FOODBORNE PATHOGENS AND DISEASE
Volume XX, Number XX, 2016
© Mary Ann Liebert, Inc.
DOI: 10.1089/fpd.2016.2140

Original Article

Food Source Prediction of Shiga Toxin–Producing *Escherichia coli* Outbreaks Using Demographic and Outbreak Characteristics, United States, 1998–2014

Alice White,[1] Alicia Cronquist,[2] Edward J. Bedrick,[3] and Elaine Scallan[1]

Abstract

Background: Foodborne illness is a continuing public health problem in the United States. Although outbreak-associated illnesses represent a fraction of all foodborne illnesses, foodborne outbreak investigations provide critical information on the pathogens, foods, and food-pathogen pairs causing illness. Therefore, identification of a food source in an outbreak investigation is key to impacting food safety.
Objective: The objective of this study was to systematically identify outbreak-associated case demographic and outbreak characteristics that are predictive of food sources using Shiga toxin–producing *Escherichia coli* (STEC) outbreaks reported to Centers for Disease Control and Prevention (CDC) from 1998 to 2014 with a single ingredient identified.
Materials and Methods: Differences between STEC food sources by all candidate predictors were assessed univariately. Multinomial logistic regression was used to build a prediction model, which was internally validated using a split-sample approach.
Results: There were 206 single-ingredient STEC outbreaks reported to CDC, including 125 (61%) beef outbreaks, 30 (14%) dairy outbreaks, and 51 (25%) vegetable outbreaks. The model differentiated food sources, with an overall sensitivity of 80% in the derivation set and 61% in the validation set.
Conclusions: This study demonstrates the feasibility for a tool for public health professionals to rule out food sources during hypothesis generation in foodborne outbreak investigation and to improve efficiency while complementing existing methods.

Introduction

Foodborne diseases remain a major public health challenge in the United States, where they cause an estimated 48 million illnesses, 128,000 hospitalizations, and 3000 deaths annually (Scallan *et al.*, 2011). The incidence of many important foodborne pathogens has not declined over the past decade, indicating that additional food safety interventions are needed (Crim *et al.*, 2014). Food safety progress relies on many factors, including the capacity of public health agencies to conduct surveillance for individual cases of illness and to detect and investigate foodborne illness outbreaks. Although outbreak-associated illnesses represent a fraction of all foodborne illnesses, foodborne outbreak investigations provide critical information on the pathogens, foods, and food-pathogen pairs causing illness. These data drive food safety regulations, policies, and practices while also informing estimates of the overall burden of foodborne illness and the attribution of illnesses to food sources (Tauxe, 1997, 2002; Reingold, 1998). However, ~60% of reported foodborne illness outbreaks identify a confirmed or suspect etiology, and of those, only half implicate a food vehicle (Gould *et al.*, 2013a).

Several recent initiatives, including the Council to Improve Foodborne Outbreak Response (CIFOR), the Integrated Food Safety Centers of Excellence, and the Foodborne Diseases Centers for Outbreak Response Enhancement (FoodCORE), aim at improving the quality of outbreak investigations nationwide by providing public health professionals at the local, state, and federal level with guidelines, tools, and resources to aid outbreak investigations (CDC, 2010, 2012; CIFOR, 2014). Examples of existing resources include guidelines for model practices during outbreak investigations (CIFOR, 2014), a national hypothesis-generating questionnaire (CDC, 2013), pathogen symptom profiles (Hall *et al.*, 2001; Turcios *et al.*, 2006; Hedberg *et al.*, 2008; Domínguez *et al.*, 2010),

[1]Department of Epidemiology, Colorado School of Public Health, University of Colorado Denver, Aurora, Colorado.
[2]Colorado Department of Public Health and Environment, Denver, Colorado.
[3]Center for Biomedical Informatics and Statistics, The University of Arizona Health Sciences, Tucson, Arizona.

and population-level food consumption data (CDC, 2007; OPHD, 2010). Using data from foodborne illness outbreak reports in the United States, our goal was to systematically identify factors predictive of outbreak food sources and to develop a tool for investigators to use for hypothesis generation. This is a novel, data-driven approach to hypothesis generation in foodborne illness outbreak investigations. For this study, the focus was Shiga toxin–producing *Escherichia coli* (STEC) outbreaks. STEC is one of the most common foodborne illness pathogens, causing ~6% of confirmed, single-etiology outbreaks (only norovirus and *Salmonella* cause more outbreaks). Most STEC outbreaks are attributed to beef, followed by leafy vegetables (Gould *et al.*, 2013a). There are many heterogeneous STEC serogroups that cause human gastrointestinal illness, the most common of which is O157:H7 (Gould *et al.,* 2013b).

