**Title** Using Outbreak Data for Hypothesis Generation—A Source Prediction Tool

**Authors**

<https://coe-foodsafetytools.shinyapps.io/sourceattribution/>

**ABSTRACT**

**INTRODUCTION**

Hypothesis generation about potential sources is a critical step in an enteric disease outbreak investigation. A thorough hypothesis narrows the scope of an investigation, making more efficient use of scarce resources and increasing the likelihood of successfully implicating a source. When generating a hypothesis, public health investigators use historical information on the pathogen and pathogen sub-types, including common vehicles implicated in past outbreaks. For example, 65% of *Salmonella* Enteritidis outbreaks are associated with eggs 1. Investigators also use descriptive data of cases to suggest a food source. For example, case demographic data, including age, gender, and ethnicity, provide clues that suggest or point away from a particular food. The geographical spread and timing of an outbreak provides evidence about the distribution and type of exposures.

Previously, we developed a model for Shiga toxin-producing *Escherichia coli* (STEC) to test the validity of using data from past outbreak to support hypothesis generation. The work endorsed using prior case and outbreak characteristics to predict food sources in STEC outbreaks (White et al. 2016). However, we noted a number of limitations. Our method required complete data on all predictors for all outbreaks included in the model, which resulted in the exclusion of many outbreaks. Because of this, the model did not include age as a predictor despite being an important predictor of outbreak sources. Finally, our analysis was restricted to STEC outbreaks, whereas non-typhoidal *Salmonella* is the leading bacterial causes of foodborne outbreaks, causing 23% of single-etiology outbreaks. In this study, we aimed to develop a source prediction tool for STEC and *Salmonella* outbreaks using alternative methods that allowed the use of incomplete data in order to improve the predictive ability. In addition, we aimed to translate the final statistical model into a user-friendly online tool for investigators.

**METHODS**

**Data Source**

Outbreak data were available from the Centers for Disease Control and Prevention’s (CDC) Foodborne Outbreak Surveillance System from 1998 to 2017. This passive surveillance system receives outbreak reports from state, local, and territorial health agencies using a standard form 3. From 1998 to 2008, data on foodborne outbreaks was collected using the Electronic Foodborne Outbreak Reporting System (eFORS). In 2009, the National Outbreak Reporting System (NORS) replaced eFORS and expanded to collect data on foodborne, waterborne, person-to-person, animal contact, environmental contamination, and undetermined transmission routes.

Each outbreak report includes information on the date and location of the outbreak, investigation methods, case demographics (e.g., the percentage of cases by sex and age group), etiology, transmission route (e.g., foodborne, animal contact), setting, and the implicated food, if applicable. CDC categorizes food commodities using the Interagency Food Safety Analytics Collaboration (IFSAC) Food Categorization Scheme (IFSAC 2013; Painter et al., 2009). The IFSAC food scheme is based on a taxonomic scheme of 17 mutually exclusive commodities. Only implicated foods composed of ingredients from a single commodity (i.e., ‘‘simple’’ foods) were categorized using this scheme. Foods with ingredients from multiple commodities were labeled as ‘‘complex’’ foods. For example, beef is a simple food, and a hamburger is a complex food4. Only foods categorized as a single commodity were used for this analysis.

**Data analysis**

We included confirmed and suspected STEC and *Salmonella* outbreaks transmitted via food or animal contact. We excluded outbreaks with multiple etiologies and outbreaks caused by multiple or complex food vehicles, outbreaks with unknown or missing food vehicles, or food vehicles that could not be classified using the IFSAC scheme. We included only foodborne or animal contact outbreaks because we assumed the risk factors and potential predictors for other modes of transmission (e.g., person-to-person) would differ substantially. We determined outbreak source categories based on number of outbreaks in that category. Outbreak data from 1998-2016 were randomly split into a training set (75%) and a testing set (25%) stratified by outbreak source to ensure balance. Outbreaks with rare food vehicles (fewer than 100 outbreaks) were excluded from the training set, but they were included in the testing set. Outbreak data from 2017 were used to evaluate the final selected model.

**Predictors**

Case demographic predictors included percentage of female cases (the number of female cases as a proportion of cases whose gender was known), and percentage of cases in each age group (<1 year, 1-4, 5-19, 20-49, ≥50). Outbreak predictors included etiology (STEC or *Salmonella* serotype), number of cases (confirmed or suspected), month of first illness onset, and geographical distribution (multi-state, multi-county in a single state, or single county in a single state). Season was based on the onset of the first case and categorized as winter (January to March), spring (April to June), summer (July to September), and fall (October to December). Missing predictors were imputed using k-nearest neighbors using the training dataset, then applied to testing dataset and validation dataset.

For *Salmonella* serotypes with fewer than ten and more than three outbreaks, we used logistic regression to cluster into two groups, plant-associated or animal-associated serotypes. Serotypes with three or fewer outbreaks were categorized into a single group, rare serotypes. Missing serotypes were treated as missing.

**Model Selection**

We selected six algorithmic methods for prediction based on their ability to predict multiple class probabilities. These methods included adaptive boosting classification trees (AdaBoost.M1), classification and regression trees (CART), weighted k nearest neighbors (knn), boosted trees (using xgboost), random forest (using ranger) and multivariate adaptive regression splines (MARS). A non-informative model with no predictor information was generated for comparison purposes. We used cross-validation in the training set to tune model parameters for each algorithmic method. We then compared Brier Scores (a measure of the difference in the predicted probability and the actual event) for each method in the testing dataset to select a final model. Brier score calculations included outbreaks with rare (<100 outbreaks) sources. We used 2017 outbreaks as an external validation step to evaluate the real-world model performance.

