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

**Authors**

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

**ABSTRACT**

**INTRODUCTION**

Hypothesis generation about the 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 (Jackson 2013). Investigators also use descriptive data of cases to suggest a food source. For example, case demographic data, including age, sex 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 STEC O157 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 2016). However, a number of limitations were noted. First, 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, age could not be used 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 2016. This passive surveillance system receives outbreak reports from state, local, and territorial health agencies using a standard form (CDC, 2009). 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 outbreaks with an implicated food vehicle were categorized into food groups using the Interagency Food Safety Analytics Collaboration (IFSAC) Food Categorization Scheme (Painter 2009, IFSAC 2013). 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 food (Painter et al., 2009). Only foods categorized as a single commodity were used for this analysis.

**Data analysis**

We included any outbreak with a completed outbreak report that listed STEC or *Salmonella* listed as the confirmed or suspected as the etiology. Only foodborne or animal contact outbreak were included. Foodborne outbreaks were excluded if they were missing IFSAC information, if they were caused by multiple sources, unclassifiable, or undetermined source.

Data were split into a training set (75%) and a testing set (25%). Outbreaks with rare food sources (fewer than 100 outbreaks) were excluded from the training set only. In order to more accurately reflect actual usage, foodborne outbreaks of other origin were included in the testing set, outbreaks with multiple, unclassifiable and no identified food sources were not.

**Predictors**

Case demographic predictors included percentage of female and male cases, percentage of cases hospitalized, and percentage of cases in each age group (<1 year, 1-4, 5-9, 10-19, 20-49, 50-74, ≥75). Outbreak predictors included etiology (STEC or *Salmonella* serotype), number of cases (confirmed or suspected), season, and 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).

**Model Selection**

We selected seven algorithmic methods for prediction based on their ability to predict multiple class probabilities. These models included bagged adaptive boosting classification trees, classification and regression trees (CART), weighted k nearest neighbors (knn), flexible discriminant analysis (FDA), weighted subspace random forest, Naive Bayes, and rule-based classifier. A non-informative model that uses no information from predictors was also generated for comparison purposes. The final model was chosen based on Brier Scores (a measure of the difference in the predicted probability and the actual event). Outbreaks with other food source origins not included in our training data were included in the brier score calculations. All analyses were performed using R version 3.5.1 (2018-07-02) with the Caret Package v(6.0.81. Parameter selection was performed using the Caret package.

**Tool Development**

We developed an online, user-friendly tool for investigators to use prospectively during an enteric disease outbreak investigation (Figure 2), which is publicly available: <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). Other optional fields included number of male and female cases, number of cases hospitalized, and number of cases in each age category. These fields were made optional because this information is not always available at the beginning of an outbreak investigation, although the tools is more reliable if all fields are entered.

We field tested the tool by asking public health outbreak investigators the Colorado Department of Public Health and Environment (CDPHE) to select several outbreaks not included in the application development to test the prospective efficacy.

**RESULTS**

From 1998 to 2016, 607 STEC outbreaks and 2,704*Salmonella* outbreaks that were primarily transmitted via a food source or animal contact were reported. Of these, 1,936 outbreaks were excluded from model development for the following reasons: missing IFSAC information (n=1,261), caused by multiple food sources (n=479), unclassifiable food sources (n=51), and undetermined food sources (n=145) leaving analysis data set of 1,375 outbreaks (305 STEC and 1070 *Salmonella*). Twelve outbreaks from a food source other than animal or plant 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) were excluded from the training data. 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).

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).

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).