Materials and Methods

Data source

National data on reported foodborne STEC outbreaks from 1998 to 2014 were available from the Centers for Disease Control and Prevention's (CDC) Foodborne Outbreak Surveillance System. The Electronic Foodborne Outbreak Reporting System (eFORS) collected data on foodborne and waterborne enteric disease outbreaks from 1998 to 2008. 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. These passive surveillance systems receive reports from state, local, and territorial health agencies using a standard form (CDC, 2009). Outbreaks reported to the CDC were extracted on August 27, 2015 using the following qualifications: foodborne mode of transmission, finalized report, onset year 1998–2014, and STEC etiology. CDC provided data in a Microsoft Access database. Relational data were merged into a single, flat file using SAS 9.4.

Prediction model

Using STEC outbreaks with a single ingredient identified, a multinomial logistic regression model was developed to predict food sources. The model was developed using a random subset of outbreaks and validated on the remaining outbreaks.

Food source categories

The Interagency Food Safety Analytics Collaboration (IFSAC) Food Categorization Scheme was used to identify food source categories (Supplementary Fig. S1; Supplementary Data are available online at www.liebertpub.com/fpd) (Painter *et al.*, 2009; IFSAC, 2013). IFSAC is a collaborative effort between CDC, the U.S. Food and Drug Administration (FDA), and the U.S. Department of Agriculture Food Safety and Inspection Service (USDA-FSIS), which aim at improving food source attribution. IFSAC developed the food scheme as a systematic method for food categorization in outbreak and attribution analyses. Categories are based on a taxonomic scheme of 17 mutually exclusive commodities. Foods composed of ingredients from a single commodity are ''simple'' foods, and foods with ingredients from multiple commodities are ''complex'' foods. For example, beef is a simple food, and a hamburger is a complex food (Painter *et al.*, 2009).

Demographic and outbreak predictors

Outbreak-associated case demographic predictors included percentage female (percentage of cases in an outbreak who were women) and age (percentage of cases in an outbreak aged <5, 5–19, 20–49, and ≥50 years). Outbreak predictors included the number of cases (both laboratory confirmed and epidemiologically linked), percentage hospitalized, multistate outbreak (i.e., cases that occurred in multiple states), exposure setting (private or non-private), season, outbreak duration (number of days between the date of illness onset for first and last reported case), and serogroup (O157:H7 or non-O157:H7). For exposure setting, a private establishment was defined as a location not subject to inspection, where a non-food worker would prepare and serve food (e.g., ''private home''). All other settings were non-private, defined as locations subject to inspection, where a food worker would prepare and serve food (e.g., a restaurant or facility). Season was based on onset date of the first case, and it was categorized as fall (September–November), spring (March–May), summer (June–August), or winter (December–February).

Statistical analysis

A split-sample approach was used for internal validation. The dataset was randomly divided into a derivation set (70%) and a validation set (30%). The association between each candidate predictor and food source category was assessed in univariate comparisons in the derivation set. For continuous predictors that met parametric assumptions, an analysis of variance (ANOVA) test was used to compare differences by foods. For non-normal continuous predictors, a Kruskal–Wallis test was used. For categorical predictors, a Pearson chi-squared (χ^2) test was used. A Fisher's exact test was used for categorical predictors with small cell sizes.

Candidate predictors with more complete data (<20% missingness) and univariate significance ($p < 0.10$) were included in multinomial logistic regression analysis, which is an extension of binary logistic regression for multi-category outcomes (Biesheuvel *et al.*, 2008; Barnes *et al.*, 2013; Ge *et al.*, 2013). A backwards method was used to select final predictors in the derivation set, selecting a minimal number of predictors that maintained adequate classification accuracy. The model predicted the probability of each food source (beef, dairy, vegetable), such that the predicted probabilities for each outbreak added to 100%.

The model was scored to the validation set, which did not impact univariate analysis or multivariable model building, and was evaluated based on diagnostic classification accuracy in both the derivation and validation sets. The model was also scored to both the entire dataset and random subsets (30%, 60%, and 90% of the total dataset). Trends for each predictor from type 3 analysis of effects based on the Wald χ^2 test were determined. Maximum likelihood estimates were obtained, along with odds ratio estimates with 95% Wald confidence limits. The predicted food was determined by the highest predicted probability for each outbreak based on the model. Predicted probabilities were plotted in triangle plots using the TRIPLOT macro for SAS (Graham and Midgley, 2000; Friendly, 2009; Barnes *et al.*, 2013).

All analyses were performed in SAS, version 9.4. Multinomial logistic regression was performed using the glogit link function in SAS PROC Logistic.

Results

STEC outbreaks

From 1998 to 2014, 470 STEC outbreaks were reported. Of these, 153 (33%) did not identify a food and were excluded (Fig. 1). In addition, 80 complex food outbreaks were excluded, which comprised 25% of the remaining 317 outbreaks with an identified food source. Of the 237 outbreaks with a single ingredient identified, 125 (53%) were beef, 44 (19%) were vegetable row crops, 30 (13%) were dairy, and 7 (3%) were sprouts. Vegetable row crops and sprouts were combined into a single ''vegetables'' category. The remaining 31 outbreaks represented a heterogeneous group of foods with rare STEC exposure and were excluded (Supplementary Table S1).

