Recency-Weighted Statistical Modeling Approach to Attribute Illnesses Caused by 4 Pathogens to Food Sources Using Outbreak Data, United States

Michael B. Batz, LaTonia C. Richardson, Michael C. Bazaco, Cary Chen Parker, Stuart J. Chirtel, Dana Cole,[1] Neal J. Golden, Patricia M. Griffin, Weidong Gu,[2] Susan K. Schmitt,[3] Beverly J. Wolpert, Joanna S. Zablotsky Kufel, R. Michael Hoekstra[4]

Foodborne illness source attribution is foundational to a risk-based food safety system. We describe a method for attributing US foodborne illnesses caused by nontyphoidal *Salmonella enterica*, *Escherichia coli* O157, *Listeria monocytogenes,* and *Campylobacter* to 17 food categories using statistical modeling of outbreak data. This method adjusts for epidemiologic factors associated with outbreak size, down-weights older outbreaks, and estimates credibility intervals. On the basis of 952 reported outbreaks and 32,802 illnesses during 1998–2012, we attribute 77% of foodborne *Salmonella* illnesses to 7 food categories (seeded vegetables, eggs, chicken, other produce, pork, beef, and fruits), 82% of *E. coli* O157 illnesses to beef and vegetable row crops, 81% of *L. monocytogenes* illnesses to fruits and dairy, and 74% of *Campylobacter* illnesses to dairy and chicken. However, because *Campylobacter* outbreaks probably overrepresent dairy as a source of nonoutbreak campylobacteriosis, we caution against using these *Campylobacter* attribution estimates without further adjustment.

Each year in the United States, nontyphoidal *Salmonella*, *Escherichia coli* O157, *Listeria monocytogenes,* and *Campylobacter* cause >2 million estimated foodborne illnesses, 31,000 hospitalizations, and 700 deaths (*1*), representing an estimated $9–$11 billion in impacts to human health (*2*,*3*). Estimating the percentage of these illnesses attributable to the consumption of specific foods (i.e., foodborne illness source attribution) is foundational to a risk-based national food safety system (*4*). Such estimates can inform strategic planning, priority setting, risk assessments, economic analyses, and evaluations of the impacts of regulations and interventions (*5*).

Numerous studies in the United States and worldwide have estimated source attribution on the basis of aggregated foodborne outbreak data (*6*–*12*). For the United States, the Centers for Disease Control and Prevention (CDC) previously estimated the number of domestically acquired foodborne illnesses, hospitalizations, and deaths attributable to food categories based on analysis of outbreaks during 1998–2008 (*13*).

Through the Interagency Food Safety Analytics Collaboration (IFSAC), CDC, the US Food and Drug Administration, and the US Department of Agriculture Food Safety and Inspection Service work in partnership to develop improved source attribution estimates through multiple interconnected projects (*14*,*15*). This study reflects a tri-agency effort to update and harmonize estimates for the United States for nontyphoidal *Salmonella*, *E. coli* O157, *L. monocytogenes,* and *Campylobacter* using data from outbreaks that occurred during 1998–2012.

Author affiliations: US Food and Drug Administration, College Park, Maryland, USA (M.B. Batz, M.C. Bazaco, C. Chen Parker, S.J. Chirtel, B.J. Wolpert); Centers for Disease Control and Prevention, Atlanta, Georgia, USA (L.C. Richardson, D. Cole, P.M. Griffin, W. Gu, R.M. Hoekstra); US Department of Agriculture, Washington, DC, USA (N.J. Golden, S.K. Schmitt, J.S. Zablotsky Kufel)

DOI: https://doi.org/10.3201/eid2701.203832

[1]Current affiliation: US Department of Agriculture, Fort Collins, Colorado, USA.

[2]Current affiliation: US Department of Defense, Bethesda, Maryland, USA.

[3]Current affiliation: US Department of Veterans Affairs, Palo Alto, California, USA.

[4]Retired.

IFSAC's approach addresses some of the limitations of prior studies. We describe this method here. We use statistical modeling to mitigate the influence of large outbreaks that might bias estimates, and we incorporate epidemiologic factors relevant to outbreak size. We weight recent outbreaks more heavily than older ones and quantify uncertainty by estimating credibility intervals around estimates. We also use an updated food categorization scheme that better meets the needs of the regulatory agencies.

Methods

Data Sources

CDC's Foodborne Disease Outbreak Surveillance System (FDOSS) collects standardized reports submitted by state, local, and territorial health departments on foodborne disease outbreaks. In FDOSS, outbreaks are defined as the occurrence of ≥2 cases of a similar illness resulting from the ingestion of a common food (*16*). We extracted data from FDOSS on reported foodborne outbreaks caused by nontyphoidal *Salmonella enterica*, *E. coli* O157 (*E. coli* O157:H7 and *E. coli* O157:NM), *L. monocytogenes,* and *Campylobacter* spp., in which the first illness occurred in a US state or the District of Columbia during 1998–2012. We extracted data on December 18, 2013. Analysis was conducted by using SAS 9.3, JMP Pro (SAS Institute, https://www.sas.com), and R (R Foundation for Statistical Computing, https://www.r-project.org).

We included only outbreaks with a single causal pathogen and for which implicated foods could be assigned to a single food category because those outbreaks have the clearest information. We excluded outbreaks caused by multiple pathogens or for which no food or ingredient was implicated, including outbreaks with a complex food vehicle (i.e., consisting of ingredients belonging to >1 food category) for which the implicated ingredient was not determined. A previously published method (*13*) for assigning the food category for complex food outbreaks could not be applied to more recent data without substantial revision.

We excluded outbreaks for which implicated foods came from >1 food category (i.e., multiple foods). For example, an outbreak for which apples and cantaloupe were both implicated would be included because both fall into the fruits category, but an outbreak for which apples and cheese were both implicated would be excluded.

By using a hierarchical scheme of 22 food categories, we assigned outbreaks to a single category on the basis of implicated foods or ingredients (*17*). Because of sparse data, outbreaks in 8 food categories were aggregated into 3 combined categories: other meat and poultry (other meat, other poultry); other seafood (shellfish, other aquatic animals); and other produce (fungi, herbs, root-underground, nuts-seeds), resulting in 17 food categories for our analysis (Appendix Figure 1, https://wwwnc.cdc.gov/EID/article/27/1/20-3832-App1.pdf).

In FDOSS, an outbreak must have ≥2 ill persons (*16*). For *Salmonella*, *E. coli* O157, and *Campylobacter*, outbreaks with confirmed etiology are defined as those in which the outbreak strain was isolated from ≥2 patients or from epidemiologically implicated food; confirmed outbreaks of *L. monocytogenes* infections must have 1 person with the outbreak strain isolated from a normally sterile site (*18*). (Cases of listeriosis can also be diagnosed based on symptoms and culture of pregnancy-associated products of conception, which are not sterile.) The etiology of an outbreak not meeting these conditions is considered to be suspected. We found no statistically significant differences in outbreak size or foods implicated between outbreaks with confirmed and those with suspected status, and therefore included outbreaks with suspected etiology in the analysis (Appendix). The final dataset used for exploratory analysis and estimates of sources included 952 outbreaks assigned to 17 food categories (Table 1).

Exploratory Analysis

We focused exploratory analysis on factors influencing outbreak size. We used the total number of reported illnesses as the measure of outbreak size. Whereas most outbreaks are small, some are very large. For example, of the 4,732 reported *Campylobacter* outbreak illnesses during the entire 15-year period, more than one third (1,644) were from a single outbreak. Large outbreaks might not be representative of the sources of sporadic illnesses and might overly influence estimates of food sources (*13*).

Untransformed outbreak size is skewed and varies across pathogens (Figure 1). Log transformation of outbreak size results in distributions that are more symmetric and normally distributed, although considerable variation remains within and across pathogens (Figure 1). We therefore used log-transformed outbreak size in statistical modeling.