All analyses were performed using R version 3.6.1 (2019-07-05). Parameter selection was performed using the Caret package. Data cleaning was done using the tidyverse package. Rsample v(0.0.5) and recipes v(0.1.7) (part of tidymodels) were used for data splitting, imputation and preprocessing. The parsnip v(0.0.3.1) (null model), adabag v(4.2) (Adaboost.M1), C50 v(0.1.2) (CART), kknn v(1.3.1) (weighted knn), xgboost v(0.90.0.2) (boosted trees), and ranger v(0.11.2) (random forest) packages were used. The caret Package v(6.0.84) was used for tuning and test set prediction. We developed a web application tool using the R Shiny package.

**RESULTS**

From 1998 to 2016, 4,059 (X STEC and Y *Salmonella*)foodborne and animal contact outbreaks were reported to CDC. Of these, X outbreaks were excluded from model development for the following reasons: missing IFSAC source information (); outbreak caused by multiple sources (), unclassifiable sources (), or undetermined sources (). The analysis dataset included 305 STEC outbreaks and 1,070 *Salmonella* outbreaks. The training set excluded 12 foodborne outbreaks that could not be classified as a plant or animal source and 203 outbreaks with rare food sources (79 dairy, 9 game, 31 other land animal, 10 grains-beans, 18 nuts-seeds, 1 oils-sugars, 19 other plant and 35 aquatic animal). Of the 305 STEC outbreaks included in model development, 23 (8%) were attributed to eggs, 121 (40%) to meat-poultry, 64 (21%) to produce, and 48 (16%) to animal contact. Of the 1,070 *Salmonella* outbreaks included in model development, 132 (12%) were attributed to eggs, 415 (39%) to meat-poultry, 218 (20%) to produce, and 139 (13%) to animal contact (Table 1). In 2017, X outbreaks were reported.

We observed differences in outbreak characteristics and demographics by outbreak source (Table 1). There were more animal contact outbreaks in winter months (35%), and more egg and meat-poultry outbreaks in summer months (38% and 35%, respectively). The proportion of multistate outbreaks was highest for produce (46%), followed by animal contact (38%). The mean percentage female was highest for produce outbreaks (55%) and lowest for egg (40%) and meat-poultry (43%) outbreaks. Animal contact was on average the most frequent source for age groups <1 year, 1-4 years, 5-9 years, and 10-19 years, with the highest in ages 1-4 (23%). Those aged 20-49 years were highest for produce (20%). Age was unknown or missing for most outbreaks. The percentage of cases hospitalized was similar across outbreak sources. Produce outbreaks tended to be higher, with a median of 19 cases per outbreak.

There were 341 outbreaks in the training data set (11% egg, 39% meat-poultry, 21% produce, 13% animal contact, and 16% ‘other’) and 989 outbreaks in the testing data set (12% egg, 41% meat-poultry, 21% produce, 14% animal contact, and 12% ‘other’) (Table 3).

Final selected tuning parameters are shown in table 2. All models performed better than the null model on the test set (table 3). Calibration curves are shown in figure 3.

Model performance varied. Two models (Naive Bayes and rule-based classifier) had a Brier score worse than the non-informative model. Weighted k-nearest neighbors (kNN) and weighted subspace random forest performed the best with Brier Scores of 0.125 and FDA of 0.127, respectively. Calibration curves based on the testing data set are shown in Figure 1. Weighted kNN was selected for the final model.

Using the kNN model, a source with a predicted probability from 0 to 20% was correct 5% of the time and a source with a predicted probability from 80 to 100% probabilities was correct 57% of the time. For example, if the model predicted an outbreak to be produce with a predicted probability of 60-80%, the outbreak was truly produce 100% of the time (Table 4).

*Validation*  
The validation data set consisted of 98 food-borne and animal contact outbreaks including 7 outbreaks whose sources was not one of our predicted categories. These were not excluded from evaluation metrics to more accurately reflect real world performance. The selected random forest model had a brier score of 0.098. The calibration plot is shown in figure 4. Fruit and animal contact outbreaks were the most under-predicted. Dairy, meat and egg outbreaks were the most over-predicted.

Overall, the correct outbreak was in the top two predicted for 64 (65%) of outbreaks. The percentage of outbreaks where the correct category was in the top two predictions by actual outbreak source is shown in figure 5.

We developed an online, user-friendly, publicly available tool for investigators to use prospectively during an enteric disease outbreak investigation (Figure 2): <https://coe-foodsafetytools.shinyapps.io/sourceattribution/>. Fields required in the tool were total cases, month of first illness onset, geography of exposures, etiology, and *Salmonella* serotype (required only if *Salmonella* was selected as the etiology). Additional optional fields included number of male and female cases, number of cases hospitalized, and number of cases in each age category. These fields were optional because this information is not always available at the beginning of an outbreak investigation, although the tools performs more reliably with information from all fields.

*Will add more about external validation (CO data, NORS data)*

**DISCUSSION**

In this study, we developed a web-based application to predict outbreak vehicles using outbreak characteristics and demographics from historical STEC and *Salmonella* outbreaks. During an outbreak investigation, investigators often use descriptive epidemiologic data to generate hypotheses about the source of an outbreak, but few studies have systematically used outbreaks to determine if outbreak characteristics and demographics are associated with outbreak sources. In a previous study, we explored the use of outbreak characteristics and demographics to predict food vehicles in STEC outbreaks (White et al. 2016). Here, we used improved statistical methods and included *Salmonella* outbreaks. We found that outbreak characteristics and demographics can predict major food and animal sources in enteric disease outbreaks, and public health investigators can use this information in a web-based application for hypothesis generation during prospective investigations. This tool provides data-driven support for using demographics and historical outbreak information, such as food-pathogen pairings, in hypothesis generation.