Univariate analysis

Of the 145 beef, dairy, and vegetable outbreaks in the derivation dataset, the median percentage female was higher for vegetables (64%) than for beef and dairy (50%) (Table 1). The median percentage aged <5 years was the highest for dairy (22%), whereas those aged 5–19 and 20–49 years were the highest for dairy (50%) and vegetables (44%), respectively. The median number of cases was the highest for vegetables ($n=18$), followed by beef ($n=9$), and dairy ($n=5$). Dairy (70%) and beef (52%) were more often reported in a private setting, whereas vegetables were more likely reported in a non-private setting (82%). The proportion of multistate outbreaks was the highest for vegetable (56%), followed by beef (36%), and dairy (10%). Seasonal trends were noted, with more vegetable outbreaks in fall (44%) and more beef in summer (40%). Outbreaks with the non-O157:H7 serogroup were the most common for dairy (25%), followed by vegetables (18%), and beef (4%) (Supplementary Table S2). The percentage of outbreaks with incomplete data was not significantly different between food sources (Table 1).

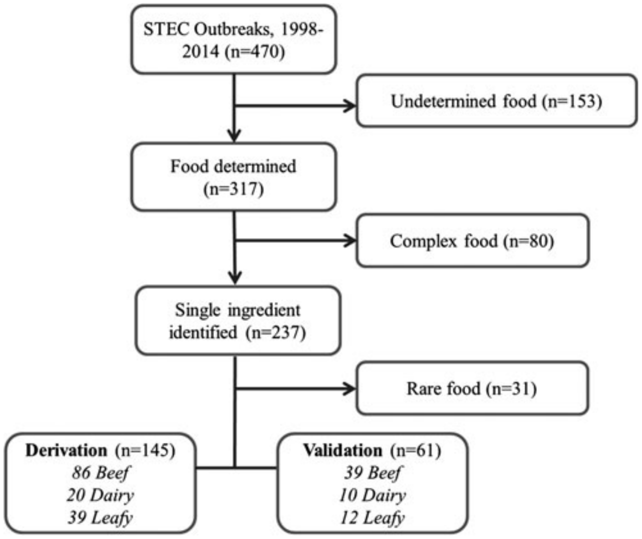

FIG. 1. Outbreaks reported to the eFORS and the NORS 1998–2014. eFORS, Electronic Foodborne Outbreak Reporting System; NORS, National Outbreak Reporting System.

Prediction model

There were 116 outbreaks in the final model after 29 outbreaks with incomplete data had been excluded. Final predictors were percentage female, number of cases, exposure setting, multistate outbreak, season, and serogroup. The model correctly classified 56 of 64 beef outbreaks (sensitivity 88%, specificity 71%); 9 of 16 dairy outbreaks (sensitivity 56%, specificity 98%); and 28 of 36 vegetable outbreaks (sensitivity 78%, specificity 93%) (Table 2). The predicted probabilities for beef outbreaks clustered in the ''beef'' apex of the plot (Fig. 2a), indicating a high predicted probability of beef for actual beef outbreaks. Similarly, vegetable outbreaks clustered in the ''leafy'' apex, indicating a high predicted probability of vegetables, with some dispersion between the beef and vegetable outbreaks. Dairy outbreaks were more dispersed between the ''dairy'' and ''beef'' apexes. Overall, 80% of outbreaks in the derivation set were correctly classified by the model's first choice, and 98% were classified by the model's first or second choice. Odds ratios for dairy versus beef outbreaks were significant and greater than 1.0 for number of cases, private setting, spring, and serogroup (Table 3). Odds ratios for leafy versus beef outbreaks were significant and greater than 1.0 for percentage female, number of cases, multistate outbreak, and serogroup.

There were 49 outbreaks in the withheld validation set scored to the model after 12 outbreaks with incomplete data had been excluded. The model correctly classified 20 of 29 beef outbreaks (sensitivity of 69%, specificity 60%), 3 of 10 dairy outbreaks (sensitivity 30%, specificity 90%), and 7 of 10 vegetable outbreaks (sensitivity 70%, specificity 82%). Distribution of predicted probabilities is shown in Figure 2b. Overall, 61% of outbreaks were correctly classified by the model's first choice, and 96% of outbreaks were classified by the model's first or second choice. The model was also scored to the entire dataset ($n=206$) and subsets of the entire dataset (30%, 60%, 90%). The sensitivity and specificity in the data subsets was comparable to the original derivation set and was not impacted by sample size.

Discussion

This study systematically identified factors predictive of food sources in STEC outbreaks reported to the CDC. Factors predictive of three major food sources (beef, vegetables, and dairy) included case demographic and outbreak characteristics. These factors were used to build and validate a prediction model to estimate the probability of each major food source for a given STEC outbreak. This study provides the groundwork for a predictive tool that investigators can use during hypothesis generation in foodborne outbreak investigations and could be applied to other foodborne pathogens.

Gender and age distributions differed between food sources in STEC outbreaks. Vegetable STEC outbreaks had the highest median percentage female, whereas there was no significant gender difference in beef and dairy outbreaks.