CDPHE investigators selected four outbreaks to prospectively test the tool. Three of the fours outbreaks were correctly predicted based on the highest predicted probability. An outbreak of *Salmonella* Typhimurium attributed to live poultry with 7 total cases in a single county (X male, Y female; age distribution), no hospitalizations, and an onset in March had the following predicted probabilities: 19% animal contact, 15% eggs, 43% meat-poultry, and 23% produce. An outbreak of *Salmonella* Oranienburg linked to a burger restaurant chain with X total cases in a single county (X male, Y female; age distribution), no hospitalizations, and an onset in August had the following predicted probabilities: 6% animal contact, 17% eggs, 69% meat-poultry, and 8% produce. An unsolved outbreak of *Salmonella* Enteritidis in a childcare center with 17 cases in a single county (X male, Y female; age distribution), no hospitalizations, and an onset in May had the following predicted probabilities: 45% animal contact, 14% eggs, 30% meat-poultry, and 11% produce. Finally, an STEC outbreak linked to a cattle branding party with 3 total cases in a single county (1 male, 2 female; 2 cases 10-19 years, 1 case 50-74 years), no hospitalizations, and onset in April had the following predicted probabilities: 47% meat-poultry, 23% animal contact, 21% produce, and 9% eggs.

**DISCUSSION**

**CONCLUSION**

**Table 1.** Etiology by food or animal source for reported *Salmonella* and STEC outbreaks, United States, 1998-2016

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Eggs** | **Meat-Poultry** | **Produce** | **Animal Contact** | **Othera** |
|  | n (%) | n (%) | n (%) | n (%) | n (%) |
| **Total** | 155 | 536 | 282 | 187 | 215 |
| **STEC** | 23 (15) | 121 (23) | 64 (23) | 48 (26) | 49 (23) |
| **Salmonella** |  |  |  |  |  |
| Braenderup | 1 (1) | 8 (1) | 7 (2) | 6 (3) | 1 (0) |
| Enteritidis | 103 (66) | 70 (13) | 20 (7) | 7 (4) | 19 (9) |
| Heidelberg | 7 (5) | 25 (5) | 1 (0) | 2 (1) | 6 (3) |
| I 4,[5],12:i:- | 0 (0) | 12 (2) | 2 (1) | 9 (5) | 4 (2) |
| Javiana | 0 (0) | 5 (1) | 14 (5) | 0 (0) | 3 (1) |
| Montevideo | 0 (0) | 6 (1) | 2 (1) | 9 (5) | 4 (2) |
| Newport | 0 (0) | 25 (5) | 33 (12) | 0 (0) | 8 (4) |
| Paratyphi B | 0 (0) | 0 (0) | 1 (0) | 3 (2) | 4 (2) |
| Saintpaul | 1 (1) | 4 (1) | 13 (5) | 1 (1) | 3 (1) |
| Typhimurium | 5 (3) | 46 (9) | 14 (5) | 28 (15) | 28 (13) |
| Nonspecific serotype | 1 (1) | 22 (4) | 20 (7) | 8 (4) | 9 (4) |
| Primary animal serotype | 5 (3) | 40 (7) | 3 (1) | 20 (11) | 8 (4) |
| Primary plant serotype | 0 (0) | 1 (0) | 14 (5) | 3 (2) | 3 (1) |
| Rare serotype | 1 (1) | 22 (4) | 12 (4) | 10 (5) | 19 (9) |
| Unknown serotype | 6 (4) | 117 (22) | 53 (19) | 23 (12) | 38 (18) |
| Multiple serotypes | 2 (1) | 12 (2) | 9 (3) | 10 (5) | 9 (4) |
| a game, grains-beans, nuts-seeds, oils-sugars, aquatic animals, and not otherwise specified land animals or plants. | | | | | |

**Table 2**. Demographic and outbreak characteristics by food or animal source for reported *Salmonella* and STEC outbreaks, United States, 1998-2016