We identified four distinct outbreak profiles: animal contact, produce, eggs, and meat-poultry. Outbreaks were not classified in these profiles if there were small sample size and they could not be combined with one of the other profiles. These included game, grains-beans, nuts-seeds, oils-sugars, aquatic animals, and not otherwise specified land animals or plants.

Animal contact outbreaks were more likely to have a high proportion of children, occur in winter months, be single county or multistate, and have a higher proportion female. Children aged 1-4 years was the age category most often associated with animal contact outbreaks, which is consistent with other reports, which attribute the higher risk in children to increased contact with animal reservoirs (e.g., petting zoos), inadequate handwashing, and more frequent hand-to-mouth activities (Dunn, 2015). We found outbreaks due to animal contact were more likely to occur in winter months. Most bacterial enteric disease outbreaks occur in summer months 8, which has been partially attributed to increased prevalence of pathogenic bacteria during these months in livestock; however, some studies have shown fecal shedding is more common in winter months when livestock are in closer proximity (Ogden, 2004; Stein 2017). Seasonal patterns are likely influenced by both environmental factors (pathogen reservoirs, transmission pathways, and pathogen-host interactions), as well as population characteristics (seasonal farming and activities, population mobility patterns, and changes in host susceptibility) 8.

We found outbreaks due to animal contact were most likely to be due to exposures in single counties (47%), although multistate outbreaks were also common (38%). This is likely due to two major types of animal contact outbreaks: those that are localized and occur in public spaces at petting zoos, farm visits, fairs 11, and those due to commercially distributed domestic pets, including live poultry, reptiles and amphibians, and rodents 12,13.

Produce outbreaks were more likely to be multistate, have a higher median number of cases, and have a higher proportion female. Multistate outbreaks are more likely to be associated with produce, which is often commercially distributed 14. As demand for fresh produce has increased, the produce industry has responded by changing agricultural, processing, and distribution practices to increase supply and quality (e.g., triple-washing pre-packaged leafy greens), which may have contributed to an increase in widespread outbreaks linked to produce 14.

Food consumption patterns vary by demographic factors, including income, gender, race/ethnicity, and age 15–17. Previous studies have shown women are more likely to follow healthy eating recommendations and consume more fiber and fruit than men 18. The FoodNet Population Survey found women reported higher rates of consumption for almost all fresh fruits and vegetables, including alfalfa sprouts and leafy greens 19.

Meat-poultry and egg outbreaks were more likely to occur in summer and in a single county. We grouped poultry with meat because predictive factors were more similar than with eggs. Differences between meat-poultry and egg outbreaks were driven by *Salmonella* serotype.

We found *Salmonella* outbreak vehicles were substantially different byserotype. Some serotypes were majority associated with produce sources (Javiana, Saintpaul, Newport), eggs (Enteritidis), animal contact (Montevideo, Paratyphi B, I 4,[5],12:i:-), and meat-poultry (Heidelberg, Typhimurium). This corresponds with similar findings from Jackson, et al, who found that certain serotypes were predominantly attributed to animal-derived food commodities, and others to plant-derived food commodities, likely associated with different animal reservoirs. For example, plant-associated outbreaks are rarely identified in *Salmonella* reservoir studies of livestock, suggesting potential non-livestock reservoirs, such as environmental, amphibian, or reptile reservoirs 1. The association between serotypes and foods is also supported by case-control studies of sporadic illness, including *Salmonella* Enteritidis and chicken consumption outside the home and undercooked eggs 20.

The final model used for the prediction tool included the following predictors: total number of cases, season of first illness onset, geography, percentage hospitalized, percentage female, and percentage age <1 year, 1-4, 5-9, 10-19, 20-49 50-74, and ≥ 75. We included fields in the final tool based on predictiveness and utility. We found duration of exposure was predictive of outbreak source, but this is not a helpful field during an ongoing investigation, where the duration of the outbreak would be unknown, so we excluded duration from the model. Some characteristics in the tool are required (total cases, month of first illness onset, geography of exposures, infectious agent, serotype for *Samonella*), and some are not (percent female, percent hospitalized, age distributions).