Table 1. Shiga Toxin–Producing *Escherichia coli* Outbreak Candidate Predictors in a Random Selection of 70% of the Dataset for Univariate Comparison Across Food Categories

Predictor	*Beef*	*Dairy*	*Vegetables*	p
Number of outbreaks (%)	86 (59)	20 (14)	39 (27)	
Total cases (%)	1531 (39)	488 (12)	1059 (27)	
Gender[a]				
% Male	50 (30–67)	50 (34–67)	36 (26–44)	0.01
% Female	50 (34–70)	50 (33–67)	64 (56–74)	0.01
Missing[b]	19 (22)	3 (15)	2 (5)	0.06
Age (years)[a]				
% <5	0 (0–13)	22 (0–33)	0 (0–4)	0.01
% 5–19	35 (12–61)	50 (50–67)	20 (15–33)	0.02
% 20–49	21 (0–33)	17 (0–33)	44 (30–54)	<0.01
% >50	14 (0–33)	0 (0–0)	19 (9–40)	<0.01
% Unknown	0 (0–0)	0 (0–0)	0 (0–0)	0.54
Missing[b]	26 (30)	3 (15)	8 (21)	0.26
% Hospitalized[a]	33 (18–59)	33 (0–56)	29 (18–60)	0.66
Missing	17 (20)	1 (5)	4 (10)	0.15
Number of cases[a]	9 (3–17)	5 (3–9)	18 (10–33)	<0.01
Exposure setting[c]				<0.01
Private	45 (52)	14 (70)	5 (13)	
Non-private	33 (38)	4 (20)	32 (82)	
Missing	8 (9)	2 (10)	2 (5)	
Multistate outbreak[c]	31 (36)	2 (10)	22 (56)	<0.01
Season[d]				0.11
Fall	17 (20)	7 (35)	17 (44)	
Winter	7 (8)	1 (5)	5 (13)	
Spring	28 (33)	6 (30)	8 (21)	
Summer	34 (40)	6 (30)	9 (23)	
Duration (days)[a]	10 (2–38)	14 (9–31)	14 (10–19)	0.25
Missing[b]	24 (28)	3 (15)	6 (15)	0.20
Non-O157:H7 Serogroup[c]	3 (4)	5 (25)	7 (18)	<0.01

[a] Data presented as median (IQR), *p*-value from Kruskal–Wallis.
[b] Incomplete data assessed independently of continuous variables, presented as proportion (%), *p*-value from chi-square.
[c] Presented as proportion (%), *p*-value from chi-square.
[d] Presented as proportion (%), Fisher's Exact.

Table 2. Sensitivity and Specificity of Predicted Food Source for Derivation and Validation Datasets

	Observed outcome			
Predicted outcome	*Beef*	*Dairy*	*Vegetables*	*Total*
Derivation				
Beef	**56**	7	8	71
Dairy	2	**9**	0	11
Vegetables	6	0	**28**	34
Total	64	16	36	116
Sensitivity	0.88	0.56	0.78	
Specificity	0.71	0.98	0.93	
Validation				
Beef	**20**	6	2	28
Dairy	3	**3**	1	7
Vegetables	6	1	**7**	14
Total	29	10	10	49
Sensitivity	0.69	0.30	0.70	
Specificity	0.60	0.90	0.82	

Bold values indicate the number of outbreaks correctly classified by the prediction model for each food source.

Vegetable STEC outbreaks also had the lowest median percentage of children and adolescents, whereas the percentage of children and adolescents was the highest for dairy outbreaks. Food consumption surveys have found similar patterns (Mun and Krebs-Smith, 1997; Shiferaw *et al.,* 2000; Patil *et al.*, 2005; Samuel *et al.*, 2007; Shiferaw *et al.*, 2012). For example, the FoodNet Population Survey found that women consume more fruits and vegetables than men and that men consume more meat and poultry; however, there was no gender difference in ground beef consumption (Shiferaw *et al.*, 2012). Other studies reported that younger children consumed less meat and vegetables than adults (Mun and Krebs-Smith, 1997).

Seasonal variation by food source was another key predictive factor noted in this study, with 40% of beef outbreaks occurring in summer and 44% of vegetable outbreaks occurring in fall. Overall, this study found that STEC outbreaks occurred in warmer seasons, with few outbreaks occurring in winter (8% for beef, 5% dairy, 13% vegetables). The incidence of foodborne illness is known to vary seasonally, generally increasing in warmer months (Lal *et al.*, 2012). Bacterial pathogens are sensitive to temperature and moisture during warmer seasons, which increase pathogen survival and proliferation (Money *et al.*, 2010). Studies have found the prevalence of STEC in cattle peaks in summer months

