Statistical Machine Learning for Foodborne Disease Source Attribution

Amos Okutse Zexuan Yu Rophence Ojiambo

23 October, 2022

Abstract

Listeria monocytogenes remains one of the most severe causes of foodborne-disease-related disease burden, particularly, due to the severity of its clinical manifestations. This project seeks to use machine learning methods, core genome multilocus sequence typing (cgMLST) data, and other selected information about the sampled Listeria monocytogenes isolates in the United States (US) obtained from the National Center for Biotechnology Information (NCBI) Pathogen Detection Database. We seek to examine these data more closely and employ statistical methods to extract patterns, as well as, substantial predictors of a food source associated with a given L. monocytogene isolate, and consider exploring the particular advantage presented by cgMLST and machine learning in food source attribution. We look to build on the resilience of foodborne pathogens, their ability to adapt both genetically and phenotypically to changing environments, and the particular interest in using this knowledge to understand pathogen responses to food management practices and ultimately reduce foodborne disease incidence by proposing ways of enhancing food safety and improve public health.

Introduction

The burden of foodborne illnesses remains substantially high across the globe. Contaminated food has been implicated in 600 million foodborne disease incidences and 420, 000 deaths per year worldwide with children below five years accounting for one-third of the total fatalities [2]. In the United States, foodborne illnesses result in about 128, 000 hospitalizations and about 3000 fatalities [3]. Known pathogens account for most of the reported cases with most illnesses being caused by noroviruses (58%), followed by non-typhoid salmonella (11%), Clostridium perfringens (10%), and *Campylobacter SPP.* (9%) whereas non-typhoid salmonella (35%), and norovirus (26%) account for the most hospitalizations [4]. *Salmonella enterica*, *E-coli*, and *Listeria monocytogenes* remain the three most common pathogens responsible for most foodborne disease outbreaks, defined as two or more cases of a similar illness resulting from the ingestion of a common food [3, 5].

Listeria monocytogenes remains one of the most severe causes of foodborne-related disease burden despite its characterization with low morbidity, particularly, due to the severity of its clinical manifestations [7]. With the immuno-compromised, pregnant women, the elderly, and infants characterized as being at high risk for listeriosis, it is ranked as the third top cause of foodborne illness-associated deaths in the US [8]. The US Centers for Disease Control and Prevention (CDC) notes that about 1600 cases of listeriosis are recorded annually with about 260 mortalities [9]. Outbreak investigations have shown links between these pathogens and specific food sources, a crucial phenomenon in identifying potential areas of food safety concern including points of contamination and the current performance of foodborne illness prevention strategies [3]. Many foodborne listeriosis outbreaks have been linked to a variety of foods, but mostly to different types of meat [10, 11]. The first laboratory-confirmed outbreak of listeriosis associated with meat products was caused by contaminated turkeys in 1988 [12]. Since then, most meat products have been associated with outbreaks or sporadic listeriosis, including processed vacuum-packed meat products [13], sausage [14], etc. Of course, there are some other types of non-meat food that have been contaminated resulting in outbreaks, such as ice cream and contaminated diced Celery [16]. Listeria outbreaks or sporadic outbreaks have also increased in developing countries in recent years [17].

In recent times, core genome multilocus sequence typing (cgMLST) has been employed to corroborate epidemiological findings, in addition to auditing the effect of public health interventions targeting the food chain on the food reservoirs [18]. This methodology enables differentiation of isolates and can be used to link them to their potential food sources in studies seeking to infer the food source of an outbreak given a pathogenic strain and ultimately result in a reduction in the incidence of foodborne illnesses [18]. Since the pathogens associated with foodborne illnesses are prone to change, understanding the role of these changes in their adaptation to food handling practices is imperative in the effective surveillance of the distribution, as well as, the occurrence of the pathogens. Moreover, the use of genomic data with machine learning methods has gained precedence due to the ability of these methods to learn patterns in high-dimensional data sets which are then exploited in predictive models [18].

Methods

Even though machine learning models promise substantial gains in outbreak investigations, particularly while thinking about the use of cgMLST profiles in foodborne disease source attribution studies, there are a limited number of studies that have explored this avenue while thinking about the gains that these methods promise in allowing exploration of how foodborne pathogens adapt to or respond to food handling practices and how this information can be analyzed and ultimately reduce the incidence of listeriosis in humans. For instance, [3] built a machine-learning model for food source attribution of Listeria monocytogenes using a boosted logit model whereas [19] employed next-generation sequencing using support vector machines with linear kernels to predict the risk of illnesses.

Additionally, Lupolova et al. [20], Munck et al. [18], Tanui et al. [3], and Karanth et al. [21] employed these methodologies in studying *Salmonella enterica*. Varied statistical methods have been employed in analyzing foodborne disease outbreak dynamics. For listeria outbreaks, advanced statistical analysis methods have also come into play including studies by Liu et al. [22] and Vangay et al., [23] which used machine learning methods to provide advice on listeria outbreaks. On the other hand, Sun et al. [24] used Markov chain Monte Carlo (MCMC) to simulate the risk of a Listeria outbreak whereas Pasonen et al. [25] employed a repeated exposures model to assess the risk of a Listeria outbreak in Finnish fish products. Mughini-Gras et al. [26] conducted a meta-analysis of sporadic infection sources of some pathogens including Listeria based on the Bayesian framework whereas Lassen et al., [27] used whole genome sequencing to analyze the risk of listeria outbreaks.