FDOSS data include epidemiologic factors that might relate to the size and scope of an outbreak, including pathogen, number of states in which outbreak exposures occurred, implicated foods and ingredients, and the type of location in which food was prepared (e.g., restaurant or private home). We explored the relationships between outbreak size and these variables.

Table 1. Number of outbreaks and outbreak illnesses caused by a single pathogen and due to a single food category for *Salmonella, Escherichia coli* O157, *Listeria monocytogenes,* and *Campylobacter*, Foodborne Disease Outbreak Surveillance System, United States, 1998–2012*

Food category	Nontyphoidal *Salmonella* spp.	*E. coli* O157	*Listeria monocytegenes*	*Campylobacter* spp.	Total
Beef	47 (1,473)	97 (1,813)	1 (4)	2 (5)	147 (3,295)
Pork	51 (1,098)	0	2 (11)	1 (27)	54 (1,136)
Chicken	114 (2,648)	1 (36)	1 (3)	24 (230)	140 (2,917)
Turkey	49 (1,308)	1 (2)	4 (124)	5 (44)	59 (1,478)
Other meat or poultry	6 (84)	2 (9)	0	2 (6)	10 (99)
Game	2 (8)	4 (18)	0	1 (2)	7 (28)
Dairy	24 (793)	18 (399)	12 (124)	106 (3,395)	160 (4,711)
Eggs	140 (5,245)	0	0	0	140 (5,245)
Fish	12 (286)	0	0	1 (3)	13 (289)
Other seafood	4 (36)	0	0	5 (344)	9 (380)
Grains, beans	7 (268)	0	0	0	7 (268)
Oils, sugars	0	0	0	1 (3)	1 (3)
Fruits	46 (2,510)	11 (893)	1 (147)	2 (29)	60 (3,579)
Seeded vegetables	34 (4,001)	0	0	3 (136)	37 (4,137)
Sprouts	33 (1,266)	6 (55)	2 (26)	0	41 (1,347)
Vegetable row crops	10 (412)	29 (1,029)	1 (10)	7 (372)	47 (1,823)
Other produce	18 (1,923)	1 (8)	0	1 (136)	20 (2,067)
Total	597 (23,359)	170 (4,262)	24 (449)	161 (4,732)	952 (32,802)

*Number of outbreak-associated illnesses in parentheses. Nontyphoidal *Salmonella* is divided into *S. enterica* serovar Enteritidis and other serovars (Appendix Table 2, https://wwwnc.cdc.gov/EID/article/27/1/20-3832-App1.pdf).

Distinct differences in distributions of outbreak size can be observed by pathogen and 3 categorical variables: food category, type of food preparation location, and whether exposures occurred in multiple states or a single state (Figure 2). For example, the mean size of multistate outbreaks is larger than single-state outbreaks for *Salmonella*, *E. coli* O157, and *L. monocytogenes*. Differences in grouped means can be observed for all 3 categorical variables for *L. monocytogenes* despite the small number of outbreaks.

Statistical Modeling

Whereas prior studies calculated attribution proportions on the basis of observed counts of reported outbreak events or outbreak illnesses (*6*,*13*), in this study we developed a model-based approach to estimate the number of outbreak illnesses for attribution. This approach mitigates the impact of large outbreaks and enables the incorporation of epidemiologic factors beyond pathogen and food category.

After considering several approaches, we chose analysis of variance (ANOVA) of log-transformed outbreak sizes as the modeling framework, partly based on simplicity and interpretability (Appendix). For each pathogen, we developed a model to estimate the log-transformed number of illnesses based on the 3 factors shown to be associated with outbreak size: food category, type of food preparation location, and whether exposures occurred in multiple states or a single state. Each of these factors was found through

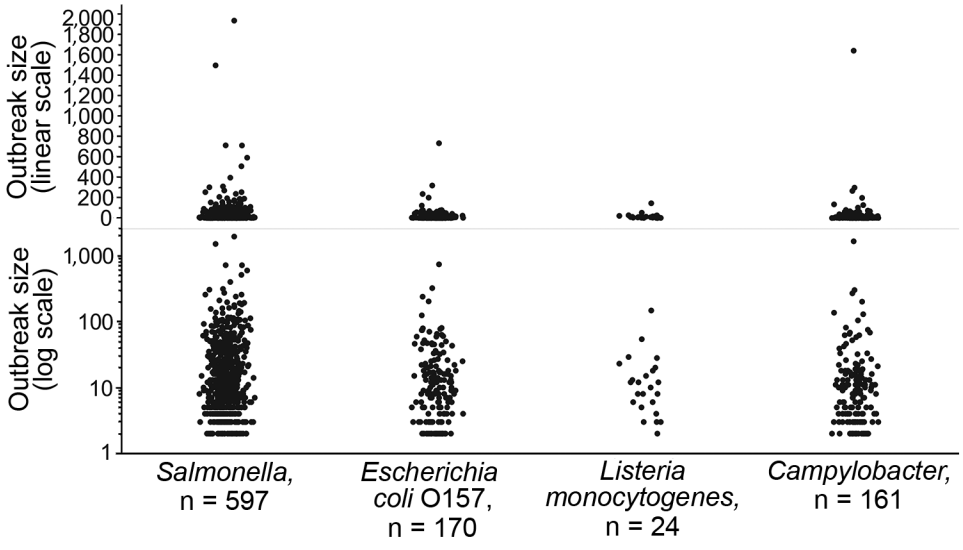


Figure 1. Number of reported illnesses for foodborne disease outbreaks caused by a single pathogen and attributable to a single food category, using linear and log scales, for *Salmonella*, *Escherichia coli* O157, *Listeria monocytogenes*, and *Campylobacter*, Foodborne Disease Outbreak Surveillance System, United States, 1998–2012.