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Eggs** | **Meat-Poultry** | **Produce** | **Animal Contact** | **Othera** |
|  | n (%) | n (%) | n (%) | n (%) | n (%) |
| **Total** | 155 | 536 | 282 | 187 | 215 |
| **Season** |  |  |  |  |  |
| Winter (Jan-Mar) | 27 (17) | 68 (13) | 45 (16) | 65 (35) | 41 (19) |
| Spring (Apr-Jun) | 37 (24) | 165 (31) | 88 (31) | 51 (27) | 54 (25) |
| Summer (Jul-Sep) | 59 (38) | 186 (35) | 83 (29) | 43 (23) | 74 (34) |
| Fall (Oct-Dec) | 32 (21) | 117 (22) | 66 (23) | 28 (15) | 46 (21) |
| **Geography** |  |  |  |  |  |
| Multi-State | 4 (3) | 69 (13) | 131 (46) | 71 (38) | 40 (19) |
| Multi-County | 12 (8) | 63 (12) | 47 (17) | 25 (13) | 38 (18) |
| Single County | 137 (88) | 401 (75) | 100 (35) | 87 (47) | 136 (63) |
| Missing | 2 (1) | 3 (1) | 4 (1) | 4 (2) | 1 (0) |
| **Hospitalized** |  |  |  |  |  |
| % Hospitalized | 16 (21) | 24 (27) | 22 (20) | 22 (24) | 24 (24) |
| Missing | 20 (13) | 65 (12) | 32 (11) | 9 (5) | 17 (8) |
|  | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) |
| **Gender** |  |  |  |  |  |
| % Male | 39 (28) | 43 (28) | 33 (21) | 37 (25) | 43 (26) |
| % Female | 40 (29) | 43 (29) | 55 (26) | 50 (29) | 45 (26) |
| % Unknown | 21 (41) | 13 (34) | 12 (32) | 13 (32) | 12 (32) |
| **Age (years)** |  |  |  |  |  |
| % <1 | 0.059 (0.41) | 0.32 (2.37) | 0.44 (1.78) | 7.3 (12.17) | 0.74 (5.02) |
| % 1-4 | 0.63 (3.2) | 2.7 (10.2) | 2.5 (8.9) | 23 (26.8) | 6 (15.5) |
| % 5-9 | 0.96 (6.2) | 2.4 (9.2) | 3 (7.7) | 12 (19.4) | 4.5 (12.6) |
| % 10-19 | 2.3 (9.3) | 4.4 (11.9) | 5.8 (11.9) | 14 (20.8) | 6.5 (14.8) |
| % 20-49 | 10 (25) | 16 (27) | 20 (27) | 16 (20) | 16 (25) |
| % 50-74 | 5 (13) | 6.6 (14) | 10 (17) | 7.9 (14) | 6.7 (16) |
| % ≥ 75 | 1.6 (8.5) | 1.7 (7.2) | 2.9 (7.8) | 2 (6.6) | 1.4 (4.6) |
| % Unknown | 58 (49) | 53 (49) | 44 (49) | 10 (27) | 46 (49) |
|  | Median (IQR) | Median (IQR) | Median (IQR) | Median (IQR) | Median (IQR) |
| **Total Cases** | 10 (4-27) | 10 (4-22) | 19 (10-46) | 8 (3-31) | 12 (5-32) |
| a game, grains-beans, nuts-seeds, oils-sugars, aquatic animals, and not otherwise specified land animals or plants | | | | | |