Given the serious threat of foodborne diseases and the high burden posed by listeriosis on human health, this research project seeks to expand the literature on foodborne disease source attribution for human listeriosis using Bayesian and ensemble-based machine learning methods and core genome multilocus sequencing typing data and other selected information about the sampled *Listeria monocytogenes* isolates in the US. In particular, the study seeks to evaluate common food categories and their link to foodborne illnesses using pathogenic isolates. The study seeks to explore the question: Given a human *Listeriosis monocytogene* isolate how likely is it to be from a particular food source? This study is informed by the need to leverage emerging technologies to identify strategies to enhance food safety, and the food production process and ultimately reduce the burden of foodborne illnesses. The prevention of the transmission of foodborne illnesses promises substantial improvements in public health. Modeling periodic human cases of diseases attributable to food sources as well as animal reservoirs informs the public health decision making process [18].

Data overview

Data source

This project uses secondary data downloaded on the 18th October, 2022 from the National Center for Biotechnology Information (NCBI) Pathogen Detection database [29] which assimilates bacterial and fungal pathogen genomic sequences from sources including food, environment, and patient samples. The data are contributed by researchers and public health agencies who sequence samples and submit them to NCBI where the sequences are analyzed and compared to identify relations between sequences and thus aid the

investigation of outbreaks including real-time surveillance of pathogens such as those for foodborne illnesses. The isolates present in the database were collected by 390 different institutions and organizations. Even though the NCBI pathogen detection allows real-time identification of clusters of related genomic sequences to aid in outbreak detection, and track the spread of resistance genes, a potential limitation of this data source is that it does not identify outbreaks or outbreak memberships and analyses rely solely on publicly available data submitted to the database.

Data and variable descriptions

The data used in analyses in this project consisted of n=53, 725 *Listeria monocytogenes* pathogens with 50 variables related to the pathogenic strains submitted, including information about who collected the isolate, its taxonomic name, its isolation source, date of collection (day, month, year), country or state from which the strain originated, among other metadata. Given the vast amount of information available in the database, these analyses employed an inclusion criteria to select strains for further analysis. In particular, for consideration and inclusion into the analysis sample, the isolate had to have a non-missing location, had been collected in the US, had a non-missing isolation source, and IFSAC category. The analysis sample based on this inclusion criteria included a total of n = 14, 810 *Listeria monocytogenes* pathogens. Figure 1 summarizes the isolate inclusion criteria for analysis in this project.

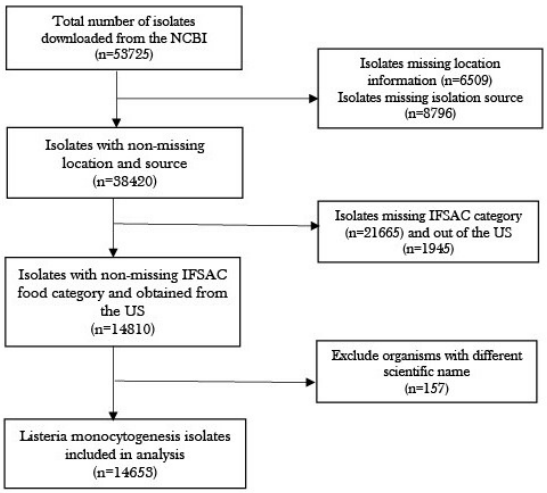

Figure 1: The flow of the data cleaning process based on the specified isolate inclusion criteria.

Data preprocessing

There were 42,794 unique strains collected across the US states represented in this data set, a reflection of the variation and heterogeneity of the data. Additionally, there were 1401 unique isolation sources which comprised of 60 clinical types and 14,474 environmental/other types. There were 296 different AMR genotypes, and 39 different outbreaks. The data had 285 unique isolate sourcing categories as developed by IFSAC category scheme, 90 unique hosts, and 67 unique host diseases. To make meaningful comparisons in relation to our objective, further work will involve aggregating the IFASC category into 7 broad categories. We also examined the collection date variable that was used to explore trends over time. Since the `Collection date` variable contained the date the sample were collected in the format the submitter supplied ranging from Month-Date-Year, Year-Month and Year only while the `Create date` was in the Year-Month-Date ISO format with time stamp the data was added into the Pathogen Detection Project, we first converted these into a standard form of Year-Month-Date. Then for `Collection date` variable with missing values, we chose to fill

in these dates by using those from the `Create date` variable. Finally we created `Year` and `Month` variables and extracted the respective years and months from the`Collection date` variable to maintain consistency in terms of available year records. For the Location variable, we reduced this to only include 50 states in the United States and 1 District (Puerto Rico). Table 1 summarizes selected variables from the data set.