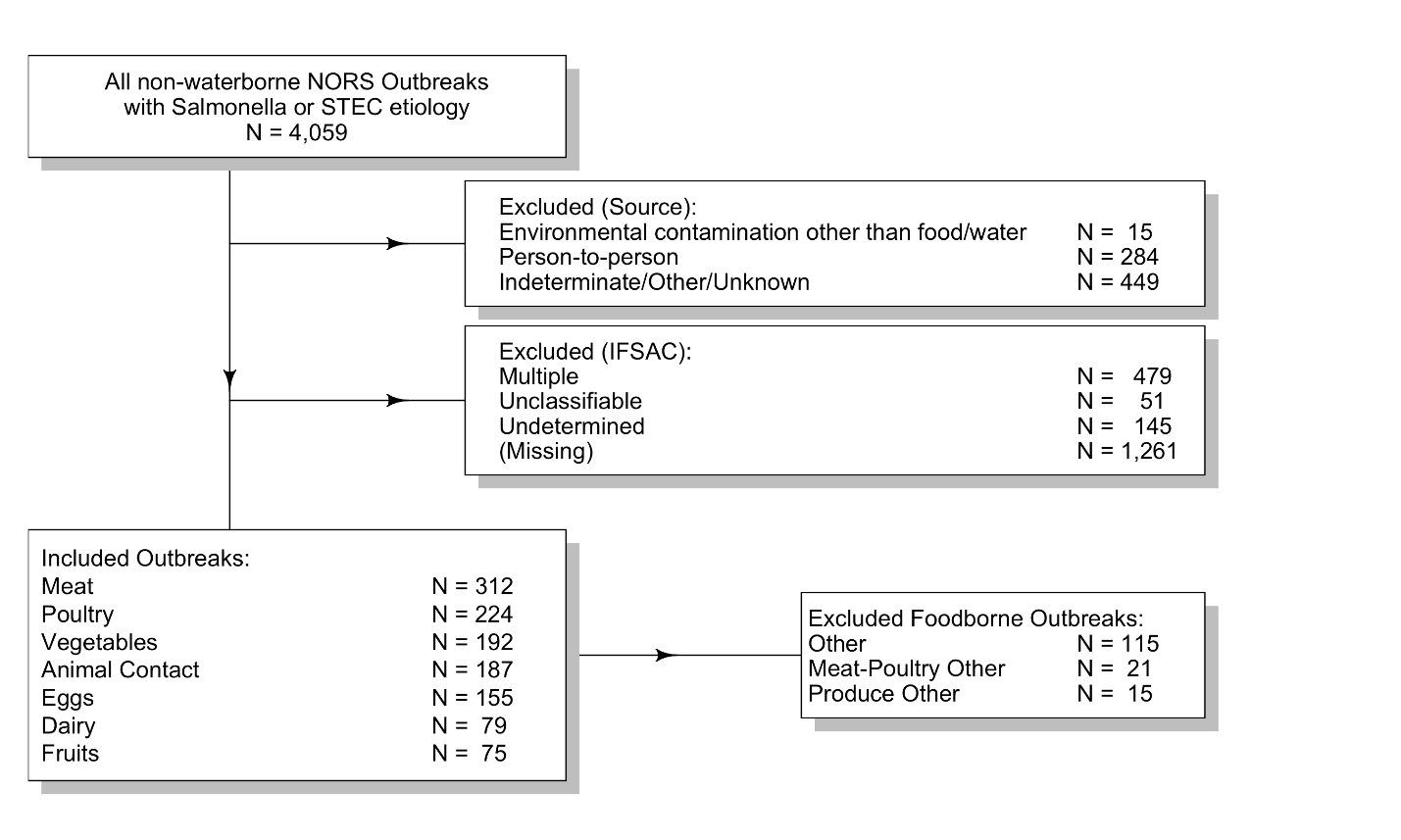
Our results indicate investigators could use the tool as a resource to help rule out potential sources, or to narrow the scope of a hypothesis. During an outbreak investigation, resources are limited and epidemiologic studies are resource and time-intensive. Developing a robust hypothesis can help direct an investigation and mitigate the use of excessive resources to do blind epidemiologic studies. When using the tool, investigators should consider the sources with the highest predicted probability in the context of the investigation and can feel more confident about excluding or spending less time investigating sources with a lower predicted probability. While investigators should not rule out potential sources without compelling evidence, an investigator might ask questions that are more thorough on an outbreak questionnaire about the top two predicted sources, for example. The tool can also serve as a reminder to consider other sources, even if there is a strong signal in the investigation.

In addition to using the tool during prospective outbreak investigations, the model developed here could be used to aid foodborne illness source attribution studies. In attribution studies, outbreaks with an undetermined source are excluded, however unsolved outbreaks represent a substantial proportion of all reported outbreaks. The model presented here could be applied to unsolved reported outbreaks to obtain additional data for attribution estimates.

We used historic outbreak data to build this tool, which had several limitations in addition to those described previously. We were limited in modeling by the number of past outbreaks in the development of the major food categories (animal contact, meat/poultry, produce, eggs). For example, we combined vegetable and fruit outbreaks to increase the sample size; however, vegetable outbreaks rarely involve children, whereas fruit outbreaks do. Missing data was a substantial issue and likely affected the accuracy of the model, particularly for age and gender distributions. The model assumes that outbreaks with missing are similar and that 'missingness' on its own is predictive. However, the reasons why predictors are missing may differ between outbreaks and during a prospective investigation. Outbreaks are reported to NORS after the outbreak has concluded. The tool is intended for use during an ongoing investigation, however many parameters change over time. For example, certain populations (e.g., children) may be more likely to seek care early in an investigation. For this analysis, we used historical outbreaks reported to CDC since 1998. There have been notable changes to surveillance during that time, as well as changes in the technologies used to detect outbreaks. The outbreak data used to build the model may be biased to older data and trends, when more meat-poultry outbreaks were detected, fewer produce outbreaks were detected, and animal contact outbreaks were not systematically reported. We included only foodborne and animal contact outbreaks in the model. At the beginning of an outbreak, investigators may not know the primary mode of transmission. Finally, some outbreaks may not be reported to NORS, and certain outbreaks may be more likely to be reported and have more complete data.