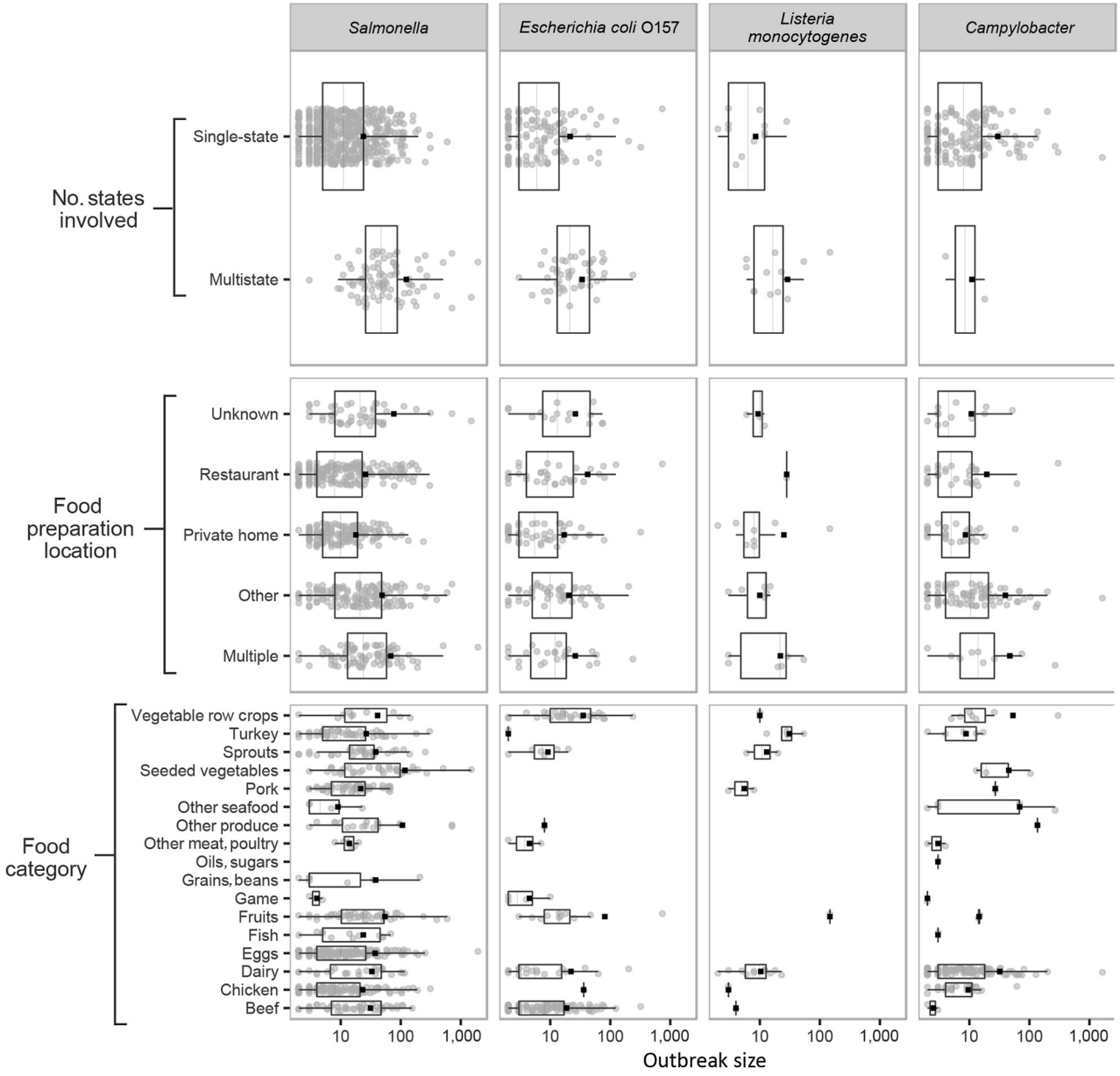


Figure 2. Number of reported illnesses (log scale) for foodborne disease outbreaks caused by a single pathogen and attributable to a single food category, for 3 outbreak characteristics, for *Salmonella, Escherichia coli O157, Listeria monocytogenes, and Campylobacter*, Foodborne Disease Outbreak Surveillance System, United States, 1998–2012. Each panel displays outbreak size for a given pathogen, grouped by 1 of 3 categorical variables. Each includes a scatterplot of individual outbreaks (indicated by solid circles), the mean (indicated by solid squares), and a boxplot showing median, interquartile range, and minimum and maximum values inside the inner and outer fences (1.5 interquartile range).

1-way ANOVA to be a statistically significant ($p<0.05$) predictor of outbreak size for ≥1 pathogens. Although not all 3 factors were significant for all pathogens, we included them to maintain uniformity across the analysis. We explored serotype-specific ANOVA models for *Salmonella*, but for most serotypes these models did not find different distributions of outbreaks across food categories or meaningful differences in outbreak size across the other factors. The exception was serotype Enteritidis, outbreaks of which did display differences from other serotypes. Therefore, we developed 2 distinct *Salmonella* ANOVA models: 1 for Enteritidis and 1 for all other serotypes.

Each of the 5 pathogen-specific models estimates the log-transformed number of illnesses for each reported outbreak on the basis of that outbreak's characteristics as defined by the categorical variables. We then back-transformed the model-estimated numbers of illnesses

(*e* raised to the transformed values) and summed the 2 sets of *Salmonella* estimates. Additional information on model selection and fit is presented in the Appendix.

As expected, ANOVA models reduce variation in outbreak size and the influence of very large outbreaks. This effect is shown in Appendix Figure 3, which compares the number of reported illnesses with the number of model-estimated illnesses. The figure also shows the wide variation in the number of outbreaks for different pathogen–food category pairs.

Recency Weighting

Because of changes over time in food consumption patterns, food production and processing practices, food safety activities, regulatory interventions, and other factors, recent outbreaks are probably more representative of current foodborne illness attribution than older outbreaks. We explored estimating attribution on the basis of only 3, 5, or 7 years of the most recent outbreaks, but data sparseness and high year-to-year variability, particularly in food categories for which outbreaks were not reported every year, led to instability and more statistical uncertainty in attribution estimates when older outbreaks were excluded (Appendix). Therefore, we included older outbreaks but down-weighted them on the basis of recency. Outbreaks older than 5 years were multiplied by an exponential decay function, an approach long used in many fields to down-weight older data in forecasting and time-series models, including public health surveillance (*19*,*20*). This approach is flexible to inclusion of additional years of data; as the number of years of data increases, the earliest years have less and less weight.

The multiplicative recency-weighting factor w for an outbreak in year y is defined as a function of decay parameter a and the most recent year of data Y:

$$w_y = \begin{cases} a^{(Y-5)-y}, y < Y-5 \\ 1, y \geq Y-5 \end{cases}$$

We used a decay parameter a of 5/7 (0.7142). This factor resulted in outbreaks occurring during 2008–2012 providing 67% of the overall information, with 28% from outbreaks occurring during 2003–2007, and 5% outbreaks occurring during 1998–2002 (Appendix Table 4).

Calculating Attribution Percentages

For pathogen p and food category c, the attribution percentage AP_{pc} is calculated by dividing the sum of recency-weighted model-estimated illnesses of that pathogen–food category pair across all years by the sum of recency-weighted model-estimated illnesses for all food categories associated with that pathogen for all years. The estimated attribution percentage AP_{pc} is defined as:

$$AP_{pc} = \frac{\sum_y \sum_i w_y \times MEI\ (o_{pcyi})}{\sum_y \sum_c \sum_i w_y \times MEI\ (o_{pcyi})} \times 100$$

where $MEI\ (o_{pcyi})$ is the number of model-estimated illnesses for a specific outbreak o_i and i is the instance in the set of outbreaks associated with a pathogen–food pair occurring in a given year. To estimate 90% credibility intervals for each AP_{pc}, Bayesian bootstrap resampling (10,000 per pathogen–food category pair) was performed on the weighted model estimates (*21*,*22*). The 5th and 95th percentiles of the bootstrap distributions were used as the lower and upper bounds for the credibility intervals.

We conducted sensitivity analyses to examine the impacts of data selection and modeling choices on estimates. These analyses included comparing our attribution estimates to those based on reported outbreak illnesses and log-transformed illnesses, evaluating the impact of alternate ANOVA model specifications, examining the impact of recency-weighting choices and approaches, and evaluating the impact of particularly large and influential outbreaks.

Results

We extracted data on 2,732 US outbreaks caused by nontyphoidal *Salmonella*, *E. coli* O157, *L. monocytogenes*, and *Campylobacter* that occurred during 1998–2012. We excluded 77 outbreaks because they were caused by multiple pathogens, excluded an additional 1,014 because they did not have an identified food vehicle, and excluded an additional 689 that could not be assigned to a single food category (Appendix Figure 2). These exclusions resulted in a dataset with 952 outbreaks (35% of 2,732), each caused by a single pathogen and assignable to 1 of 17 food categories (Table 1).

The final estimates (Table 2; Figure 3) attributed *Salmonella* illnesses more broadly than other pathogens, with nonzero estimates for 16 categories; of those, 4 food categories had estimated percentages >10%: seeded vegetables (e.g., tomatoes), eggs, fruits, and chicken. Cumulatively, the top 7 food categories accounted for 77% of illnesses. Credibility intervals for *Salmonella* were largely overlapping but comparatively narrow, attributable in part to the high number of *Salmonella* outbreaks in the analysis. In contrast, 82% of illnesses caused by *E. coli* O157 were attributed to only 2 food categories, beef and vegetable row crops (e.g., leafy greens). Only 2 other food categories, dairy and fruits, had estimated attribution percentages >1%. Similarly, 81% of illnesses caused by *L. monocytogenes* were attributed to 2 food categories, dairy and fruits.