Table 1: Variable descriptions

Variable	Description
Organism group	The name of the taxonomy group that the isolate belongs to and is represented by the Genus species name, for our case we shall focus on Listeria monocytogenes.
Isolate	The unique Pathogen Detection accession of the isolate where each accession has a prefix (PDT), which stands for Pathogen Detection Target.
IFSAC category	Categories of isolate sourcing information as developed by The Interagency Food Safety Analytics Collaboration (IFSAC).
Isolation Source	Provides information on the physical, environmental and/or local geographical source of the biological sample from which the sampled was derived.
Isolation Type	Contains categories of the isolation sources into either clinical or environmental/other groups.
Strain	Denotes the microbial strain name used to distinguish a genetically distinct lineage separated from another strain by one or two mutations.
Host	Refers to the host species of the isolate such as Animal, Homo sapiens, Sheep, Pigeon, Horse and Guinea pig.
Host Disease	Host disease matches the identified isolate to a disease origin, for example Listeriosis, gastroenteritis, Meningitis and Septicaemia.
Collection Date	Gives the date the sample was collected.
Create Date	Gives the date on which the isolates were first seen by the Pathogen Detection system.
Outbreak	Defines a way to group isolates that originated due to the same breakout among a specific group of people or within a specific area over a period of time.
BioSample	Describes the biological source materials used in experimental assay.
Lat/Lon	Provides the geographical coordinates (latitude and longitude) of the location where the sample was collected.
Location	Provides the geographical origin of the sample (Country or Region).
Min-Same	Represents the minimum single nucleotide polymorphism (SNP) distance to another isolate of the same isolation type for example, the minimum SNP distance from one clinical isolate to another clinical isolate.
Min Diff	Represents the minimum SNP distance to another isolate of a different isolation type. For example, the minimum SNP difference from a clinical isolate to an environmental isolate.
Serovar	Represents the combined field of sub-species, serotype, or serovar
AMR Genotypes	Provides information on the antimicrobial resistance (AMR) genes found in each isolate.
SNP Cluster	Represents single nucleotide polymorphisms (SNP) clusters, where the genome assemblies are closely linked to each other.

Missing Data

We shall start by assessing the missingness in our data set. Figure 2 shows us the overall missingness of our selected variables ordered from the least to the largest missing percentage. Key to note here is that there is over 86% missing observations in the variables. `Host Disease`, `AST Phenotypes`, `Computed types`, `Virulence genotypes`, `Source type` and `Outbreak` have approximately 100% of missing values. Variables such as `Isolation source`, `Isolation type`, `BioSample`, `Location`, `AMR genotypes` and `Isolates` have no missingness present. The missingness observed in this data set is likely due to under-reporting or non-response by the researchers or clinicians who submit this information on the NCBI website. Missing information in the collection date variable could be due to data entry errors when entering this information or just an oversight by the submitter. Generally, the missingness can be attributed to the high variations in the reporting practices or amount of time taken for lab processing prior to submission to the NCBI pathogen detection database. We acknowledge that this amount of missingness will be pose a major limitation in our study as these variables may not be informative in our analysis and may limit the interpretation and generalizability of our study findings.