Given these limitations, this tool should be used as intended for hypothesis generation and in conjunction with other outbreak investigation methods to build quality hypotheses and test hypotheses using rigorous epidemiological methods. This tool complements other methods to generate hypotheses, including case exposure assessment using hypothesis generating questionnaires and binomial probability calculations 21,22. This tool integrates traditional hypothesis generating methods, including descriptive data and food-pathogen pairs. It is particularly powerful in combining these methods in one algorithm. However, there are additional techniques, such as mapping the geographical spread of cases that would be useful additions. The predictive results of the tool could be particularly useful in deciding which sources to emphasize when developing an outbreak questionnaire, particularly when there are time limitations. However, it is important to remember that the results are based on historical outbreaks and probabilities, and a potential source should never be eliminated based on the results of the tool, particularly when there is compelling evidence suggesting it should be considered. Because historical data was used and there were limited outbreak source categories, investigators should be cautious about using the tool and eliminating novel outbreak vehicles or uncommon vehicles. Finally, hypothesis generation should always be followed with rigorous and appropriate hypothesis testing.

**CONCLUSION**

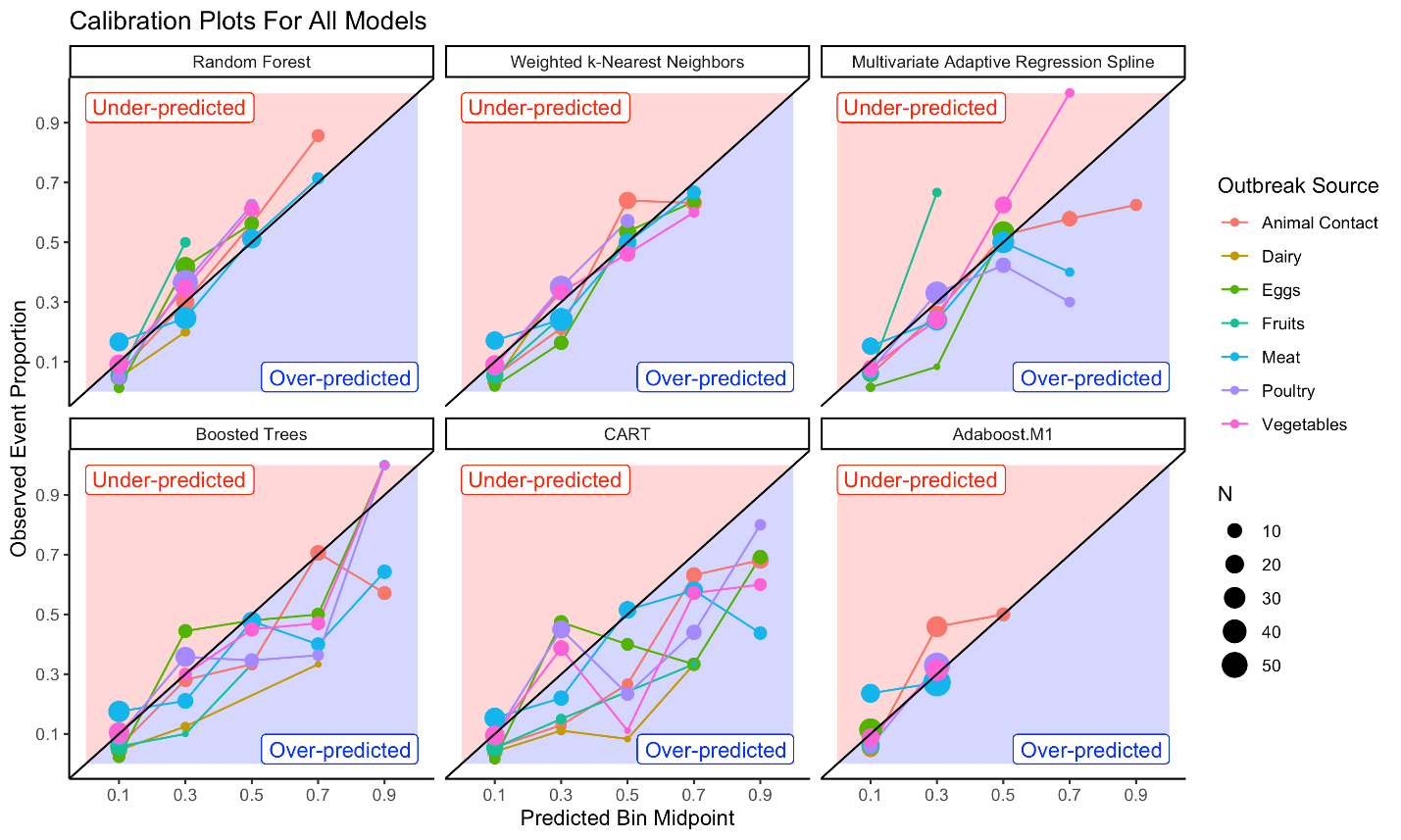
****

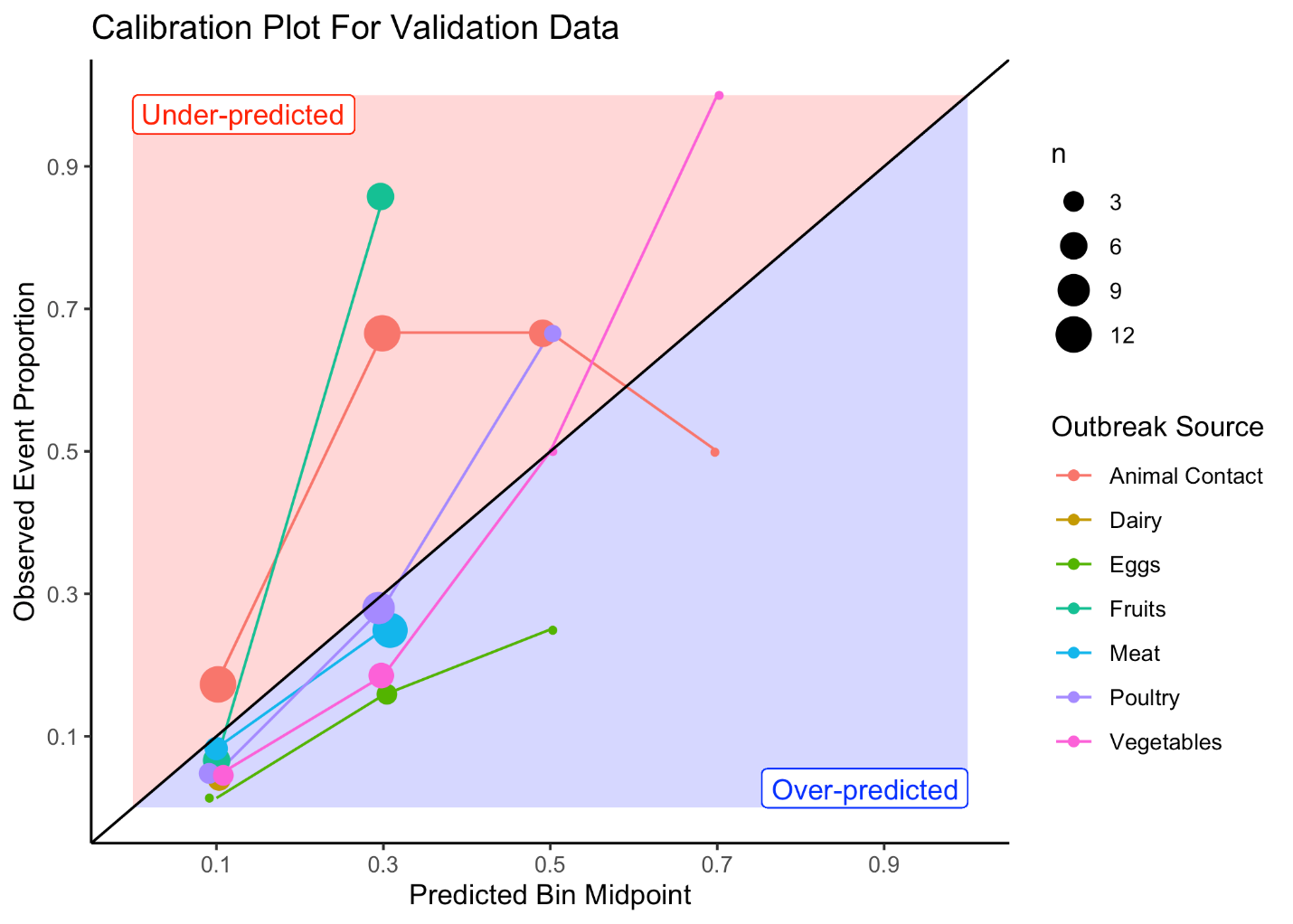
|  |
| --- |
| **Table 1.** Demographic and outbreak characteristics by source for reported *Salmonella* and STEC foodborne and animal contact outbreaks, United States, 1998-2016 |
|  |