Table 2. Estimated percentages of foodborne illnesses attributed to 17 food categories and 90% credibility intervals for *Salmonella*, *Escherichia coli* O157, *Listeria monocytogenes*, and *Campylobacter*, based on analysis of single pathogen, single food category outbreaks, Foodborne Disease Outbreak Surveillance System, 1998–2012*

Food category	% (90% credibility interval)			
	Salmonella	*E. coli* O157	*L. monocytogenes*	*Campylobacter*
Land animals				
Beef	9 (6–13)	46 (36–55)	0 (0–1)	1 (<1–1)
Pork	8 (6–10)	–	2 (<1–8)	3 (<1–8)
Chicken	10 (7–13)	0 (0–1)	0 (0–2)	8 (5–12)
Turkey	7 (5–10)	0 (0–<1)	6 (2–16)	2 (1–4)
Other meat or poultry	0 (<1–1)	0 (0–1)	–	1 (<1–1)
Game	0 (0–<1)	1 (<1–3)	–	0 (0–<1)
Dairy	3 (1–5)	9 (5–14)	31 (12–64)	66 (57–74)
Eggs	12 (9–17)	–	–	–
Aquatic animals				
Fish	2 (1–3)	–	–	0 (0–<1)
Other seafood	0 (0–<1)	–	–	6 (2–11)
Plants				
Grains, beans	1 (<1–2)	–	–	–
Oils, sugars	–	–	–	0 (0–1)
Fruits	12 (8–16)	7 (3–12)	50 (5–77)	1 (<1–2)
Seeded vegetables	18 (13–25)	–	–	6 (1–13)
Sprouts	8 (5–12)	1 (<1–1)	8 (1–22)	–
Vegetable row crops	3 (1–6)	36 (26–46)	3 (<1–13)	6 (2–11)
Other produce	7 (3–11)	1 (0–2)	–	2 (<1–6)

*Estimates calculated by using analysis of variance model–estimated outbreak illnesses for single pathogen, single food category outbreaks occurring during 1998–2012, with down-weighting of outbreaks that occurred during 1998–2007. Because of rounding, 0 indicates nonzero estimates <0.5. Dashes indicate pathogen–food category pairs for which we did not estimate attribution percentages because of a lack of outbreaks.

Only 4 other food categories (sprouts, turkey, vegetable row crops, and pork) had estimated attribution percentages >1%. The wide and overlapping credibility intervals reflect the very small number of *L. monocytogenes* outbreaks in the analysis (n = 24).

An estimated 66% of *Campylobacter* outbreak illnesses were attributed to the dairy category. This percentage was substantially higher than for any other food category. About 8% of illnesses are attributed to chicken; 6% were attributed respectively to vegetable row crops, seeded vegetables, and other seafood.

We found estimates to be robust across a wide variety of scenarios. We assessed the sensitivity of estimates to particularly influential outbreaks and to modeling decisions, such as choices in statistical modeling, down-weighting of older outbreaks, and consideration of etiology status (Appendix).

Discussion

Although only a small proportion of foodborne illnesses are part of recognized outbreaks, outbreak investigations can provide insights into the causes and contributing factors leading to infection. Because linking an illness to a particular food is rarely possible except during an outbreak, aggregated data from foodborne outbreaks have been used to estimate the food sources of all illnesses caused by specific pathogens in numerous countries (*6*–*12*).

Whereas our approach addresses numerous challenges with estimating attribution percentages on the basis of outbreak data, some issues must be considered when using these estimates to inform food safety decision-making. Our analysis does not indicate the point of contamination, because outbreak investigations implicate only the food vehicle that was consumed. Moreover, the outbreaks included in this analysis include only 35% of the reported foodborne disease outbreaks caused by these pathogens during the study period, and they might not be representative of all foodborne outbreaks caused by these pathogens. The exclusion of outbreaks attributable to complex foods for which the contaminated ingredient was not determined could result in underrepresentation of food categories containing foods often eaten as part of complex dishes (e.g., leafy greens and eggs) (*23*). However, because the published method for assigning food categories to these complex food outbreaks is somewhat subjective and relies on internet searches for recipes (*24*), excluding these outbreaks provides results based on the most accurate available data. A method to incorporate data from these outbreaks is being developed.

Foods are implicated in outbreaks through epidemiologic analyses, by isolation of the causal pathogen from implicated food, through examination of supply chain records or environmental assessments, or by other information. The strength of evidence implicating foods varies widely across outbreaks.

For some pathogens and pathogen–food pairs, the number of outbreaks available for analyses was

quite low. For example, our data include only 24 outbreaks caused by *L. monocytogenes*, so 1 fruits-linked outbreak (a very large outbreak that occurred in 2011 and was associated with contaminated cantaloupe) had a profound influence on attribution estimates for this pathogen. However, our approach reflects the uncertainties associated with sparse data in wider credibility intervals.

Although we weighted recent outbreaks more heavily than older ones because recent outbreaks are probably more representative of current attribution, we did not formally account for possible changes in underlying factors over time in the main effects model. Examples of such factors could include changes in pathogen-specific disease incidence, outbreak investigation practices, or outbreak reporting by states. Generalizing outbreak-based attribution to overall foodborne disease assumes, implicitly, that the foods implicated in outbreaks reflect the food sources of illness in the general population. However, these assumptions might not always hold. For example, ≈10% of outbreaks occurred among institutionalized populations, such as those in correctional facilities, hospitals, and nursing homes. In these populations, case-ascertainment rates, food options, and sources of food contamination might not be representative of the general population. However, such outbreaks might elucidate setting- or subpopulation-specific contamination problems that are difficult to identify among the general population.

Campylobacter attribution presents a specific challenge. Our outbreak-based model attributes 66% (90% credibility interval 57%–74%) of foodborne campylobacteriosis to dairy, which is in line with other outbreak-based estimates for the United States (*6*,*13*). However, most foodborne *Campylobacter* outbreaks in this study were associated with unpasteurized fluid milk, which is not widely consumed by the general population. For example, in a Foodborne Active Surveillance Network population survey of food exposures, only 3% reported consuming unpasteurized milk in the preceding week (*25*). Moreover, outbreak-based estimates are not consistent with other lines of evidence. An analysis of 38 case–control studies of sporadic campylobacteriosis found a much smaller percentage of illnesses attributable to consumption of raw milk than chicken (*12*). For example, 1 of these studies, a Foodborne Active Surveillance Network