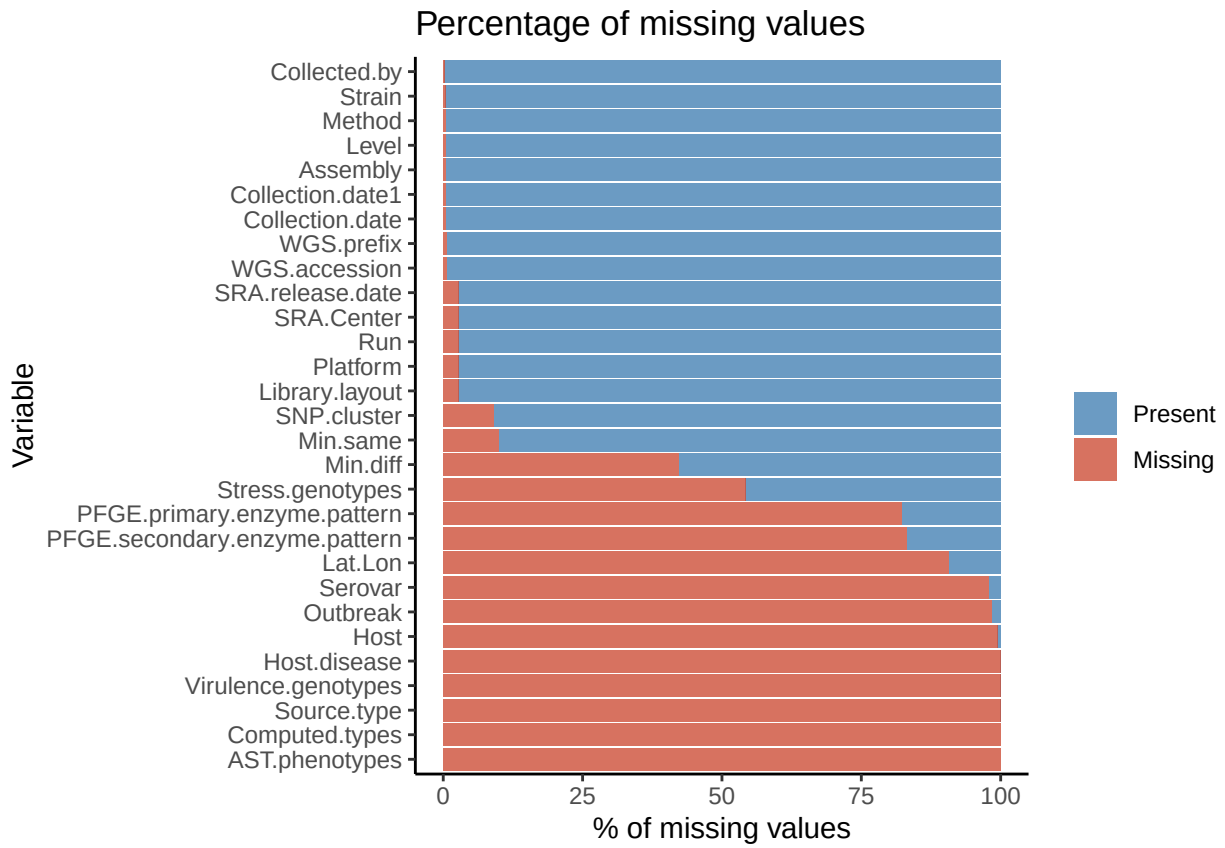

Figure 2: Missing values in variables

Exploratory data analysis

Overall trends in the counts of *L. monocytogenes* through time

We used descriptive statistics to first examine the proportion represented by our main variable of interest, the isolation source which originally had 1401 unique values which were as a result of punctuation, case sensitivity as well as many variations of the naming conventions of a general source. For example, 'cheese', 'white

cheese', 'ham cheese', and 'double cheeseburger'. For simplicity and for comparative purposes, we grouped the isolation sources into broader categories based on the patterns observed in this variable. Ultimately, the number of isolation sources was reduced to 38 broad categories including environmental, food, pork, chicken, beef,turkey, stool, water, other/unspecified. Environmental sources were highest at 54.34% followed by other/unspecified sources (9.65%). Water, dairy, and food sources represented 9.65%, 9.24% and 6.43%, respectively, while fish, beef, and pork represented 1.67%, 1.56%, and 1.47%, respectively.

We then used line plots to show an initial exploration of the trends in the number of *Listeria monocytogenes* over time. We filtered our data to work with a time frame from the year 2000 to 2022. The line plots in Figure 3 show a non-linear trend over time. There was a moderate increase in samples collected from the year 2000 to 2008, which sharply increased until about the year 2018. From 2018 to 2020, there was variation in terms of steady decrease/increase that was later followed by another sharp decrease in the samples collected. However, we also observed a slight increase in the counts following the year 2020. Grouped by isolation types, we observed a higher count in the environmental/other source types compared to the clinical type which remained relatively lower throughout the entire period of sample collection.

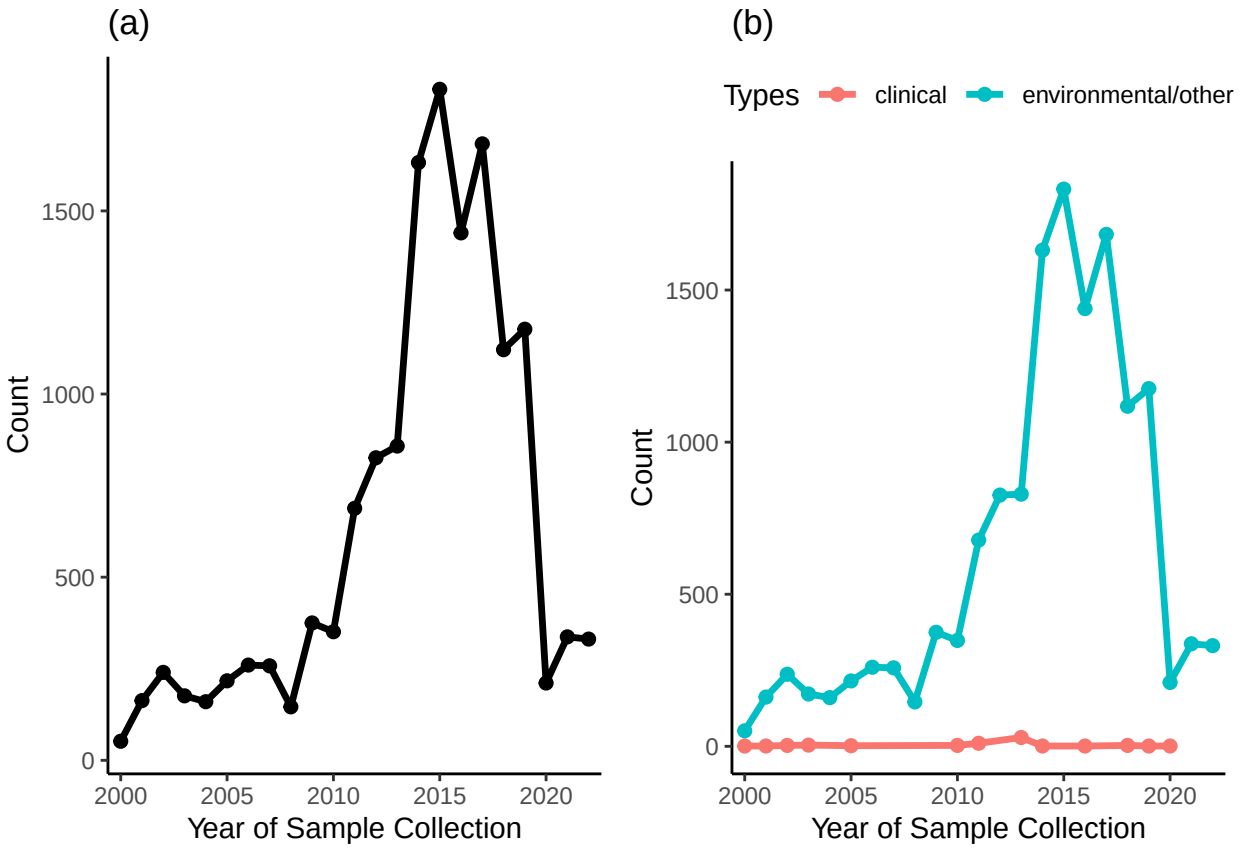

Figure 3: **(a)** Trends in the total counts of collected *L. monocytogene* pathogens. **(b)** Trends in the total counts of *L. monocytogenes* by isolation type.

