Manuscript for Foodborne-Outbreak Project

Weifan Wu & Kailin (Kai) Chen

5/1/23

here() starts at C:/Users/Kai/Documents/School/Colleges/UGA/MPH Year/Spring 2023/Modern Data Analysis/Weifan-Kailin-MADA-project

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.1 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.1 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.1   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

# Summary/Abstract

# Introduction

## General Background Information

Foodborne pathogens typically cause serious gastrointestinal infections. On average, 1 in 6 Americans will acquire foodborne illness each year; 128,000 will be hospitalized, and 3,000 will die (Food Safety & Applied Nutrition, 2022). Economic loss associated with foodborne illness is estimated to be greater than 50 billion US dollars annually. Restaurants are frequently sites attributed to mass-outbreaks, and typical foods associated with these outbreaks include improperly-prepared meats and vegetables (Mun, 2020; Sarno, Pezzutto, Rossi, Liebana, & Rizzi, 2021). Disease etiologies range from bacteria (ex: *Salmonella* spp. and *E. coli*) to viruses (norovirus and rotavirus), to parasites (*Cryptosporidium* spp. and *Cyclospora* spp. (CDC, 2022). To verify pathogens of greatest concern, common settings associated with foodborne illness, and foods typically associated with such outbreaks, this study was launched as a research project at the University of Georgia using information gathered over a 17-year period from the Centers for Disease Control and Prevention (CDC).

### Questions & Hypotheses

From this research project, we would like to answer the following questions: - What are the most frequent locations that foodborne outbreaks occur? - What pathogen species are most frequently associated with foodborne outbreaks? - What type of food is most often seen for foodborne outbreaks?

In other words, immediate outcomes of interest are location, disease etiology, and food category. These will act as predictors for the health outcomes of interest: illnesses and hospitalizations.

Our hypotheses include: - Outbreaks occur most often in homes rather than restaurants. - The most common pathogen associated with foodborne disease outbreaks is *E. coli*. - The food most commonly associated with foodborne outbreaks consists of salad.

# Methods

## Data Sources, Acquisition, and Description

Two datasets were adopted for this research. The first dataset–which was used to build statistical models for the study–was uploaded onto *Kaggle* in 2017, with 12 variables and more than 16,950 observations (CDC, 2017). The second dataset–which was used for machine learning–was retrieved from the CDC’s National Outbreak Reporting System (NORS) website. This second dataset covers reports of foodborne and waterborne disease outbreaks, as well as enteric disease outbreaks spread by contact with environmental sources, infected animals or persons, and other means from 2009 to 2020.

## Data Importation and Cleaning

The datasets were downloaded from their respective sites and imported into R/RStudio using the tidyverse and readxl packages. Afterwards, the data structure for each file was analyzed, including how many “NA” values were present for each column in the dataset. Only observations pertaining to foodborne disease were kept. Variables containing less than 80% “NA” values, including predictors and outcomes of interest, were retained. For the designated “training data”, cases that did not have laboratory-confirmed tests were filtered out. Variables for pathogen type (virus, bacteria, parasite, and poison) and Simplified Etiology were created based on pathogen species. For both datasets, Location was formatted identically to maintain compatibility for eventual machine learning. IFSAC Category was also created in the training data based on the formatting of the “test data” and modified to include salads, desserts, and poultry. To address coinfections, observations containing multiple pathogen etiologies were separated into different rows. Finally, both datasets were modified to contain identical columns prior to being saved as “processed data” in the appropriate subdirectory.

For more information, please see the “processingfile\_v1.qmd” file located in the “processing\_code” folder of the “code” subdirectory.

## Statistical Analysis

### Exploratory

After processing procedures, an exploratory data analysis was performed to view general relationships between predictors (time, pathogen type, pathogen etiology, state, and food type) and the health outcomes of interest (illnesses and hospitalizations).

### Machine Learning

Linear and Poisson regression models were afterwards constructed using the aforementioned predictors and outcomes from the training data. Data was resampled using the bootstrapping approach 1,000 times for better estimates of accuracy and variability of model coefficients. Random forest and bagging models were also trained to predict health outcomes using the Kaggle dataset for the NORS-CDC data. RMSE was used to compare each of the models. The best model built from the training data was then employed for the test data.

For more information, please refer to the “statistical\_analysis.qmd” file inside the “analysis\_code” folder of the “code” subdirectory.

# Results

## Exploratory Analysis

### Training Data

|  |
| --- |
| Figure 2: Hospitalizations by Food Type. |

### Test Data

|  |
| --- |
| Figure 1. Year vs Illnesses Caused by Pathogen Type |

|  |
| --- |
| Figure 3. Illnesses by outbreak location. |

|  |
| --- |
| Figure 4. Bacterial etiology by outbreak frequency. |

## Statistical Analysis

### Basic

Between simple and multivariate regressions, multivariate models performed better with lower associated RMSEs. For both health outcomes, the multivariate linear regression models (RMSE = 58.4, 5.50) constructed performed better than multivariate poisson regression models (RMSE = 59.7, 5.64).

|  |
| --- |
| Figure 5. Predicted illnesses versus actual illness count for multivariate linear regression model. |

### Machine Learning

Overall, between itself and the other models (linear regression, poisson regression, and bagging) generated, the random forest model performed the best for both health outcomes with the lowest RMSEs ()

### Final Fit with Test Data

Fitting the finalized random forest model to the test data resulted in RMSEs of 36.6 and 6.70 for the health outcomes of illnesses and hospitalizations, respectively.

# Discussion

## Summary and Interpretations

## Strengths and Limitations

# Conclusion

# References

CDC. (2017). *Foodborne Disease Outbreaks, 1998-2015*. Retrieved from <https://www.kaggle.com/datasets/cdc/foodborne-diseases>

CDC. (2022). *Foodborne illnesses and germs*. Retrieved from <https://www.cdc.gov/foodsafety/foodborne-germs.html>

Food Safety, C. for, & Applied Nutrition. (2022). *What You Need to Know about Foodborne Illnesses*. Retrieved from <https://www.fda.gov/food/consumers/what-you-need-know-about-foodborne-illnesses>

Mun, S. G. (2020). The effects of ambient temperature changes on foodborne illness outbreaks associated with the restaurant industry. *International Journal of Hospitality Management*, *85*, 102432. <https://doi.org/10.1016/j.ijhm.2019.102432>

Sarno, E., Pezzutto, D., Rossi, M., Liebana, E., & Rizzi, V. (2021). A Review of Significant European Foodborne Outbreaks in the Last Decade. *Journal of Food Protection*, *84*(12), 2059–2070. <https://doi.org/10.4315/JFP-21-096>