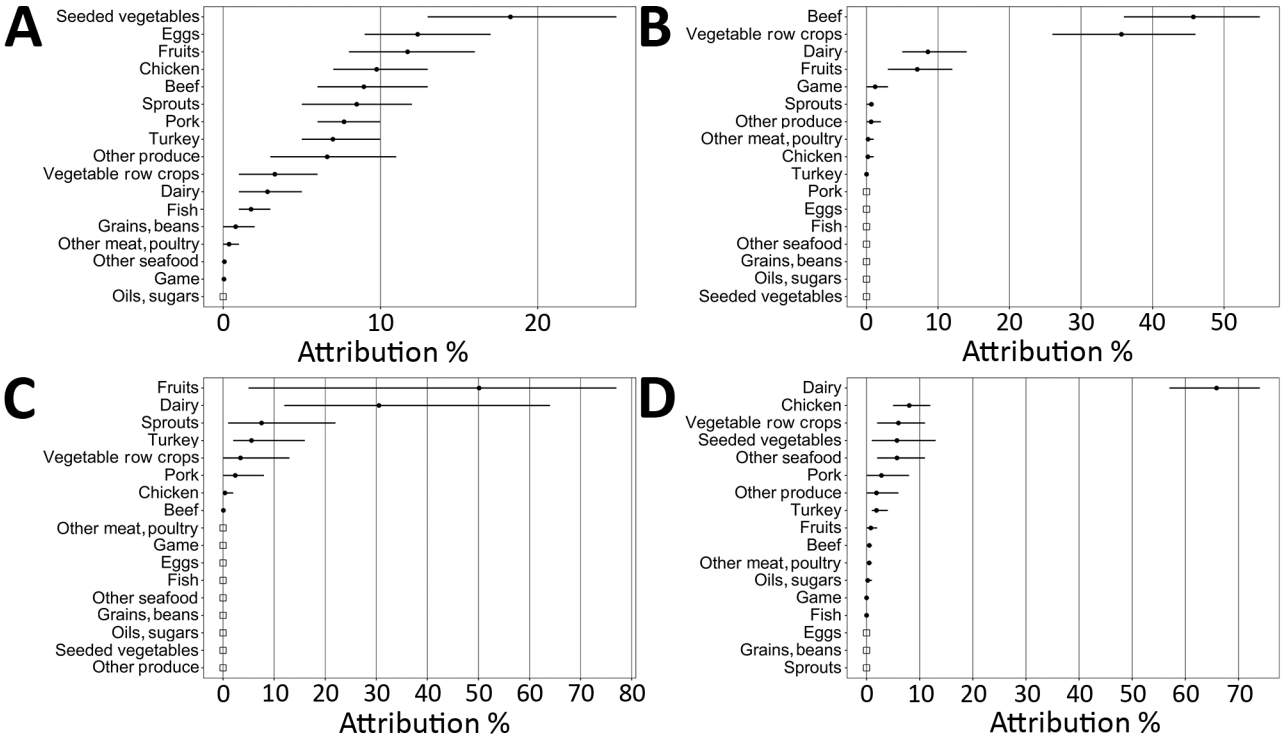


Figure 3. Estimated percentages of foodborne illnesses attributed to food categories and 90% credibility intervals (error bars) for *Salmonella* (A)*, Escherichia coli* O157 (B)*, Listeria monocytogenes* (C)*, and Campylobacter* (D), based on analysis of single-pathogen, single-food category outbreaks, Foodborne Disease Outbreak Surveillance System, United States, 1998–2012. Percentages are presented in descending order. Open squares indicate that no illnesses were attributed to that food category because no outbreaks were reported for that pathogen in that food category during the study period. Estimates calculated by using analysis of variance model–estimated outbreak illnesses for single pathogen, single food category outbreaks occurring during 1998–2012, with down-weighting of outbreaks that occurred during 1998–2007.

case–control study, attributed 1.5% of campylobacteriosis cases to consumption of unpasteurized milk, compared with 24% to consumption of chicken prepared in a restaurant (*26*). Structured expert judgment studies conducted in the United States and in other countries estimate 8%–10% of foodborne campylobacteriosis to be attributable to dairy products (principally, raw milk), compared with 33%–72% to chicken (*27*–*30*).

Because *Campylobacter* outbreaks appear to overrepresent dairy as a source of sporadic *Campylobacter* illness, we do not advise using these attribution percentages without further adjustment or without considering additional information. Removing the dairy category entirely might be an appropriate adjustment, given that the resulting distribution of *Campylobacter* attribution estimates across other food categories is more consistent with the published literature (*31*). When the dairy category is excluded from this analysis, 29% of *Campylobacter* illnesses are attributed to poultry (23.5% to chicken and 5.5% to turkey), 18% to vegetable row crops, 17% to seeded vegetables, 17% to other seafood, 8% to pork, 6% to other produce, and 6% to other food categories.

Our estimates reflect data on outbreaks that occurred during 1998–2012 because those were the most recent data available at the outset of this effort. We do not include more recent outbreaks in this analysis because substantial preparation of the data was needed, and because the primary purpose of this report is to describe our methods and explain modeling decisions. IFSAC has published reports based on more recent outbreaks using the methodology described in this article (*32*).

To address some of the challenges with using outbreak data to estimate the food categories responsible for foodborne illnesses, we developed an approach that reduces the influence of large outbreaks, adjusts for important epidemiologic characteristics, and weights recent data more heavily than older data. We also incorporate an updated food categorization scheme better aligned to the needs of regulatory agencies and provide statistical uncertainty around the estimates. This approach can be used for routine updating of estimates by incorporating additional years of data.

The resulting estimates of attribution percentages for *Salmonella*, *E. coli* O157, *L. monocytogenes*, and *Campylobacter* can play an important role in science- and risk-based decision-making because they can be used alongside other data to inform regulatory decisions, to prioritize food safety efforts, and to evaluate the effectiveness of prevention measures. Further, federal agency consensus on a single set of outbreak-based attribution estimates improves the transparency of governmental efforts to inform and engage stakeholders, such as industry and consumers, about food safety strategies.

Acknowledgments

The authors are indebted to Christopher Alvares, Christopher Aston, Marc Boyer, Chris Braden, Beau Bruce, Eric Ebel, David Goldman, Chuanfa Guo, Kristin Holt, Shacara Johnson, Sherri McGarry, Kara Morgan, Debra Street, Robert Tauxe, Curtis Travis, Iris Valentin-Bon, Antonio Vieira, Katherine Vierk, Christopher Waldrop, Michael Williams, and others within Centers for Disease Control and Prevention (CDC), the Food and Drug Administration, and the US Department of Agriculture's Food Safety and Inspection Service who provided input on this work. We also thank the many other persons from outside these agencies who shared feedback. Special thanks also go to the CDC National Outbreak Reporting System Team for access and guidance in using outbreak data, and to the state, local, tribal, and territorial health departments who report these outbreaks to CDC.

About the Author

Mr. Batz is a senior policy advisor in the Office of Analytics and Outreach, Center for Food Safety and Applied Nutrition, US Food and Drug Administration. His research focus is on improving public health decision-making by quantifying foodborne disease risks.

References

1. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, et al. Foodborne illness acquired in the United States—major pathogens. Emerg Infect Dis. 2011;17:7–15. https://doi.org/10.3201/eid1701.P11101
2. Hoffmann S, Maculloch B, Batz MB. Economic burden of major foodborne illnesses acquired in the United States. Washington: US Department of Agriculture Economic Research Service; 2015. Economic Information Bulletin No. (EIB-140) [cited 2018 Jun 6]. https://www.ers.usda.gov/publications/pub-details/?pubid=43987
3. Minor T, Lasher A, Klontz K, Brown B, Nardinelli C, Zorn D. The per case and total annual costs of foodborne illness in the United States. Risk Anal. 2015;35:1125–39. https://doi.org/10.1111/risa.12316
4. Institute of Medicine and National Research Council Committee on the Review of Food and Drug Administration's Role in Ensuring Safe Food. Enhancing food safety: the role of the Food and Drug Administration. Wallace RB, Oria M, editors. Washington: National Academies Press; 2010.
5. Mangen MJ, Batz MB, Käsbohrer A, Hald T, Morris JG Jr, Taylor M, et al. Integrated approaches for the public health prioritization of foodborne and zoonotic pathogens. Risk Anal. 2010;30:782–97. https://doi.org/10.1111/j.1539-6924.2009.01291.x