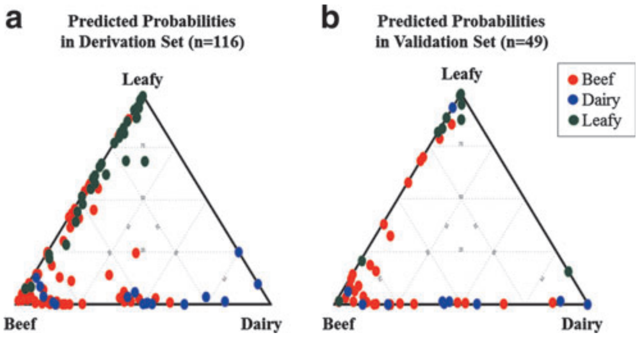

FIG. 2. Predicted probability of beef, dairy, and leafy vegetable food sources as a function of actual food sources. The three-way predicted probabilities of beef, dairy, and leafy vegetable food sources in the derivation set **(a)** and the validation set **(b)** such that the total predicted probability for each outbreak sums to 100%. The mean predicted probability in the derivation set was 0.70 for beef (standard deviation: ±0.21); 0.45 ± 0.28 for dairy; and 0.68 ± 0.27 for leafy vegetable outbreaks. Eighty percent of outbreaks in the derivation set were correctly classified by the model's first choice, and 98% were correctly classified by the model's first or second choice. The mean predicted in the validation set was 0.65 ± 0.27 for beef; 0.41 ± 0.34 for dairy; and 0.67 ± 0.38 for leafy vegetable outbreaks. Sixty percent of outbreaks in the validation set were correctly classified by the model's first choice, and 96% were correctly classified by the model's first or second choice.

(Edrington *et al.*, 2006; Ferens and Hovde, 2011). Food consumption patterns may differ as well due to seasonal events and holidays, availability, and cost (Ravel *et al.*, 2010; Wilson, 2015; Stelmach-Mardas *et al.*, 2016).

Vegetable STEC outbreaks had the highest average number of cases per outbreak. This correlated with the fact that a higher proportion of vegetable outbreaks are multistate, involving large numbers of outbreak-associated cases. Conversely, many dairy outbreaks are due to unpasteurized milk and tend to be local, because unpasteurized dairy products cannot be sold at retail stores, and, therefore, have limited distribution (Angulo *et al.*, 2009; Newkirk *et al.*, 2011). Localized outbreaks with a single exposure site are typically contaminated during preparation and served at a single setting (Murphree *et al.*, 2012). Conversely, multistate outbreaks have contamination points earlier in the production chain, and foods that are prone to contamination at these stages may differ from foods that are contaminated at a single setting (Nguyen *et al.*, 2015).

In the multivariable model, six factors (percentage female, number of cases, multistate outbreak, exposure setting, season, and serogroup) were predictive of food sources. Model classification accuracy was the highest for beef, moderate for vegetable, and poor for dairy. Misclassification of dairy and vegetable outbreaks almost exclusively classified them as beef, with limited misclassification between dairy and vegetable outbreaks. This was likely driven by unequal sample sizes. There were more beef outbreaks; whereas dairy outbreaks were much fewer and, therefore, contributed less information to the model estimates. Each food category represented an aggregated group of food items. For example, dairy outbreaks included food sources that were pasteurized or unpasteurized, and they included products such as ice cream, cheese, and fluid milk. Each of these items may have unique profiles that do not necessarily justify aggregation, therefore making prediction difficult.

The results indicate that the model could be used the most effectively as a rule-out tool. Most misclassified outbreaks had a similar predicted probability for the model's first and second choice. For example, if a dairy outbreak was predicted to be related to beef, the predicted probability could be 48% beef, 47% dairy, and 5% vegetable. Although the model misclassified outbreaks, 98% were correctly classified by the second choice in the derivation set. During hypothesis generation, an investigator would consider the first or second food sources, but could be more confident about excluding the third food source.

The results of this study could be translated into a tool for public health professionals to use during outbreak investigation. In the early stages of an outbreak, descriptive profiles based on univariate results could be used when data are only available for a limited number of variables. For example, if an STEC outbreak occurs in fall, has a higher percentage female, aged 20–49, vegetables may be highly suspected. Public health professionals could use a tool adapted from the multivariable model at the point of food hypothesis generation after non-foodborne routes of transmission have been ruled

TABLE 3. ODDS RATIOS OF FOOD SOURCE OUTCOME BY PREDICTOR FROM THE MULTIVARIABLE LOGISTIC REGRESSION MODEL

	Dairy outbreaks (reference beef outbreaks)		*Vegetable (reference beef outbreaks)*	
Predictor	*OR (95% CI)*	p	*OR (95% CI)*	p
% Female[a]	0.93 (0.71–1.21)	0.56	1.58 (1.16–2.15)	<0.01
Number of cases[a]	1.75 (1.20–2.54)	<0.01	1.41 (1.02–1.96)	0.04
Private setting	22.60 (2.33–218.98)	<0.01	0.03 (0.01–0.18)	<0.01
Multistate outbreak	0.01 (0.00–0.25)	0.01	3.48 (0.86–14.05)	0.08
Season				
Fall	1.0		1.0	
Winter	0.11 (0.00–3.29)	0.20	0.30 (0.05–2.02)	0.22
Spring	1.56 (0.25–9.74)	0.64	0.27 (0.06–1.24)	0.09
Summer	0.21 (0.04–1.21)	0.08	0.18 (0.03–0.92)	0.04
Non-O157:H7	30.53 (1.97–473.74)	0.01	17.92 (1.47–218.23)	0.02

[a] Covariates presented by increments of 10%.

out. For STEC outbreaks, the investigator would enter percentage female, number of cases, multistate outbreak, exposure setting, season, and serogroup. The tool would output a predicted probability of an STEC outbreak being beef, dairy, or vegetables. The predicted probability would help direct hypothesis generation, or give additional evidence for existing hypotheses.

