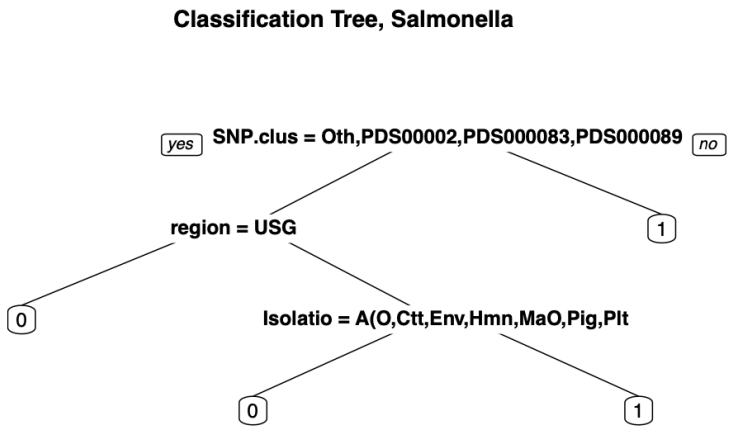

Figure 4: Classification Tree for Salmonella

Figure 4 shows the tree for Salmonella, we note that the tree first splits on SNP cluster (indicating it is the most important factor), the SNP clusters 4723.723, 30237.975, 32668.817 and 120941.3 are classified as outbreaks regardless of their other attributes. For the other SNP clusters if the case does not belong a specific region of the US it is classified as not being an outbreak, whereas cases that have this information are split once more. This final split is determined based on the isolation source, if a case is from produce or food then it is classified as being part of an outbreak, other levels of this isolation source indicate that the case is not part of an outbreak. We observe that the results between our tree and risk score model show good agreement in terms of the most important factors for determining that if the case is part of an outbreak. We now consider our models for E. coli starting with the risk score model.

Table 4: Score coefficents for E. coli Risk model

	Score
(Intercept)	1
SNP.clusterPDS000076681.93	2
AMR.genotypesacrF=MISTRANSLATION,blaEC=COMPLETE,mdtM=COMPLETE	2
Isolation.source.categoryHuman	-1
Isolation.source.categoryPoultry	-3
regionNortheast	3
regionUSA, General	-7

We can see from table 4 that for E. coli the coefficient of the highest magnitude is not having location data at the region level deducting 7 points from the risk score. Though lack of information on region is not the only coefficient that reduces the risk of being in an outbreak as isolates from humans and poultry reduce the risk of the case being part of an E. coli outbreak. Though we do observe that being in the 76681.93 SNP cluster, having a mistranslation in AMR.genotype and being a case in the northeast increase the risk of a case being part of an outbreak.

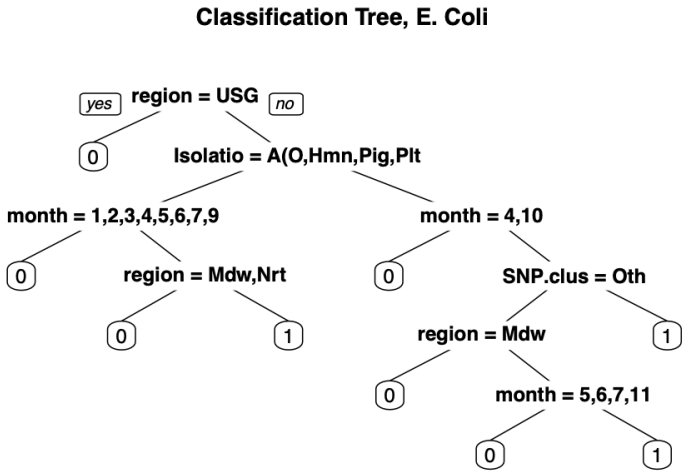

Figure 5: Classification Tree for E. Coli

We now turn our attention to the classification tree for E. Coli, shown above in figure 5, which first splits on the case having information on the region from which it was collected. If the case does not have information regarding the region then it is classified as not being part of an outbreak, where as cases with this information are split further. The tree splits the cases with region information based on the isolation sources, if the case is from an Animal (other), human, pig, or poultry then splits based on month and region occur to classify an outbreak. If the Isolation source is Cattle, Environmental, Other, or from Food then the tree makes a split based on month, then SNP cluster before splitting on region and month again to classify the case. The variables that have the most impact in the E. coli models are generally the same between the two (region and Isolation source). It is evident from tables 3 and 4, and figure 4 that this tree (figure 5) is the most complex model that we have, yet at most requires 6 decisions before classifying an outbreak. Showing that our resulting models are simple and parsimonious, achieving the goal of creating simple models that could be used by individuals with no statistical training to predict an outbreak.