6. Batz MB, Hoffmann S, Morris JG Jr. Ranking the disease burden of 14 pathogens in food sources in the United States using attribution data from outbreak investigations and expert elicitation. J Food Prot. 2012;75:1278–91. https://doi.org/10.4315/0362-028X.JFP-11-418
7. Greig JD, Ravel A. Analysis of foodborne outbreak data reported internationally for source attribution. Int J Food Microbiol. 2009;130:77–87. https://doi.org/10.1016/j.ijfoodmicro.2008.12.031
8. Jackson BR, Griffin PM, Cole D, Walsh KA, Chai SJ. Outbreak-associated *Salmonella enterica* serotypes and food commodities, United States, 1998–2008. Emerg Infect Dis. 2013;19:1239–44. https://doi.org/10.3201/eid1908.121511
9. King N, Lake R, Campbell D. Source attribution of nontyphoid salmonellosis in New Zealand using outbreak surveillance data. J Food Prot. 2011;74:438–45. https://doi.org/10.4315/0362-028X.JFP-10-323
10. Pires SM, Vigre H, Makela P, Hald T. Using outbreak data for source attribution of human salmonellosis and campylobacteriosis in Europe. Foodborne Pathog Dis. 2010;7:1351–61. https://doi.org/10.1089/fpd.2010.0564
11. Ravel A, Greig J, Tinga C, Todd E, Campbell G, Cassidy M, et al. Exploring historical Canadian foodborne outbreak data sets for human illness attribution. J Food Prot. 2009;72:1963–76. https://doi.org/10.4315/0362-028X-72.9.1963
12. Domingues AR, Pires SM, Halasa T, Hald T. Source attribution of human campylobacteriosis using a meta-analysis of case-control studies of sporadic infections. Epidemiol Infect. 2012;140:970–81. https://doi.org/10.1017/S0950268811002676
13. Painter JA, Hoekstra RM, Ayers T, Tauxe RV, Braden CR, Angulo FJ, et al. Attribution of foodborne illnesses, hospitalizations, and deaths to food commodities by using outbreak data, United States, 1998–2008. Emerg Infect Dis. 2013;19:407–15. https://doi.org/10.3201/eid1903.111866
14. Interagency Food Safety Analytics Collaboration. About IFSAC. 2017 [cited 2017 Aug 7]. https://www.cdc.gov/foodsafety/ifsac/overview/index.html
15. Ebel ED, Williams MS, Cole D, Travis CC, Klontz KC, Golden NJ, et al. Comparing characteristics of sporadic and outbreak-associated foodborne illnesses, United States, 2004–2011. Emerg Infect Dis. 2016;22:1193–200. https://doi.org/10.3201/eid2207.150833
16. Gould LH, Walsh KA, Vieira AR, Herman K, Williams IT, Hall AJ, et al.; Centers for Disease Control and Prevention. Surveillance for foodborne disease outbreaks—United States, 1998–2008. MMWR Surveill Summ. 2013;62:1–34.
17. Richardson LC, Bazaco MC, Parker CC, Dewey-Mattia D, Golden N, Jones K, et al. An updated scheme for categorizing foods implicated in foodborne disease outbreaks: a tri-agency collaboration. Foodborne Pathog Dis. 2017;14:701–10. https://doi.org/10.1089/fpd.2017.2324
18. Centers for Disease Control and Prevention. Guide to confirming an etiology in foodborne disease outbreak. 2015 [cited 2018 Jul 12]. https://www.cdc.gov/foodsafety/outbreaks/investigating-outbreaks/confirming_diagnosis.html
19. Brown RG. Smoothing, forecasting and prediction of discrete time series. Englewood Cliffs (NJ): Prentice-Hall; 1963.
20. Ngo L, Tager IB, Hadley D. Application of exponential smoothing for nosocomial infection surveillance. Am J Epidemiol. 1996;143:637–47. https://doi.org/10.1093/oxfordjournals.aje.a008794
21. Rubin DB. The Bayesian bootstrap. Ann Stat. 1981;9:130–4. https://doi.org/10.1214/aos/1176345338
22. Davison AC, Hinkley DV. Bootstrap methods and their application. Cambridge and New York: Cambridge University Press; 1997.
23. St Louis ME, Morse DL, Potter ME, DeMelfi TM, Guzewich JJ, Tauxe RV, et al. The emergence of grade A eggs as a major source of *Salmonella enteritidis* infections. New implications for the control of salmonellosis. JAMA. 1988;259:2103–7. https://doi.org/10.1001/jama.1988.03720140023028
24. Painter JA, Ayers T, Woodruff R, Blanton E, Perez N, Hoekstra RM, et al. Recipes for foodborne outbreaks: a scheme for categorizing and grouping implicated foods. Foodborne Pathog Dis. 2009;6:1259–64. https://doi.org/10.1089/fpd.2009.0350
25. Centers for Disease Control and Prevention. Foodborne Active Surveillance Network (FoodNet) population survey atlas of exposure, 2006–2007. Atlanta: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2011 [cited 2019 Dec 12]. https://www.cdc.gov/foodnet/surveys/foodnetexposureatlas0607_508.pdf
26. Friedman CR, Hoekstra RM, Samuel M, Marcus R, Bender J, Shiferaw B, et al.; Emerging Infections Program FoodNet Working Group. Risk factors for sporadic *Campylobacter infection* in the United States: a case-control study in FoodNet sites. Clin Infect Dis. 2004;38(Suppl 3):S285–96. https://doi.org/10.1086/381598
27. Hoffmann S, Fischbeck P, Krupnick A, McWilliams M. Using expert elicitation to link foodborne illnesses in the United States to foods. J Food Prot. 2007;70:1220–9. https://doi.org/10.4315/0362-028X-70.5.1220
28. Havelaar AH, Galindo AV, Kurowicka D, Cooke RM. Attribution of foodborne pathogens using structured expert elicitation. Foodborne Pathog Dis. 2008;5:649–59. https://doi.org/10.1089/fpd.2008.0115
29. Tam CC, Larose T, O'Brien SJ. Costed extension to the Second Study of Infectious Intestinal Disease in the Community: identifying the proportion of foodborne disease in the UK and attributing foodborne disease by food commodity. Liverpool (UK): University of Liverpool; 2014. Project B18021 (FS231043) [cited 2019 Dec 12]. https://livrepository.liverpool.ac.uk/3014609
30. Butler AJ, Pintar KD, Thomas MK. Estimating the relative role of various subcategories of food, water, and animal contact transmission of 28 enteric diseases in Canada. Foodborne Pathog Dis. 2016;13:57–64. https://doi.org/10.1089/fpd.2015.1957
31. Interagency Food Safety Analytics Collaboration. Foodborne illness source attribution estimates for 2013 for *Salmonella, Escherichia coli* O157, *Listeria monocytogenes*, and *Campylobacter* using multi-year outbreak surveillance data, United States. Atlanta and Washington: Centers for Disease Control and Prevention, US Food and Drug Administration, US Department of Agriculture Food Safety and Inspection Service; 2017 [cited 2020 May 4]. https://www.cdc.gov/foodsafety/ifsac/annual-reports.html
32. Interagency Food Safety Analytics Collaboration. Foodborne illness source attribution estimates for 2017 for *Salmonella, Escherichia coli* O157, *Listeria monocytogenes*, and *Campylobacter* using multi-year outbreak surveillance data, United States. Atlanta and Washington: Centers for Disease Control and Prevention, US Food and Drug Administration, US Department of Agriculture Food Safety and Inspection Service; 2019 [cited 2020 May 4]. https://www.cdc.gov/foodsafety/ifsac/annual-reports.html

Address for correspondence: Michael B. Batz, Center for Food Safety and Applied Nutrition, Food and Drug Administration, 5001 Campus Dr, College Park, MD, 20740, USA; email: michael.batz@fda.hhs.gov

Article DOI: https://doi.org/10.3201/eid2701.203832

Recency-Weighted Statistical Modeling Approach to Attribute Illnesses Caused by 4 Pathogens to Food Sources Using Outbreak Data, United States

Appendix

IFSAC Food Categorization Scheme

Appendix Figure 1 shows the categorization scheme used to classify foods implicated in outbreaks. The scheme and the associated methodology used to assign outbreaks to food categories based on implicated foods and ingredients, as well as examples of foods assigned to each category, are described in Richardson et al. (*1*).