For this analysis, all predictors were included that improved predictive value for the model using data from finalized investigations. However, some predictors are unlikely to be available at the start of an outbreak, whereas others may change over the course of an investigation. Exposure setting is often determined by the investigation and, therefore, unlikely to be known at the time that this tool would be used. Other predictors change over the investigation. For example, the number of ill cases may increase as investigators actively find additional cases. Values for predictors may be skewed at the start of an outbreak and may not be representative of the total number of cases in an outbreak, which was used to derive the model.

In addition to use as a practical tool during outbreak investigations, the statistical model developed here could be used in foodborne illness source attribution by estimating the food source distribution in outbreaks with a previously undetermined vehicle. Attribution estimates exclude outbreaks with undetermined food vehicles (Painter *et al.*, 2013; IFSAC, 2015). By using a model to estimate the food source distribution in outbreaks with a previously undetermined vehicle, this would increase the sample size of available outbreaks as well as provide more accurate estimates for the relative contribution of food sources.

This study has several limitations in addition to those already discussed. First, age variables were excluded because of incomplete data and each age category was a separate variable, which resulted in non-convergence of the model when included. Exclusion of age, consequently, impacted classification accuracy of dairy outbreaks, which decreased considerably. Second, serogroup was included as a predictor, because it varied by food source; however, inclusion may not be justified. Non-O157, which is increasing in incidence of sporadic illnesses (Gould *et al.*, 2013b), is generally less severe, differs geographically from O157, and is associated with different modes of transmission (animal contact is more common) and different foods (Mathusa *et al.*, 2010; Gould *et al.*, 2013b; Luna-Gierke *et al.*, 2014). Third, there are many other factors that could be predictive of a food source in an outbreak investigation; however, we were limited by what is collected by the surveillance system, as well as by the completeness and consistency of reporting. Finally, an external dataset (e.g., an independent population), often used in clinical rules for validation, was unavailable.

It is recommended that future studies build on this work to apply to other pathogens, such as *Salmonella*, and refine the model to increase accuracy and generalizability. The model should be translated into a user-friendly online tool for investigators. Future work should explore methods to incorporate novel and complex foods and to further characterize outbreaks with undetermined sources. Additional data sources should be explored to supplement outbreak surveillance data and to explore additional potentially important predictors. Finally, additional work should use outbreak surveillance data to build tools for foodborne outbreak investigators.

Conclusion

This study provides evidence for the feasibility of using prior case and outbreak characteristics to predict food sources in foodborne outbreak investigations. This is the first study to demonstrate statistically that case demographics are associated with food sources in a foodborne outbreak investigation, which are commonly used by experienced investigators to generate hypotheses. In addition, this is the first study to propose a model to consider all factors simultaneously in a single prediction model. The complexity of the global food industry and modern challenges to food safety mean that outbreak investigations are increasingly important to identifying sources of foodborne illness. To combat the challenges inherent in public health investigations, analytical, data-driven tools are essential to efficient outbreak investigations and for improving food source identification.

Acknowledgments

This study was funded by the Centers for Disease Control and Prevention (CDC) through the Colorado Integrated Food Safety Center of Excellence (CoE). The authors thank the National Outbreak Reporting System (NORS) group at CDC for providing the data used in this analysis and for their consultation.

Disclosure Statement

No competing financial interests exist.

References

Angulo FJ, LeJeune JT, Rajala-Schultz PJ. Unpasteurized milk: A continued public health threat. Clin Infect Dis 2009;48: 93–100.

Barnes DE, Mehta KM, Boscardin WJ, Fortinsky RH, Palmer RM, Kirby KA, Landefeld CS. Prediction of recovery, dependence or death in elders who become disabled during hospitalization. J Gen Intern Med 2013;28:261–268.

Biesheuvel CJ, Vergouwe Y, Steyerberg EW, Grobbee DE, Moons KGM. Polytomous logistic regression analysis could be applied more often in diagnostic research. J Clin Epidemiol 2008;61:125–134.

[CDC] Centers for Disease Control and Prevention. FoodNet Population Survey. 2007. Available at: www.cdc.gov/foodnet/studies/population-surveys.html, accessed November 2015.

[CDC] Centers for Disease Prevention and Control. The National Outbreak Reporting System (NORS). About NORS. 2009. Available at: www.cdc.gov/NORS/about.html, accessed November 2015.

[CDC] Centers for Disease Control and Prevention. Foodborne Diseases Centers for Outbreak Response Enhancement (FoodCORE). 2010. Available at: www.cdc.gov/foodcore/, accessed January 2016.