Additionally, we explored summary frequencies of *Listeria monocytogenes* grouping by `Month` and `Isolation type`. Table 2 summarizes the counts of the *L. monocytogenes* where we observe that most cases of *Listeria monocytogenes* were observed in the early month of January; with 40% for clinical isolation type and 22.92% for environmental/other types. Additionally, for the clinical isolation types, frequent cases were observed in the months of September and October at 26.67% and 11.67% respectively. For the environmental/other isolation type, during the warmer months of April to August, we observed a moderate number of cases of *Listeria monocytogenes* ranging between 6.99% and 9.08%. There was missing clinical isolation types

Table 2: Counts of *Listeria monocytogenes* by month

Isolation	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
clinical	40%	3.33%	NA	NA	6.67%	3.33%	NA	5%	26.67%	11.67%	1.67%	1.67%
Environmental	22.92%	6.94%	7.16%	9.08%	7.28%	7.38%	7.52%	6.99%	6.71%	7.02%	5.31%	5.69%

cases observed during the months of March, April and July. Looking at the trends by State, California had the largest number of *Listeria monocytogenes* cases throughout our study time frame, $N = 2672$ (18.38%), followed by New York and Washington DC at 12.78% and 8.67% respectively. Nevada, West Virginia and Puerto Rico had the least number of cases each at 0.02%.

Trends in the counts of *L. monocytogenes* over time by the top isolation sources

The `Isolation source` was re-categorized into 38 broad categories. Figure 4 presents the trends over time in the counts of *Listeria monocytogenes* for the following sources: beef, chicken, dairy, pork, fish, food, potato, water. We observe that the most common isolate source in our data is dairy, water, followed by food and and pork.

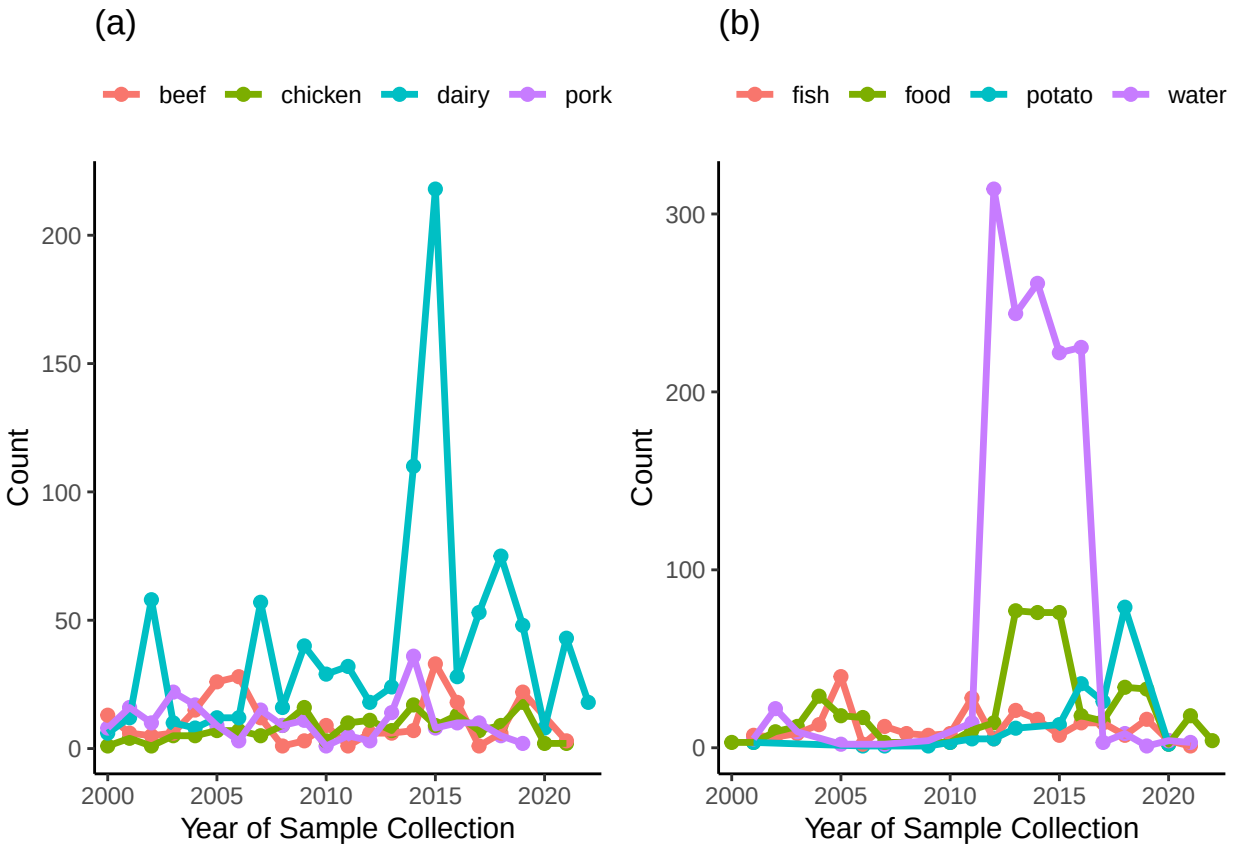

Figure 4: Line plots of top isolation surces for listeria monocytogenes counts over time. **(a)** Trends in the counts of pathogens from beef, chicken, dairy, and pork. **(b)** Trends in the pathogen counts from fish, food, potatoes, and water.