Description of Data Set

This section provides a general description of the data used in the source attribution model described in this article.

We extracted data on 16,584 foodborne disease outbreaks reported in the 15 years from 1998 through 2012 from CDC's Foodborne Disease Outbreak Surveillance System (FDOSS) (https://www.cdc.gov/foodsafety/fdoss) (*2*). These data were extracted on December 18, 2013.

Appendix Figure 2 shows the stages of data preparation. Specifically, it shows the number of outbreaks excluded from the analysis at each stage (shaded boxes) and the number remaining (unshaded boxes).

First, we excluded outbreaks that occurred in outlying U.S. territories (e.g., Puerto Rico). Next, we excluded the 83% of remaining outbreaks not caused by the 4 priority pathogens: nontyphoidal *Salmonella*, *Escherichia coli* O157 (namely, *E. coli* O157:H7 and *E. coli* O157:NM), *Listeria monocytogenes*, and *Campylobacter* spp. Of these 2,732 outbreaks, we further excluded 77 outbreaks caused by multiple pathogens.

Of the resulting 2,655 outbreaks caused by one of the 4 priority pathogens as the single etiology, we excluded 38% (n = 1,014) because investigators did not identify an implicated food and 26% (n = 689) because the implicated food(s) could not be assigned to a single food category. Implicated foods could not be assigned to a single food category because the identified food was complex (composed of ingredients belonging to more than one food category) (n = 448); or foods from more than one food category were implicated or suspected (i.e., multiple foods) (n = 142); or the food was too vaguely described to be assigned to any category (e.g., "buffet," "appetizer") (n = 50); or the food was too vaguely described to be assigned to the specific food categories used in the analysis (e.g., could only be assigned to "Produce" or "Meat-Poultry") (n = 49).

We focus on single-pathogen single-food category outbreaks because the appropriate categorization of both pathogens and foods is known. A method is in development for assigning to multiple food categories those outbreaks due to complex foods for which the implicated ingredient was unknown. The previously published approach could not be applied to our data series without substantial revisions (*3*). This approach used "recipes" developed using internet searches and these recipes would need to be updated to reflect current online recipes and to incorporate changes to food categories (*4*).

Thus, our final dataset for analysis included outbreaks caused by a single pathogen that could be assigned to one of 22 specific food categories; these were 952 (36%) of the 2,655 single etiology outbreaks. The pathogen with the most outbreaks in the resulting data was *Salmonella* (n = 597), the predominant serotype of which was Enteritidis (n = 184) (Appendix Table 1). There were 170 outbreaks caused by *E. coli* O157, 24 by *L. monocytogenes*, and 161 by *Campylobacter* (Appendix Figure 2).

As part of preliminary analyses, we assessed the quality of information on etiology status. In FDOSS, an outbreak must have at least 2 ill persons (*2*). For *Salmonella*, *E. coli* O157, and *Campylobacter*, outbreaks with "confirmed" etiology are defined as those in which the outbreak strain was isolated from at least 2 patients or from epidemiologically implicated food; "confirmed" outbreaks of *L. monocytogenes* infections must have 1 person with the outbreak strain isolated from a normally sterile site (*5*). (Cases of listeriosis can also be diagnosed based on symptoms and culture of products of conception, which are not sterile.) The etiology of an

outbreak not meeting these conditions is considered to be "suspected." Of the 2,732 outbreaks associated with the 4 priority pathogens, 90% (2,462) were coded as having confirmed etiology. We found that 12% of outbreaks coded as having confirmed etiology did not have sufficient data to fulfill the confirmed etiology definition, but also found that over 95% of outbreaks coded as having suspected etiology had at least one laboratory-confirmed illness. Outbreaks occurring early in the study period were more likely to have insufficient data to confirm an etiology.

We decided to include outbreaks with either confirmed or suspected etiology status in the analysis so as not to lose information associated with those outbreaks, following the decision made in Painter et al. (*3*). We also conducted a sensitivity analysis on this decision, as described elsewhere in this appendix.

Of the 952 outbreaks in the data used to estimate attribution percentages, 83 (8.7%) did not have a confirmed etiology, including 8.9% (n = 53) of *Salmonella* outbreaks, 4.1% (n = 7) of *E. coli* O157 outbreaks, 12.5% (n = 3) of *L. monocytogenes* outbreaks, and 12.4% (n = 20) of *Campylobacter* outbreaks.

NORS includes 3 variables related to outbreak size: the number of lab-confirmed primary cases (ConfirmedPrimary), the number of additional illnesses that were not laboratory confirmed (ProbablePrimary), and the total of both confirmed and probable illnesses (EstimatedPrimary). In our attribution estimates, we use the estimated total illnesses as our measure of outbreak size.

Statistical Model Development

This section provides additional details about the models used to estimate the food sources of illnesses. Specifically, it describes development of pathogen-specific statistical models of outbreak size, and the approach used to weight recent outbreaks more heavily than older outbreaks.

Analysis of variance models

Log-transforming outbreak size resulted in relatively normally distributed outbreak illness numbers that could be modeled using straightforward analysis of variance (Te) modeling techniques. We explored several modeling approaches, including analysis of covariance (ANCOVA), generalized linear models, and least absolute shrinkage and selection operator (LASSO) models, among others. We decided to use ANOVA based on structural simplicity and

interpretability, and because our data were not sufficient to credibly describe the complexity of interactions between the epidemiologic characteristics of reported outbreaks.

We developed pathogen-specific models because we did not want to smooth over differences in outbreak size by pathogen, as this variation likely results from epidemiologic factors, not random variation. We found outbreaks caused by different *Salmonella* serotypes varied in foods implicated and other epidemiologic factors. In particular, serotype Enteritidis outbreaks had some distinct patterns. Thus, we decided to model serotype Enteritidis separately from all the other serotypes for estimating outbreak size; when we calculate attribution percentages in a subsequent stage, we do so after summing the 2 sets of model estimates.

Based on preliminary modeling analysis and considerations of epidemiologic importance, we included 3 variables as the predictors of outbreak size in all 5 pathogen-specific models: food category, the type of location at which the food was prepared, and whether outbreak exposures occurred in a single state or in multiple states. Each outbreak was assigned to 1 of 17 food categories, as described previously. The food preparation location variable used in the model included 5 categories, based on 24 individual location types identified in outbreak reports, as shown in Appendix Table 2. We reduced the number of categories to 5 to address the relatively sparse data across most locations other than restaurant or private home. A dichotomous variable was used to indicate whether exposures occurred in multiple states or a single state.

We desired a model that was portable in that it could be similarly described across the 4 etiologies included in the study and expandable to additional pathogens. Summary measures for model fit are shown in Appendix Table 3, including traditional lack-of-fit, R-squared, overall model significance, and significance of each predictor. Appendix Table 3 also shows, in the last 3 columns, variance explained via random forest decomposition using identical predictors. Appendix Figure 3 compares the number of reported illnesses with the number of model-estimated illnesses and shows that, as expected, our ANOVA models reduce variation in outbreak size and the influence of very large outbreaks.

Recency weighting

The decision to down-weight older data was made because recent outbreaks are likely to be more representative of current foodborne illness attribution than older outbreaks. Changes in attributable risk may result from changes over time in food consumption patterns, food

production and processing practices, food safety activities, regulatory interventions, and other factors.

This decision is supported by characteristics of the underlying data. Appendix Figure 4 presents a heat map with the number of outbreaks by pathogen and food category over time. White cells indicate no outbreaks due to that pathogen-food category pair in that year, with color from pale orange to red indicating between 1 and 25 outbreaks in that year. Appendix Figure 4 illustrates the variability in data sparseness across many pathogen-food categories.