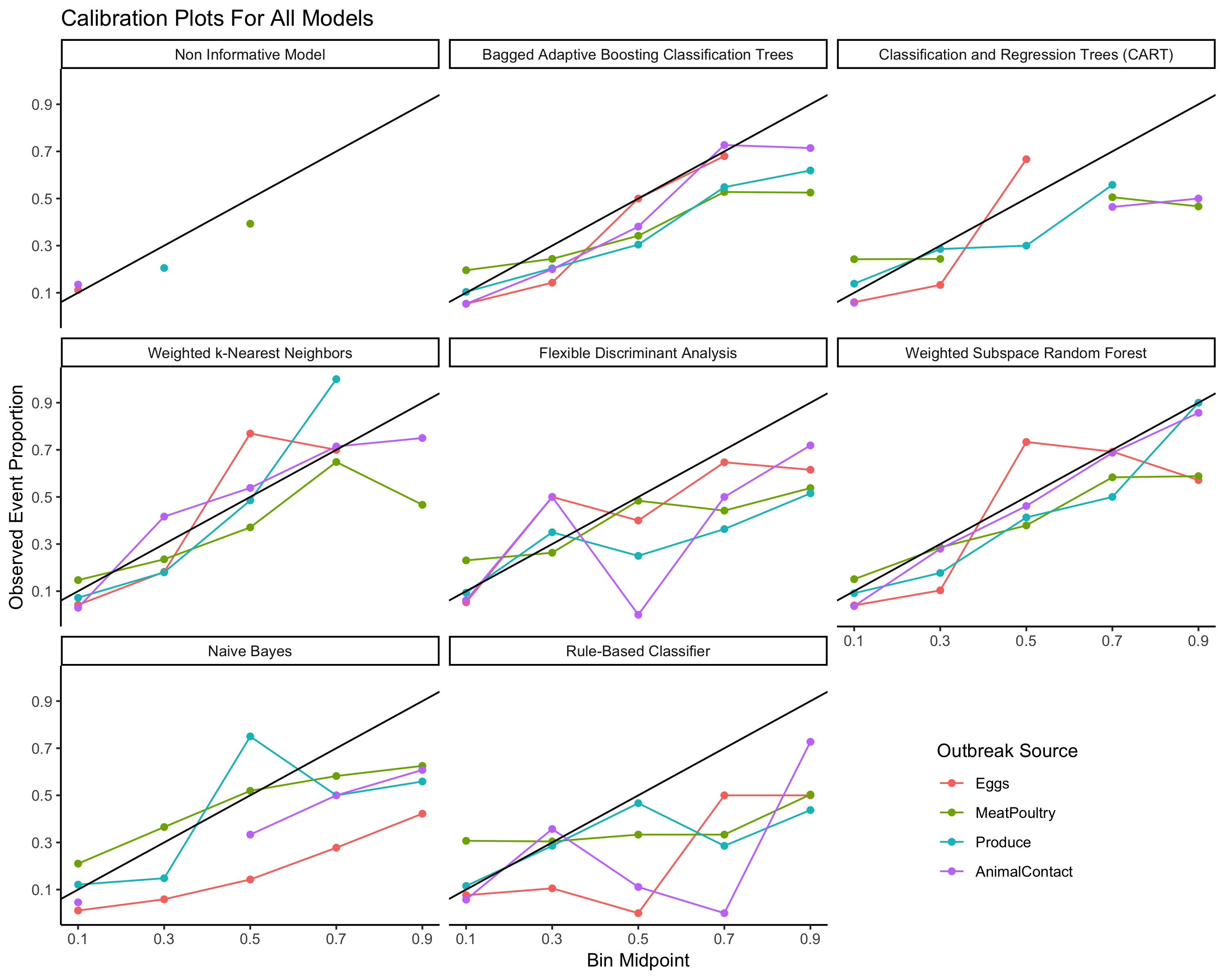
**Table 3**. STEC and *Salmonella* outbreak sources in testing and training dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Eggs** | **Meat-Poultry** | **Produce** | **Animal Contact** | **Other** |
|  | n (%) | n (%) | n (%) | n (%) | n (%) |
| Testing Dataset | 38 (11) | 134 (39) | 70 (21) | 46 (14) | 53 (16) |
| Training Dataset | 117 (12) | 402 (41) | 212 (21) | 141 (14) | 117 (12) |

**Table 4**. Percent of time source and the given predicted probability was correct.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0-20% | 20-40% | 40-60% | 60-80% | 80-100% |
| **Eggs** | 4 | 18 | 77 | 70 | NA |
| **Meat-Poultry** | 15 | 24 | 37 | 65 | 47 |
| **Produce** | 7 | 18 | 49 | 100 | NA |
| **Animal Contact** | 3 | 42 | 54 | 71 | 75 |
| **All Outbreaks** | 5 | 23 | 44 | 68 | 57 |

**Figure 1**. Calibration plots for statistical models



**Figure 2**. Picture of tool

**References**