Serovar, AST phenotypes, AMR genotypes, and SNP Clusters

Our data had 14,534 unique isolates for *Listeria monocytogenes* with 14517 distinct `Biosamples` and no missing data on this information. Additionally, we looked at the distribution of `Serovar` and noted a relatively high percentage of missing data (n = 14275; 98.22%). `Serovar` information was entered using free text as there are many variations of names that could be representing similar information such as 1, 1a, 1/2a. On the other hand, the `AST phenotypes` variable, denoting the Antimicrobial Susceptibility Test was recorded in a raw string form. This variable represents the antibiotics that each isolate is either susceptible or resistant to. The `AMR genotypes` variable represents the Antimicrobial resistance (AMR) genes found in the isolate during analysis. We found 184 unique AMR genes in our data with no missing information. There were 1, 474 SNP clusters whose genome assemblies were closely related.

Distributions of `Min Same` and `Min Difference` variables.

Next we examined the distributions of `Min Same` and `Min Diff` variables. `Min-diff` is the minimum SNP distance to another isolate of a different isolation type (from an environmental isolate to a clinical isolate). Figure 5 shows that `Min Diff` approximately follows a bi-modal distribution suggesting that a transformation of this variable may be useful prior to using it in further analysis. On the other hand, `Min-same` was the minimum SNP distance to another isolate of the same isolation type (clinical to clinical or environmental to environmental). Additionally, Figure 5 shows that `Min Same` approximately follows an exponential distribution. A log transformation of this variable did not suggest a substantial deviation from the exponential distribution.

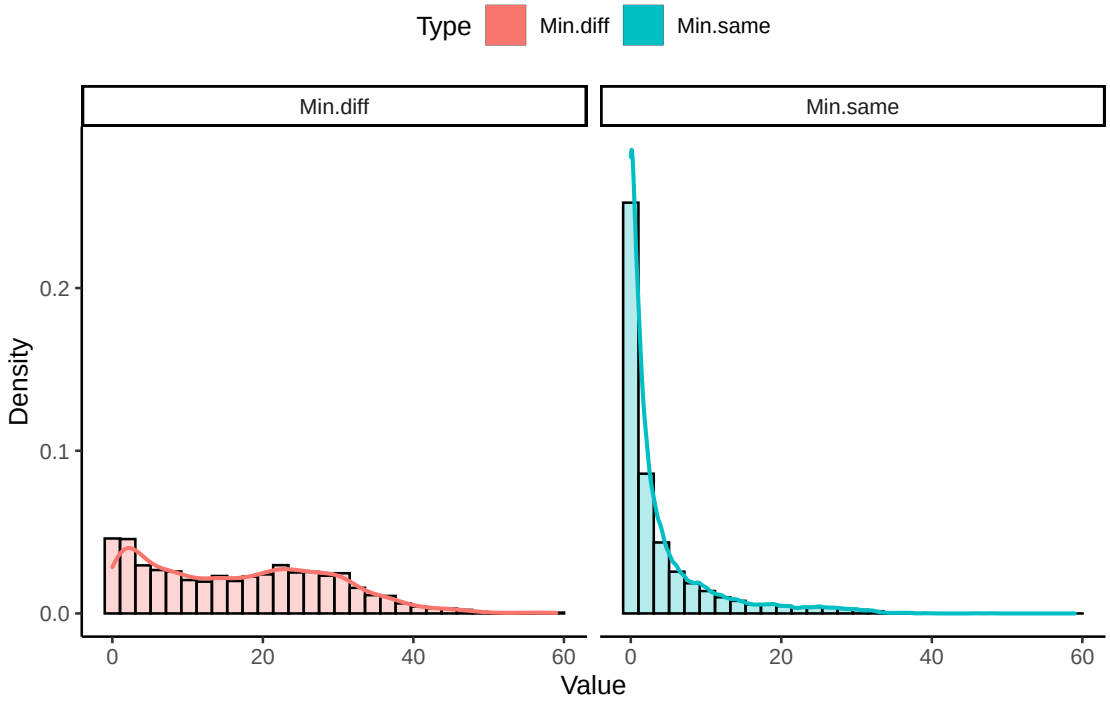

Figure 5: The distributions of the minimum SNP distance to another isolate of a different isolation type and the minimum SNP distance to another isolate of a similar isolation type.

Statistical modeling

This research project seeks to expand the literature on foodborne disease source attribution for human listeriosis using Bayesian and ensemble-based machine learning methods and core genome multilocus sequencing typing data, in addition to other selected information about the sampled *Listeria monocytogene* isolates in the US. The study seeks to evaluate the common food categories and their contribution to the overall foodborne illness disease burden using pathogenic isolates. The study seeks to explore the question: Given a human *Listeria monocytogene* isolate how likely is it to be from a particular food source? This study is informed by the need to leverage emerging technologies to identify strategies to enhance food safety, and the food production process and ultimately reduce the burden of foodborne illnesses. The paper seeks to compare the performance of Bayesian additive regression trees and random forest models to develop a foodborne-illness source attribution system; methods which have been shown to yield robust results in previous studies.