[CDC] Centers for Disease Control and Prevention. Integrated Food Safety Centers of Excellence (CoE). 2012. Available at: www.cdc.gov/foodsafety/centers/, accessed January 2016.

[CDC] Centers for Disease Control and Prevention. National Hypothesis Generating Questionnaire. 2013. Available at: www.cdc.gov/foodsafety/outbreaks/surveillance-reporting/investigation-toolkit.html, accessed January 2016.

[CIFOR] Council to Improve Foodborne Outbreak Response. *Guidelines for Foodorne Disease Outbreak Response*, 2nd ed.

Atlanta: Council of State and Territorial Epidemiologists, 2014.

Crim SM, Iwamoto M, Huang JY, Griffin PM, Gilliss D, Cronquist AB, Cartter M, Tobin-D'Angelo M, Blythe D, Smith K, Lathrop S, Zansky S, Cieslak PR, Dunn J, Holt KG, Lance S, Tauxe R, Henao OL; Centers for Disease Control and Prevention. Incidence and trends of infection with pathogens transmitted commonly through food—Foodborne Diseases Active Surveillance Network, 10 U.S. sites, 2006–2013. MMWR Morb Mortal Wkly Rep 2014;63:328–332.

Domínguez A, Broner S, Torner N, Martínez A, Jansà JM, Alvarez J, Barrabeig I, Cayla J, Godoy P, Minguell S, Camps N, Sala MR; Working Group for the Study of Outbreaks of Acute Gastroenteritis in Catalonia. Utility of clinical-epidemiological profiles in outbreaks of foodborne disease, Catalonia, 2002 through 2006. J Food Prot 2010;73:125–131.

Edrington TS, Callaway TR, Ives SE, Engler MJ, Looper ML, Anderson RC, Nisbet DJ. Seasonal shedding of *Escherichia coli* O157:H7 in ruminants: A new hypothesis. Foodborne Pathog Dis 2006;3:413–421.

Ferens WA, Hovde CJ. *Escherichia coli* O157:H7: Animal reservoir and sources of human infection. Foodborne Pathog Dis 2011;8:465–487.

Friendly M. Visualizing Categorical Data: Triplot. 2009. Available at: www.datavis.ca/sasmac/triplot.html, accessed August 2015.

Ge WJ, Mirea L, Yang J, Bassil KL, Lee SK, Shah PS; Canadian Neonatal Nework. Prediction of neonatal outcomes in extremely preterm neonates. Pediatrics 2013;132:876–885.

Gould LH, Mody RK, Ong KL, Clogher P, Cronquist AB, Garman KN, Lathrop S, Medus C, Spina NL, Webb TH, White PL, Wymore K, Gierke RE, Mahon BE, Griffin PM. Increased recognition of non-O157 Shiga toxin–producing *Escherichia coli* infections in the United States during 2000–2010: Epidemiologic features and comparison with *E. coli* O157 infections. Foodborne Pathog Dis 2013a;10:453–640.

Gould LH, Walsh KA, Vieir AR, Herman K, Williams IT, Hall AJ, Cole D. Surveillance for foodborne disease outbreaks—United States, 1998–2008. MMWR Morb Mortal Wkly Rep 2013b;62:1–34.

Graham DJ, Midgley NG. Graphical representation of particle shape using triangular diagrams: An excel spreadsheet method. Earth Surf Process Landf 2000;25:1473–1477.

Hall JA, Goulding JS, Bean NH, Tauxe RV, Hedberg CW. Epidemiologic profiling: Evaluating foodborne outbreaks for which no pathogen was isolated by routine laboratory testing: United States, 1982–1989. Epidemiol Infect 2001;127:381–387.

Hedberg CW, Palazzi-Churas KL, Radke VJ, Selman CA, Tauxe RV. The use of clinical profiles in the investigation of foodborne outbreaks in restaurants: United States, 1982–1997. Epidemiol Infect 2008;136:65–72.

[IFSAC] Interagency Food Safety Analytics Collaboration. Food Categorization Scheme. 2013. Available at: www.cdc.gov/foodsafety/ifsac/projects/food-categorization-scheme.html, accessed November 2015.

[IFSAC] Interagency Food Safety Analytics Collaboration. Foodborne Illness Source Attribution Estimates for *Salmonella, Eschericia coli* O157 (*E. coli* 0157), *Listeria monocytogenes (Lm),* and *Campylobacter* using Outbreak Surveillance Data. 2015. Available at: www.cdc.gov/foodsafety/pdfs/ifsac-project-report-508c.pdf, accessed January 2016.

Lal A, Hales S, French N, Baker MG. Seasonality in human zoonotic enteric diseases: A systematic review. PLoS One 2012;7:e31883.

Luna-Gierke RE, Griffin PM, Gould LH, Herman K, Bopp CA, Strockbine N, Mody RK. Outbreaks of non-O157 Shiga toxin–producing *Escherichia coli* infection: USA. Epidemiol Infect 2014;142:2270–2280.

