School of Public Health
BROWN UNIVERSITY

# Foodborne pathogen detection of listeria isolation source and seasonal/regional patterns using machine learning methods

Tianye Cui, Yifan Zhao, Yingjie Zhou
Department of Biostatistics

## Overview

**We predicted sources of listeria cases from NCBI Pathogen Detection genetic, location, and time data using K-nearest neighbors, multinomial logistic regression, and neural networks. Then, we evaluated the accuracy and interpretability of the models.**

## Background

- Listeria monocytogenes is the third most costly foodborne pathogen (per case) in the United States, with 1600 infections annually.
- A sample of a pathogen (isolate) has genetic makeup and other properties that can be analyzed and grouped for commonality
- NCBI Pathogen Detection database provides SNP cluster, AMR genotype, geographical, and time data for each isolate sample.
- Machine learning tools can help public health researchers narrow down the investigation of outbreaks by identifying isolation source.

## Methods

- Data Source: NCBI Pathogen Detection Database
- Outcome variable, Isolation Source, is grouped into 5 categories using the Level 1 Hierarchy of CDC's IFSAC Food Categorization Scheme.

| Aquatic animal | Clinical human | Environ -ment | Land animal | Plants |
|---|---|---|---|---|
| 228 (1.7%) | 1583 (11.9%) | 8363 (62.8%) | 1978 (14.9%) | 1160 (8.71%) |

- Inverse probability weighting was used to balance class.
- Predictors: SNP cluster, AMR genotype, Region or State, Season or Month and Year, Min SNP Diff.
- Different classification models used to see if results were consistent
- Methods:
  - Multinomial: Best model was selected using backward selection based on AIC and contained all main effects of predictors above with significant interactions of region:season and region:year. Model fit was checked using LRT and prediction on test set.
  - K-nearest neighbors (KNN): A 10-fold cross-validation was used in the KNN classifier to determine the isolation source by the voting for the nearest neighbors.
  - Neural Network - 3 Layers, Softmax Activation Function, Categorical Cross-Entropy Loss Function, Adam Optimizer, L2 Regularization, Dropout Layer (0.2), Validation Split (0.2)

## Multinomial Logistic

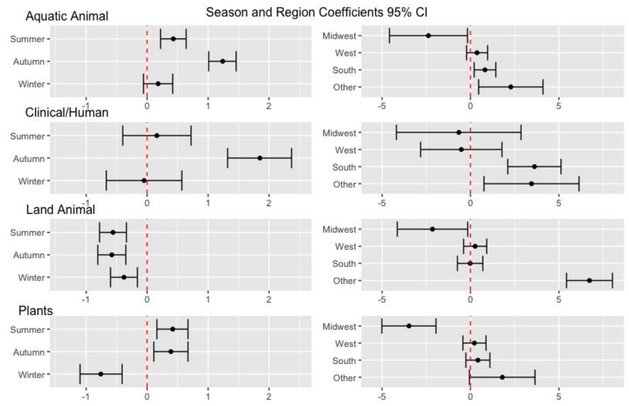


**Figure 1.** The 95% confidence interval (CI) illustrates different effects of season and region on risk of Listeria sources. CI's that don't contain zero (red line) indicate significant effects. Specifically, Aquatic Animal and Plant source have higher risk to occur in summer and autumn, whereas Land Animal source is only risky in Spring. Comparing with Northeast, South has higher risk of Listeria sourcing from Aquatic Animal and Human, and Midwest has lower risk sourcing from Land Animal and Plants. (*Reference level: Spring, Northeast, Environment)*

## K-nearest neighbors

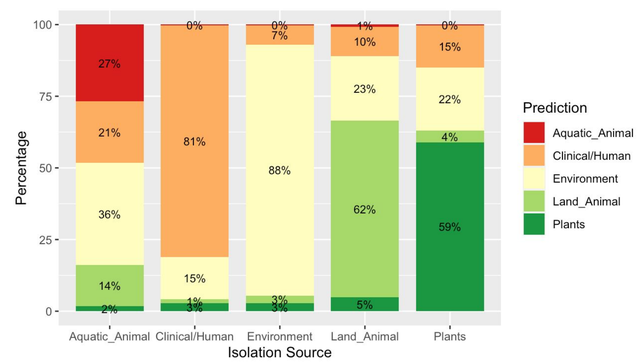


**Figure 2.** Actual versus predicted isolation source based on the KNN model. The aquatic animal group, which had the lowest predictive accuracy of any of the isolation sources at 27%, only made up roughly 2% of all the cases combined. All other sources have acceptable predictive accuracy. The environment category especially, making up 65% of the dataset, had the highest level of accuracy with a success rate of 88%. The overall maximum accuracy was achieved at 80.4% when the number of nearest neighbors k=5.

## Neural Network Results

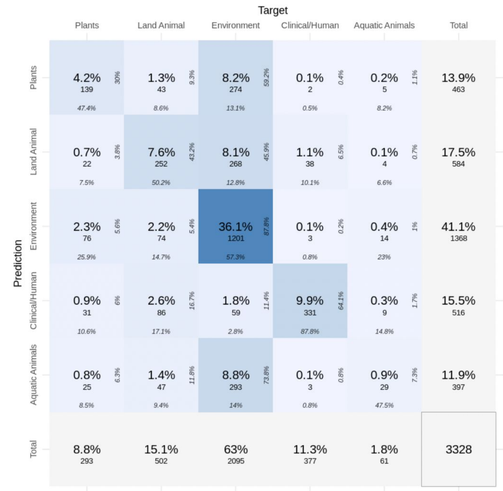


**Figure 3.** Confusion Matrix. Accuracy: 58.65%; 95% CI: [56.95%, 60.33%]. There is poor prediction accuracy for minority classes even when weighted.

## Limitations

- Consistent underestimation/erroneous predictions of minority classes
- Predictions more difficult with higher number of levels
- Conclusion and prediction accuracy were sensitive to isolation source variable grouping criteria and data cleaning/preprocessing steps. Open-entry responses make grouping arbitrary.
- Choice of reference level of categorical variables dependent on public health research input

## Conclusion

- Broad seasonal and regional patterns among Listeria sources identified by multinomial logistic regression. Utility of these patterns need improvement. For example, higher risk of listeria from aquatic animals and plants in the summer and autumn (relative to spring) reinforces literature but isn't specific enough to inform policy change.
- Using KNN to cluster cases by characteristic similarities can help narrow down pathogen source for unlabelled samples.
- Prediction of isolation source by neural net may not viable/useful.
- Future Works:
  - Repeat for other levels of ISFAC Food Categorization Scheme
  - Compare Random Forest predictions with those from NN. Draw insights from Variable Importance Plots.
  - Include other variables (e.g., host species) into the models