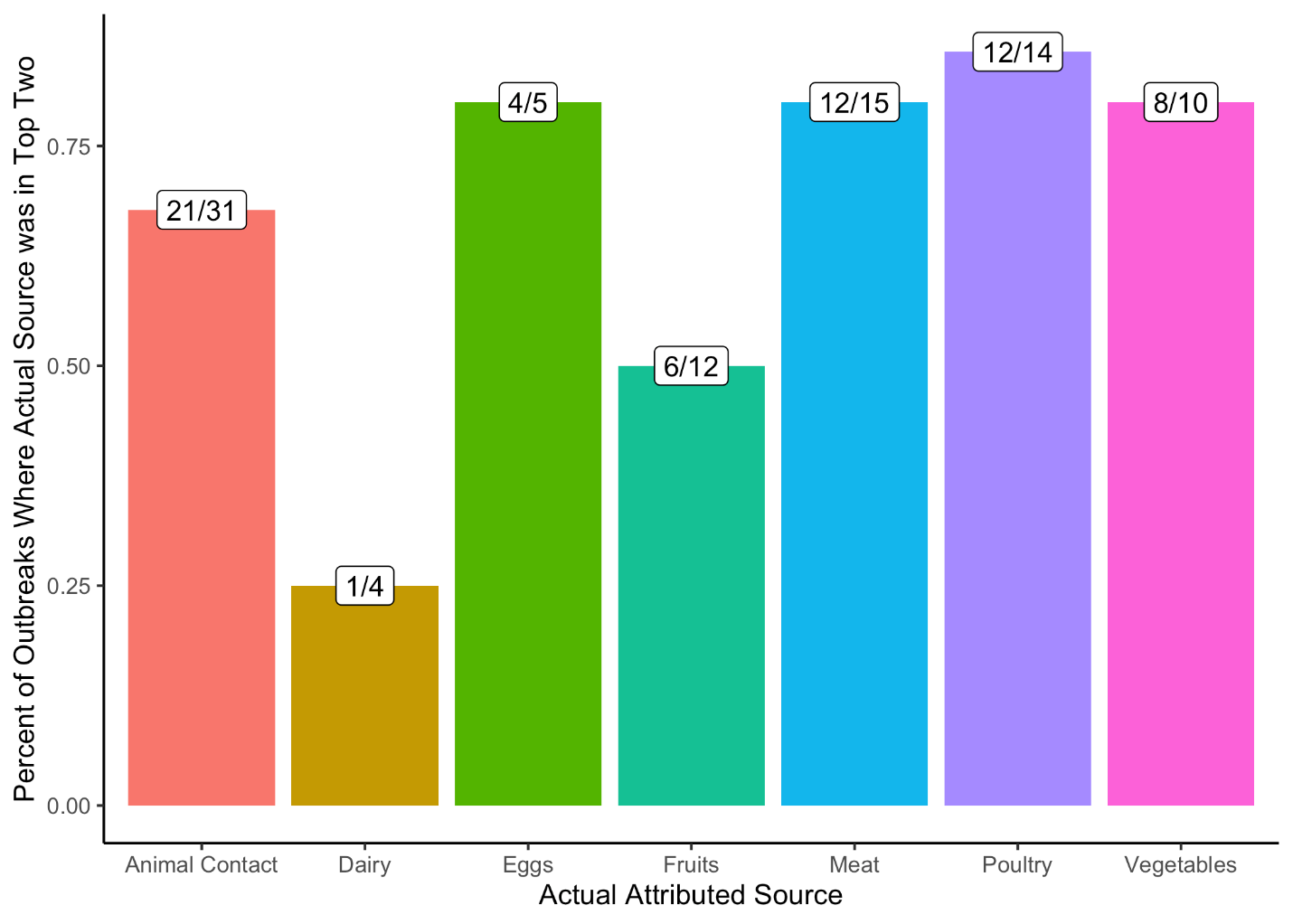
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Animal Contact** | **Dairy** | **Eggs** | **Fruits** | **Meat** | **Poultry** | **Vegetables** | **Other** |
|  |  | **N= 187** | **N= 79** | **N= 155** | **N= 75** | **N= 312** | **N= 224** | **N= 192** | **N= 151** |
| **Season** | Winter | 65(35) | 16(20) | 27(17) | 8(11) | 41(13) | 27(12) | 37(19) | 25(17) |
|  | Spring | 51(27) | 19(24) | 37(24) | 27(36) | 104(33) | 61(27) | 57(30) | 39(26) |
|  | Summer | 43(23) | 27(34) | 59(38) | 24(32) | 112(36) | 74(33) | 50(26) | 56(37) |
|  | Fall | 28(15) | 17(22) | 32(21) | 16(21) | 55(18) | 62(28) | 48(25) | 31(21) |
| **Geography** | Multi County | 25(13) | 30(38) | 12(8) | 12(16) | 49(16) | 14(6) | 34(18) | 9(6) |
|  | Multi State | 71(38) | 8(10) | 4(3) | 33(44) | 55(18) | 14(6) | 97(51) | 33(22) |
|  | Single County | 87(47) | 41(52) | 137(88) | 28(37) | 206(66) | 195(87) | 60(31) | 107(71) |
|  | Missing | 4(2) | – | 2(1) | 2(3) | 2(1) | 1(0) | 1(1) | 2(1) |
| **Total Cases** | Mean (Sd) | 30.99(55.88) | 24.00(35.96) | 36.94(158.52) | 54.28(119.66) | 22.46(38.42) | 26.08(58.45) | 56.57(139.31) | 47.23(127.68) |
| **Percent Male** | Mean (Sd) | 42.12(23.19) | 49.11(22.41) | 49.81(23.39) | 36.21(17.70) | 50.61(23.51) | 49.02(26.53) | 37.94(19.08) | 47.17(21.30) |
| **Percent Female** | Mean (Sd) | 57.88(23.19) | 50.89(22.41) | 50.19(23.39) | 63.79(17.70) | 49.39(23.51) | 50.98(26.53) | 62.06(19.08) | 52.83(21.30) |
| **Percent Under 1yr** | Mean (Sd) | 8.92(13.20) | 2.45(8.56) | 0.68(4.63) | 1.73(3.60) | 0.76(4.06) | 0.69(3.16) | 0.69(2.43) | 1.30(6.53) |
| **Percent 1yr to 4yr** | Mean (Sd) | 27.24(27.23) | 19.65(21.22) | 2.77(6.73) | 11.37(16.50) | 7.55(16.47) | 5.64(12.53) | 4.63(11.09) | 6.82(14.16) |
| **Percent 20yr to 49yr** | Mean (Sd) | 20.36(21.92) | 27.52(24.82) | 46.73(30.64) | 25.71(20.86) | 41.72(29.70) | 50.53(30.04) | 50.53(22.96) | 49.59(29.52) |
| **Percent 5yr to 19yr** | Mean (Sd) | 31.53(29.37) | 36.16(26.06) | 17.12(24.49) | 25.91(25.29) | 27.35(28.99) | 19.33(25.60) | 19.71(18.48) | 19.00(22.71) |
| **Percent 50yr or older** | Mean (Sd) | 11.95(17.62) | 14.22(19.83) | 32.70(31.06) | 35.29(31.76) | 22.63(24.66) | 23.81(28.02) | 24.44(20.83) | 23.29(26.57) |
| **Percent Hospitalized** | Mean (Sd) | 0.28(0.26) | 0.31(0.28) | 0.18(0.24) | 0.27(0.19) | 0.31(0.30) | 0.23(0.28) | 0.27(0.23) | 0.24(0.25) |
| **Length (in days) of Outbreak** | Mean (Sd) | 43.92(125.49) | 33.53(57.55) | 8.77(20.34) | 7.23(11.26) | 7.48(19.54) | 5.88(42.01) | 10.47(18.43) | 8.13(15.06 |

| Table 2. Tuning Parameters | |
| --- | --- |
| **Model** | **Parameters** |
| Random Forest | mtry = 2; split rule = gini; min node size = 1 |
| Weighted k-nearest neighbors | k = 44 |
| Multivariate Adaptive Regression Spline | degree = 2; nprune = 18 |
| Boosted Trees | eta = 0.3; max depth = 6; gamma = 0 |
| CART | trials = 20; model type = ‘rules’; no winnowing |
| Adaboost.M1 | mfinal = 9, max depth = 1, coefficient type = ‘Zhu’ |

| Table 3. Brier Scores for each model | |
| --- | --- |
| **Model** | **Brier Score** |
| Random Forest | 0.098 |
| Weighted k-Nearest Neighbors | 0.098 |
| Multivariate Adaptive Regression Spline | 0.100 |
| Boosted Trees | 0.103 |
| CART | 0.106 |
| Adaboost.M1 | 0.110 |
| Null | 0.118 |