Mathusa EC, Chen Y, Enache E, Hontz L. Non-O157 Shiga toxin–producing *Escherichia coli* in foods. J Food Prot 2010; 9:1596–1773.

Money P, Kelly AF, Gould SWJ, Denholm-Price J, Threlfall EJ, Fielder MD. Cattle, weather and water: Mapping *Escherichia coli* O157:H7 infections in humans in England and Scotland. Environ Microbiol 2010;12:2633–2644.

Mun KA, Krebs-Smith SM. Food intakes of US children and adolescents compared with recommendations. Pediatrics 1997; 100:323–329.

Murphree R, Garman K, Phan Q, Everstine K, Gould LH, Jones TF. Characteristics of foodborne disease outbreak investigations conducted by foodborne diseases active surveillance network (Foodnet) sites, 2003–2008. Clin Infect Dis 2012;54: S498–S503.

Newkirk R, Hedberg C, Bender J. Establishing a milkborne disease outbreak profile: Potential food defense implications. Foodborne Pathog Dis 2011;8:433–437.

Nguyen VD, Bennett SD, Mungai E, Gieraltowski L, Hise K, Gould LH. Increase in multistate foodborne disease outbreaks—United States, 1973–2010. Foodborne Pathog Dis 2015;12:867–872.

[OPHD] Oregon Public Health Division. Binomial Probability Worksheet. 2010. Available at: https://public.health.oregon.gov/DiseasesConditions/CommunicableDisease/Outbreaks/Gastroenteritis/Pages/Outbreak-Investigation-Tools.aspx#binomial, accessed January 2016.

Painter JA, Ayers T, Woodruff R, Blanton E, Perez N, Hoekstra RM, Griffin PM, Braden C. Recipes for foodborne outbreaks: A scheme for categorizing and grouping implicated foods. Foodborne Pathog Dis 2009;6:1259–1264.

Painter JA, Hoekstra RM, Ayers T, Tauxe RV, Braden CR, Angulo FJ, Griffin PM. Attribution of foodborne illnesses, hospitalizations, and deaths to food commodities by using outbreak data, United States, 1998–2008. Emerg Infect Dis 2013;19:407–415.

Patil SR, Cates S, Morales R. Consumer food safety knowledge, practices, and demographic differences: Findings from a meta-analysis. J Food Prot 2005;11:1884–1894.

Ravel A, Smolina E, Sargeant JM, Cook A, Marshall B, Fleury MD, Pollari F. Seasonality in human salmonellosis: Assessment of human activities and chicken contamination as driving factors. Foodborne Pathog Dis 2010;7:785–794.

Reingold AL. Outbreak Investigations—a perspective. Emerg Infect Dis 1998;4:21–27.

Samuel MC, Vugia DJ, Koehler KM, Marcus R, Deneen V, Damaske B, Shiferaw B, Hadler J, Henao OL, Angulo FJ. Consumption of risky foods among adults at high risk for severe foodborne diseases: Room for improved targeted prevention messages. J Food Safety 2007;27:219–232.

Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, Jones JL, Griffin PM. Foodborne illness acquired in the United states-major pathogens. Emerg Infect Dis 2011;17:7–15.

Shiferaw B, Verrill L, Booth H, Zansky SM, Norton DM, Crim S, Henao OL. Sex-based differences in food consumption: Foodborne diseases active surveillance network (FoodNet) population survey, 2006–2007. Clin Infect Dis 2012;54(Suppl 5): S453–S457.

Shiferaw B, Yang S, Cieslak P, Vugia D, Marcus R, Koehler J, Deneen V, Angulo F. Prevalence of high-risk food consumption and food-handling practices among adults: A multistate survey, 1996 to 1997. The Foodnet Working Group. J Food Prot 2000;63:1538–1543.

Stelmach-Mardas M, Kleiser C, Uzhova I, Peñalvo JL, La Torre G, Palys W, Lajko D, Nimptsch K, Suwalska A, Linseisen J, Saulle R, Colamesta V, Boeing H. Seasonality of food groups and total energy intake: A systematic review and meta-analysis. Eur J Clin Nutr 2016;70:700–708.

Tauxe RV. Emerging foodborne diseases: An evolving public health challenge. Emerg Infect Dis 1997;3:425–434.

Tauxe RV. Surveillance and investigation of foodborne diseases; roles for public health in meeting objectives for food safety. Food Control 2002;13:363–369.

Turcios RM, Widdowson MA, Sulka AC, Mead PS, Glass RI. Reevaluation of epidemiological criteria for identifying outbreaks of acute gastroenteritis due to norovirus: United States, 1998–2000. Clin Infect Dis 2006;42:964–969.

Wilson E. Foodborne illness and seasonality related to mobile food sources at festivals and group gatherings in the state of Georgia. J Environ Health 2015;77:8–11.

Address correspondence to:
Alice White, MS
Department of Epidemiology
Colorado School of Public Health
University of Colorado Denver
Anschutz Medical Campus
12477 E. 19th Avenue
Mail Stop B119
Aurora, CO 80045

E-mail: alice.white@ucdenver.edu