Random forests

Random forests are an ensemble of multiple decision trees which employ bagging to reduce over-fitting and thus enhance the generalizability of the resulting model. Multiple trees are fitted on the training data using randomly selected predictors to ensure that the resulting trees are uncorrelated. In the training of each tree, some samples are not used in the training process and are used in the estimation of the model performance by averaging. Each of the trees building up the random forest makes its prediction/classification to results in multiple predictions which are then averaged to inform the final result. For an in-depth analysis of the method, see Breiman [30].

Bayesian additive regression trees (BART)

BART is an ensemble machine learning methodology that allows flexible modeling of interactions in non-linear regressions and classification problems in the Bayesian context. The method involves sums of classification or regression trees and is able to flexibly capture non-linearities. The estimation of the unknown function, f, involves recursive partitioning of the variable space using a fully Bayesian probability modeling approach [31]. Every tree in this process is a weak learner that can only explain a portion of the variability in the outcome. The trees form decision trees which are highly interpretable and flexible but are prone to overfitting. Overfitting, in BART, is controlled using a regularization prior that forces every tree in the model to explain only a small amount of the relationships between the outcome and the predictor variables. The model can be summarized as:

$$\begin{aligned} E[Y|X] &= f(\mathbf{X}) + \epsilon \\ &= T_1^M(\mathbf{X}) + \cdots + T_m^M(\mathbf{X}) + \epsilon \,\text{where}\, \epsilon = N(0, \sigma^2\mathbf{I}_n) \end{aligned} \tag{1}$$

where Y denotes the vector of outcomes, $\mathbf{X}$ denotes the $n \times p$ design matrix where p denotes the number of predictors in the model and ϵ denotes some noise. m is the number of unique trees, T denotes tree structure, and M denotes the terminal node or leaf parameters. BART models consist of three prior components including tree structure, leaf parameter, and error variance. Chipman et al. [31] presents an in-depth discussion of this methodology.

Potential data limitations

Even though the NCBI pathogen detection allows real-time identification of clusters of related genomic sequences to aid in outbreak detection, and track the spread of resistance genes, a potential limitation of this data source is that it does not identify outbreaks or outbreak memberships and analyses rely solely on publicly available data submitted to the database. Additionally, the database allows a lot of flexibility in the naming conventions which results in substantial heterogeneity that make it difficult to query and extract meaningful patterns for microbial risk assessments [32]. For instance, the `collected by` and `isolation source` are fields that were entered as free text which are very extreme in the options they present for analysis. Moreover, there is a lot of missing data on potentially useful fields, a scenario that makes it difficult to derive inferences that could inform food policies.

Conclusion

In this part of the paper, we have presented literature related to the use of statistical machine learning methods in foodborne illness source attribution. Given the high burden associated with human listeriosis and the particular severity posed on people with compromised immune systems, we focus on developing a system for linking *Listeria monocytogenes* pathogens to particular food sources following our categorization informed by the IFSAC categorization. Exploratory data analysis revealed substantial missing data as well as overall heterogeneity in the representation of the NCBI data particularly with regards to naming conventions. The lack of a standard criteria in the names of the levels of the variables is limiting in terms of querying and extracting useful patterns for risk assessment. Given our interest in listeria foodborne illness source attribution, our categorizations reveal beef, chicken, dairy, pork, fish, water, and potatoes as potential food sources of interest in our modeling given their high presence in the analyzed *Listeria monocytogene* pathogens analyzed herein. In the next steps, we examine these data more closely and employ statistical methods to extract patterns as well as substantial predictors of a food source associated with a given isolate and consider exploring the particular advantage presented by cgMLST and machine learning. We look to build on the resilience of foodborne pathogens, their ability to adapt both genetically and phenotypically to changing environments and the particular interest in using these knowledge to understand pathogen responses to food handling practices and ultimately reduce foodborne disease incidence.

Data and code availability

The data and other analysis files associated with this project are available online on GitHub at https://github.com/okutse/foodborne_diseases.

References

[1] Organization WH. Estimates of the global burden of foodborne diseases, https://www.who.int/multimedia/details/estimates-of-the-global-burden-of-foodborne-diseases (2015).

[2] Lee H, Yoon Y. Etiological agents implicated in foodborne illness world wide. *Food Science of Animal Resources*; 41. Epub ahead of print 2021. DOI: 10.5851/KOSFA.2020.E75.

[3] Tanui CK, Karanth S, Njage PM, et al. Machine learning-based predictive modeling to identify genotypic traits associated with salmonella enterica disease endpoints in isolates from ground chicken. *LWT* 2022; 154: 112701.

[4] Scallan E, Hoekstra RM, Angulo FJ, et al. Foodborne illness acquired in the united states-major pathogens. *Emerging Infectious Diseases*; 17. Epub ahead of print 2011. DOI: 10.3201/eid1701.P11101.

[5] Gourama H. Foodborne pathogens. In: *Food safety engineering*. Springer, 2020, pp. 25–49.

[6] Chlebicz A, Śliżewska K. Campylobacteriosis, salmonellosis, yersiniosis, and listeriosis as zoonotic foodborne diseases: A review. *International journal of environmental research and public health* 2018; 15: 863.