4.1.2. Model Performance

With an understanding of the models and the decisions that they make we sought to examine how the models performed with the training data. For the training data sets the models all showed high AUC (range: .88-.97) and low Brier Scores (.06-.13), with Accuracy between .86 and .93, indicating good discrimination and calibration for all models. The sensitivity was also high for all our training data sets the only pace where we observe something concerning is the risk score model for salmonella having a specificity of .57. Though the models do well on the training data our objective is to develop tools that could be used in practice. Given the need for practical utility we find it of utmost importance to see how the models perform on our testing set. This was done by using the models to predict if cases in the test set were outbreaks or not, and assessing their performance with the same metrics.

Table 5: Metrics for models on testing data

	Risk score; Salmonella	Tree; Salmonella	Risk score; E. Coli	Tree; E. coli
AUC	0.56	0.54	0.79	0.58
Brier Score	0.18	0.25	0.08	0.19
Accuracy	0.76	0.70	0.91	0.72
Sensitivity	0.99	0.91	1.00	0.76
Specificity	0.00	0.00	0.00	0.27

We observe that in our with held test sets that both models for Salmonella have an AUC

below .6 and above .5, with the risk socre model performing marginally better. Though these AUCs indicate that the model is doing only a little better than random guessing for classifying Salmonella outbreaks. Though when we look at our models for E.coli the risk score model has an AUC of .79 for while the tree has an AUC of .58. This indicates that out of all our models the one with the best predictive accuracy is the Risk score model for E. coli. We can see for all the models that even though sensitivity is high, specificity is not, meaning the models are struggling to predict outbreaks in the test set. We note that the models for E. coli have lower Brier scores, indicating that these models have better calibration and therefore is doing a better job predicting the probabilities of being part of an outbreak than our models for Salmonella. We lastly assess the performance of our risk score models by examining the percent of cases with a given risk score that are classified as outbreaks in tables 6-9 below.

Table 6: Salmoella test set

score	Percent outbreaks
-2	0.00
0	26.41
1	0.00
2	0.00
3	0.00
4	0.00

Table 7: Salmonella training data

score	Percent outbreaks
-2	0.92
-1	9.19
0	15.14
1	31.63
2	58.02
3	80.00
4	91.43
5	100.00

Table 8: E. coli test set

score	Percent outbreaks
-8	0.00
-7	0.00
-6	0.00
-3	0.00
-1	0.00
0	18.93
1	92.31
2	21.43
3	0.00

Table 9: E. coli training data

score	Percent outbreaks
-8	0.00
-7	0.36
-6	0.00
-5	0.00
-4	0.00
-3	0.00
-1	7.24
0	13.57
1	47.37
2	80.19
3	76.90
5	0.00

We note from table 7 that when looking at the scores for Salmonella in the training set the percent of cases that are part of outbreaks steadily increases as the score does, though the negative scores do have some cases that are part of outbreaks. When looking at table 6 it is easy to note that all the cases that are part of outbreaks in the data are given a score of 0, meaning that the scores had no bearing on predicting if a case was part of an outbreak in this data. Table 8 shows the scores for the training data in E. coli, which has many scores that are negative, which for the most part do not contain cases associated with outbreaks. But as the score increases from -1 the percent of cases that are part of outbreaks increases, but after a score of 2 the percent of cases classified as being part of an outbreak decreases swiftly. Lastly we note from table 8 that in our test set no negative scores contain a case that is part of an outbreak, interestingly most of the cases with a score of 1 are classified as being part of an outbreak.

5. Discussion

Foodborne illness outbreaks are not clearly defined, usually resulting in outbreaks being loosely defined as a relatively high number of cases for a given location and season [16]. Though, attempting to meaningfully detect an outbreak requires a more stringent set of criteria. In our work we propose initial data-defined definitions of outbreaks for use in analysis of foodborne illness data. Although our definitions would benefit from better validation, including guidance from domain experts, they are intuitive approaches to this problem of outbreak classification where no clear solution exists.

Our goal of predicting if a case was part of an outbreak may not have resulted in the most robust models in terms of future prediction, as evidenced by tables 3 and 4 above, though they did illuminate some interesting patterns. In all models SNP cluster was important in classifying if a case was part of an outbreak or the risk that it was part of an outbreak. In the models for salmonella the only thing that differed was that the risk score model included a computed type coefficient for calculating the score and another SNP cluster was included in the tree to to classify an outbreak. Though the results of the E. coli models show more drastic differences, with the tree taking into account time and region across many splits.

We note that generally the models for E. coli perform better than the models for Salmonella, and we suspect that this may be due to salmonella's emphasis on SNP cluster in the models (high scores and first split). Which makes sense as we plot the frequency if high observed SNP clusters in Salmonella we see that there are many SNP clusters that thrive for the early years and the number of cases shrinks over time, which makes the prediction of future cases difficult when SNP cluster is the most important piece of information. Where as the models for E. coli put more weight on more easily observable variables, such as region and Isolation source and showed better discrimination and calibration.