****

****

**..**

**References**

1. Jackson, B. R., Griffin, P. M., Cole, D., Walsh, K. A. & Chai, S. J. Outbreak-associated salmonella enterica serotypes and food commodities, united states, 1998-2008. *Emerg. Infect. Dis.* **19**, 1239–1244 (2013).

2. White, A., Cronquist, A., Bedrick, E. J. & Scallan, E. Food Source Prediction of Shiga Toxin–Producing *Escherichia coli* Outbreaks Using Demographic and Outbreak Characteristics, United States, 1998–2014. *Foodborne Pathog. Dis.* **13**, 527–534 (2016).

3. Prevention, C. for D. C. and. The National Outbreak Reporting System (NORS). About NORS. Available at: http://www.cdc.gov/NORS/about.html. (Accessed: 20th November 2015)

4. Painter, J. a *et al.* Recipes for foodborne outbreaks: a scheme for categorizing and grouping implicated foods. *Foodborne Pathog. Dis.* **6**, 1259–1264 (2009).

5. (IFSAC), I. F. S. A. C. Food Categorization Scheme. *Centers for Disease Control and Prevention* (2013). Available at: http://www.cdc.gov/foodsafety/ifsac/projects/food-categorization-scheme.html.

6. White, A., Cronquist, A., Bedrick, E. & Scallan, E. Food Source Prediction of Shiga Toxin – Producing Escherichia coli Outbreaks Using Demographic. *Foodborne Pathog. Dis.* **13**, 527–534 (2016).

7. Dunn, J. R., Behravesh, C. B. & Angulo, F. J. Diseases Transmitted by Domestic Livestock: Perils of the Petting Zoo. *Microbiol. Spectr.* 1–8 (2015).

8. Lal, A., Hales, S., French, N. & Baker, M. G. Seasonality in human zoonotic enteric diseases: A systematic review. *PLoS One* **7**, (2012).

9. Ogden, I. D., MacRae, M. & Strachan, N. J. C. Is the prevalence and shedding concentrations of E. coli O157 in beef cattle in Scotland seasonal? *FEMS Microbiol. Lett.* **233**, 297–300 (2004).

10. Stein, R. A. & Katz, D. E. Escherichia coli, cattle and the propagation of disease. *FEMS Microbiol. Lett.* **364**, 1–11 (2017).

11. Conrad, C., Stanford, K., Narvaez-Bravo, C., Callaway, T. & McAllister, T. Farm Fairs and Petting Zoos: A Review of Animal Contact as a Source. *Foodborne Pathog. Dis.* **14**, (2017).

12. Basler, C., Nguyen, T.-A., Anderson, T. C., Hancock, T. & Behravesh, C. B. Outbreaks of Human Salmonella Infections Associated with Live. *Emerg. Infect. Dis.* **22**, 1705–1711 (2016).

13. Burke, C. *et al.* Multistate oubtreak of human Salmonella Typhimurium infections associated with pet turtle exposure - United States, 2008. *MMWR* **59**, (2010).

14. Berger, C. N. *et al.* Minireview Fresh fruit and vegetables as vehicles for the transmission of human pathogens. *Environ. Microbiol.* **12**, 2385–2397 (2010).

15. French, S. A., Tangney, C. C., Crane, M. M., Wang, Y. & Appelhans, B. M. Nutrition quality of food purchases varies by household income : the SHoPPER study. *BMC Public Health* 231 (2019).

16. Shiferaw, B. *et al.* Antimicrobial susceptibility patterns of shigella isolates in foodborne diseases active surveillance network (foodnet) sites, 2000-2010. *Clin. Infect. Dis.* **54**, (2012).

17. Kim, S. A., Moore, L. V, Galuska, D., Wright, A. P. & Harris, D. Vital Signs : Fruit and Vegetable Intake Among Children — United States , 2003 – 2010. *Morb. Mortal. Wkly. Rep.* **63**, 671–676 (2014).

18. Wardle, J. *et al.* Gender Differences in Food Choice: The Contribution of Health Beliefs and Dieting. *Ann. Behav. Med.* **27**, 107–116 (2004).

19. Shiferaw, B. *et al.* Sex-based differences in food consumption: Foodborne diseases active surveillance network (FoodNet) population survey, 2006-2007. *Clin. Infect. Dis.* **54**, 2006–2007 (2012).

20. Kimura, A. C. *et al.* Chicken Consumption Is a Newly Identified Risk Factor for Sporadic Salmonella enterica Serotype Enteritidis Infections in the United States: A Case-Control Study in FoodNet Sites. *Clin. Infect. Dis.* **38**, S244-52 (2004).

21. Jervis, R. *et al.* *Moving away from population-based case-control studies during outbreak investigations*. (2018).

22. Keene, W. The use of binomial probabilites in outbreak investigations. in *15th Annual PulseNet Conference and 7th Annual OutbreakNet Conference* (2011).