[7] Filipello V, Mughini-Gras L, Gallina S, et al. Attribution of listeria monocytogenes human infections to food and animal sources in northern italy. *Food microbiology* 2020; 89: 103433.

[8] Lomonaco S, Nucera D, Filipello V. The evolution and epidemiology of listeria monocytogenes in europe and the united states. *Infection, Genetics and Evolution* 2015; 35: 172–183.

[9] Disease Control C for, Prevention. Listeria (listeriosis) | listeria | CDC, https://www.cdc.gov/listeria/index.html (2022).

[10] Matle I, Mbatha KR, Madoroba E. A review of listeria monocytogenes from meat and meat products: Epidemiology, virulence factors, antimicrobial resistance and diagnosis. *Onderstepoort Journal of Veterinary Research* 2020; 87: 1–20.

[11] Hilliard A, Leong D, O'Callaghan A, et al. Genomic characterization of listeria monocytogenes isolates associated with clinical listeriosis and the food production environment in ireland. *Genes* 2018; 9: 171.

[12] Schwartz B, Hexter D, Broome CV, et al. Investigation of an outbreak of listeriosis: New hypotheses for the etiology of epidemic listeria monocytogenes infections. *Journal of Infectious Diseases* 1989; 159: 680–685.

[13] Chen J-Q, Regan P, Laksanalamai P, et al. Prevalence and methodologies for detection, characterization and subtyping of listeria monocytogenes and l. Ivanovii in foods and environmental sources. *Food Science and Human Wellness* 2017; 6: 97–120.

[14] Jensen AK, Björkman JT, Ethelberg S, et al. Molecular typing and epidemiology of human listeriosis cases, denmark, 2002–2012. *Emerging infectious diseases* 2016; 22: 625.

[15] Gaul LK, Farag NH, Shim T, et al. Hospital-acquired listeriosis outbreak caused by contaminated diced celery—texas, 2010. *Clinical infectious diseases* 2013; 56: 20–26.

[16] Pouillot R, Klontz KC, Chen Y, et al. Infectious dose of listeria monocytogenes in outbreak linked to ice cream, united states, 2015. *Emerging Infectious Diseases* 2016; 22: 2113.

[17] Noordhout CM de, Devleesschauwer B, Angulo FJ, et al. The global burden of listeriosis: A systematic review and meta-analysis. *The Lancet Infectious Diseases* 2014; 14: 1073–1082.

[18] Munck N, Njage PMK, Leekitcharoenphon P, et al. Application of whole-genome sequences and machine learning in source attribution of salmonella typhimurium. *Risk Analysis*; 40. Epub ahead of print 2020. DOI: 10.1111/risa.13510.

[19] Njage PMK, Henri C, Leekitcharoenphon P, et al. Machine learning methods as a tool for predicting risk of illness applying next-generation sequencing data. *Risk Analysis*; 39. Epub ahead of print 2019. DOI: 10.1111/risa.13239.

[20] Lupolova N, Dallman TJ, Holden NJ, et al. Patchy promiscuity: Machine learning applied to predict the host specificity of salmonella enterica and escherichia coli. *Microbial genomics*; 3.

[21] Karanth S, Tanui CK, Meng J, et al. Exploring the predictive capability of advanced machine learning in identifying severe disease phenotype in salmonella enterica. *Food Research International* 2022; 151: 110817.

[22] Liu Y-Y, Chen C-C. A machine learning-based typing scheme refinement for listeria monocytogenes core genome multilocus sequence typing with high discriminatory power for common source outbreak tracking. *PloS one* 2021; 16: e0260293.

[23] Vangay P, Steingrimsson J, Wiedmann M, et al. Classification of listeria monocytogenes persistence in retail delicatessen environments using expert elicitation and machine learning. *Risk analysis* 2014; 34: 1830–1845.

[24] Sun W, Liu Y, Wang X, et al. Quantitative risk assessment of listeria monocytogenes in bulk cooked meat from production to consumption in china: A bayesian approach. *Journal of the Science of Food and Agriculture* 2019; 99: 2931–2938.

[25] Pasonen P, Ranta J, Tapanainen H, et al. Listeria monocytogenes risk assessment on cold smoked and salt-cured fishery products in finland-a repeated exposure model. *International journal of food microbiology* 2019; 304: 97–105.

[26] Mughini-Gras L, Benincà E, McDonald SA, et al. A statistical modelling approach for source attribution meta-analysis of sporadic infection with foodborne pathogens. *Zoonoses and Public Health*.

[27] Lassen SG, Ethelberg S, Björkman J, et al. Two listeria outbreaks caused by smoked fish consumption—using whole-genome sequencing for outbreak investigations. *Clinical Microbiology and Infection* 2016; 22: 620–624.

[28] Pires SM, Vieira AR, Hald T, et al. Source attribution of human salmonellosis: An overview of methods and estimates. *Foodborne pathogens and disease* 2014; 11: 667–676.

[29] National Library of Medicine (US) National Center for Biotechnology Information B (MD): The NCBI pathogen detection project [internet], https://www.ncbi.nlm.nih.gov/pathogens/ (2016).

[30] Breiman L. Random forests. *Machine learning* 2001; 45: 5–32.

[31] Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 2010; 4: 266–298.

[32] Sanaa M, Pouillot R, Vega FG, et al. GenomeGraphR: A user-friendly open-source web application for foodborne pathogen whole genome sequencing data integration, analysis, and visualization. *PloS one* 2019; 14: e0213039.