In part due to a lack of a well defined outbreak definition, we found it to be very difficult to correctly classify or determine the risk of single cases being part of outbreaks. In addition, year to year changes in genetic information of foodborne illnesses complicates the predictive ability of genetic variables in future data. We found that varying the conditions determining inclusion in an outbreak did produce different results. Though this approach could work well if information linking new SNP clusters to relatives was available, adjusting for this could lead to more robust results and even improve our ability to predict an outbreak.

It is our recommendation that the outbreaks be more thoroughly defined in the NCBI database. Currently, there is an outbreak variable included that references an outbreak if applicable, but this variable is highly missing (upwards of 99% for Salmonella and E. coli), greatly limiting its use in data analysis projects. Again, this is due to a lack of a validated measure for outbreak detection. Development of a validated tool to define an outbreak could be beneficial to foodborne illness surveillance systems.

Our analysis suffered from many limitations, the first of which is by having a user defined outbreak variable that is dependent on the SNP cluster. The use of SNP cluster allows for genetically related cases occurring at the same time to be captured, which does add a level of granularity that would not be afforded by simply examining the location that a number of cases is observed. Though due to the large number of SNP clusters observed throughout the time we analyzed led to issues when modeling as we only consider the top SNP clusters as to not overwhelm and overfit our model. Another limitation is that we did not have enough data from the collection dates to look at outbreaks on a more granular scale (weekly) which could have improved our definition of an outbreak, and possibly improved our models performance. A limitation from the perspective of someone not trained in statistics is that the tree models shown here may not be as interpretable (in terms of how it is displayed) as a simple risk score tabulation, the trees with explicit splits are available in our github, though were deemed to be to overwhelming to present here. In future work it would be helpful to take the resulting trees and make them readable in another software, as presentation is limited by the packages available in R.

Though our analysis was limited by many factors it did have its strengths, we were able to derive models that incorporate different aspects of a case, showing that there is signal in this data that can be captured for our stringent definition of an outbreak. Though the greatest strength

is that we were able to derive models are simplistic in nature and take in information that we believe could be obtained by public health officials with relative ease and allow for them to make a quick determination on if a case is part of an outbreak. Though these models need further work and refinement before any clinical use it is reassuring to see that simplistic models can be derived and these models provided reasonable scores and splits for the foodborne determinants such as location, isolation source and biological factors (SNP cluster and AMR genotype). In future work we would like to expand our definition of what cases would be part of an outbreak and compare how different definitions perform to find easily interpretable models that perform well for public health officials. It would also be of interest to expand the pathogens we examine as we explore more outbreak definitions, as different pathogens could show variable performance based on how an outbreak is defined.

5.1. Conclusions

Our analysis highlights an important struggle in the development of easily interpretable models for the detection of foodborne pathogen outbreaks, especially when the definition of outbreak is self-defined due to lack of data and heterogeneity in definition. In the models that were developed for salmonella we note that SNP cluster was identified to be the most important predictive factor, which lead to these models not performing well on testing data as the top SNP clusters prevalence wane over time. Where as the models for E. coli do not place as much emphasis on SNP cluster, and rather place more emphasis on easily accessible information such as the region of the country, and Isolation source, which leads to the risk score model for E. coli being the best in terms of discrimination and calibration of all models that we developed. The relative success of the E. coli Risk score model (high AUC, low Brier score, better correspondence between scores and outbreaks) is a promising first effort in this work, as an easily intrepretable scoring system that places most emphasis on general information could be of great utility for public health officials that do not have advanced statistical training.

References

[1] Health people 2030-foodborne illness.

[2] Home - pathogen detection - ncbi.

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

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

[5] Y. Baik, H. M. Rickman, C. F. Hanrahan, L. Mmolawa, P. J. Kitonsa, T. Sewelana, A. Nalutaaya, E. A. Kendall, L. Lebina, N. Martinson, A Katamba, and D. W. Dowdy. A clinical score for identifying active tuberculosis while awaiting microbiological results: Development and validation of a multivariable prediction model in sub-saharan africa. *PLOS Medicine*, 17:1–23, 11 2020.

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

[7] P. Leekitcharoenphon, E. M. Nielsen, R. S. Kaas, Ole Lund, and F. M. Aarestrup. Evaluation of whole genome sequencing for outbreak detection of salmonella enterica. *PLOS One*, 9, 02 2014.

[8] W. Li, T. Gojobori, and M. Nei. Pseudogenes as a paradigm of neutral evolution. *Emerging Infectious Diseases*, 292(1), 1987.

[9] 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.

[10] K. Ramanathan, M. Thenmozhi, S. George, S. Anandan, B. Veeraraghavan, E. N. Naumova, and L. Jeyaseelan. Assessing seasonality variation with harmonic regression: Accommodations for sharp peaks. *Int. J. Environ. Res. Public Health*, 17, 02 2020.

[11] P. M Ridker, J. E. Buring, N. Rifai, and N. R. Cook. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women. *JAMA Network*, 297:611–619, 02 2007.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[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.