We examined the impacts of excluding older data by estimating attribution for 3, 5-year time frames: 1998–2002, 2003–2007, and 2008–2012. Appendix Figure 5 displays the estimated attribution percentages (y-axis) by food category (lines) and timeframe (x-axis). There are notable differences across these timeframes. The variability of underlying data leads to instability in estimated percentages based on short time windows. Excluding older data entirely results in estimates of zero attribution for some categories with known nonzero risk.

Based on this and other analyses, we decided that outbreaks older than 5 years should be included in estimates of attribution but down-weighted to increase the relative influence of more recent outbreaks on attribution estimates.

As described in the article, we determined that the most appropriate approach would be to use an exponential decay function to define the recency-weighting multiplier *w* for an outbreak in year *y*, as a function of decay parameter *a*:

$$w_y = \begin{cases} a^{2008-y}, y < 2008 \\ 1, y \geq 2008 \end{cases}$$

We evaluated various options for the decay parameter *a* and the resulting weighting factor by year, as shown in Appendix Figure 6. Our preference was for more than half of the information in our estimates to come from the most recent 5 year period, and a small amount – around 5% – from data older than 10 years. Because the distribution of outbreak illnesses is not constant over time or by pathogen (as shown in Appendix Figure 4), we selected a decay parameter that best met our preferences for all pathogens. As shown in Appendix Table 4, with a decay parameter value of 0.7142 (5/7), 67% of the total down-weighted model-estimated outbreak illnesses used in the attribution calculation were from outbreaks that occurred during

the most recent 5-year period (2008–2012), with ≈28% from the middle 5-year period, and 5% from the oldest 5-year period.

Sensitivity Analyses

This appendix describes sensitivity analyses conducted to assess the robustness of our attribution estimates. We compare our model-based estimates to those derived used in prior studies and explore sensitivities to modeling decisions and underlying data.

Sensitivity to Use of Statistical Modeling and Recency-weighting

Prior estimates of foodborne illness source attribution based on outbreaks have summed the raw number of reported outbreaks or outbreak illnesses associated with a given pathogen-food category pair and divided this by the total number of outbreaks or outbreak illnesses associated with that pathogen (*3*,*6*). By contrast, our estimates are based on statistical modeling of log-transformed outbreak size, with exponential down-weighting of older outbreaks (we refer to these as our "baseline" estimates).

Appendix Figure 7 compares our model-based attribution percentages (with and without down-weighting of older outbreaks) to those based on raw numbers of reported outbreaks and outbreak illnesses. Attribution percentages are shown in log scale to better highlight differences. Differences reflect dependencies between multinomial estimates; because attribution percentages sum to 100%, a downward shift in the percentage for one food category results in higher percentages elsewhere.

Appendix Figure 7 illustrates 3 points. First, it shows the range of estimates using the methods in the published literature, namely that there are notable differences between attribution estimates based on numbers of reported outbreaks (purple lines) and numbers of outbreak illnesses (pink lines). Pathogen-food category pairs with the largest differences are *Salmonella* in Seeded Vegetables and Chicken, *E. coli* O157 in Fruits and Sprouts, *L. monocytogenes* in Fruits and Turkey, and *Campylobacter* in Chicken and Other Seafood. These differences reflect the outbreak size variation across food categories, as well as the impact of very large outbreaks.

Second, Appendix Figure 7 also shows that although our attribution estimates are generally similar to estimates based on the approaches used in the published literature, there are some differences. For *Salmonella*, the Seeded Vegetables category has the highest attribution

percentage in our baseline estimates, whereas the Eggs category has the highest percentage based on numbers of reported outbreaks (purple) or outbreak illnesses (pink). Chicken also has higher attribution percentages based on counts of reported outbreaks or outbreak illnesses, compared to the baseline, whereas the Sprouts category has a lower estimate. For *E. coli* O157, baseline estimates for the Vegetable Row Crops category are higher than those based on reported outbreaks or outbreak illnesses. For *L. monocytogenes*, our baseline estimates are closest to those based on counts of reported outbreak illnesses, though the baseline estimate for Turkey is lower than those based on counts of reported outbreaks or outbreak illnesses, and the baseline estimate for Fruits is respectively higher. There are fewer differences in *Campylobacter* estimates, though the Seeded Vegetables category has a notably higher attribution percentage when based on model-estimated illnesses.

Lastly, Appendix Figure 7 shows that eliminating recency-weighting affects attribution percentages, though not drastically. For *Salmonella*, attribution percentages calculated without recency-weighting (green) are higher for Chicken and Eggs, and lower for Seeded Vegetables and Vegetable Row Crops, compared to baseline estimates with recency-weighting (red). For *E. coli* O157, estimates for Dairy and Vegetable Row Crops are marginally lower than the baseline without recency-weighting; estimates for Sprouts and Fruits are marginally higher. For *L. monocytogenes*, removing recency-weighting results in lower estimates for Fruits and higher estimates for Turkey, reflecting the impact of the large cantaloupe outbreak in 2011, its recency, and the fact that there were not any *L. monocytogenes* outbreaks traced to turkey luncheon meat between 2005 and 2012 (Appendix Figure 4). For *Campylobacter*, removing recency-weighting results in a higher estimate for Other Produce, reflecting the impact of a large 2002 prison outbreak associated with potatoes (the only outbreak in this pathogen-food category (Appendix Figure 4).

Sensitivity to ANOVA Model Specifications

As noted in the text and other appendices, we conducted exploratory analyses to determine which predictors should be included in the pathogen-specific ANOVA models. The final 3-predictor model specifications were based both on epidemiologic reasoning and our findings that these variables were statistically significant predictors of outbreak size.

We conducted sensitivity analyses around the final model specification by estimating attribution percentages using 3 alternative ANOVA models: one without the dichotomous multi-state variable, one without the categorical preparation location variable, and one without either. The results (Appendix Figure 8) show that our model is robust to model specification decisions in comparison with the baseline model specification.

Sensitivity to Etiology Status

As described previously, we included outbreaks with "suspected" etiology in addition to those with laboratory-confirmed isolates from patients or food in the analysis. Those without confirmed etiology comprise ≈9% of the outbreaks used in the analysis, though of these, most had at least one laboratory-confirmed illness. We conducted a sensitivity analysis around this decision. Appendix Figure 9 presents our baseline attribution estimates and 90% credibility intervals alongside estimates based on data excluding the 83 outbreaks with suspected etiology. Appendix Figure 9 shows that for all but a few pathogen-food category pairs, the differences in point estimates are minimal, though credibility intervals are wider when outbreaks of suspected etiology are excluded.

Sensitivity to Influential Outbreaks

We conducted a series of analyses to identify which outbreaks are most influential on our attribution estimates and to assess model sensitivity to these outbreaks. This was done in part to ascertain the extent to which our estimates were sensitive to very large outbreaks, though because our estimates are based on a 3-parameter statistical model of log-transformed outbreak size, with recency-weighting, we needed a systematic approach to identify influential outbreaks.

The first step was to define an influence metric for each outbreak based on the aggregate difference in attribution estimates when that outbreak was excluded from the analysis. That is, for each of 952 outbreaks, we estimated attribution percentages without that single outbreak. We defined an "influence metric" as the sum of mean differences squared across all pathogen-food category pairs between the baseline estimate and the estimate without that outbreak; the attribution percentages change only for the pathogen for which an outbreak was excluded. We then calculated the overall "influence rank" for each outbreak based on the rank order of the "influence metric."

Appendix Figure 10 presents, for each pathogen, the calculated influence metric for each outbreak, in descending order. These plots show that most outbreaks have influence metrics at or very close to zero, but a small number do have measurable influence metrics. Appendix Figure 10 shows that the 10 outbreaks most influential on attribution estimates were caused by *L. monocytogenes* and *Campylobacter*. Appendix Table 5 provides details for the 5 outbreaks most influential on attribution estimates for each pathogen. Although the plots in Appendix Figure 10 show that some outbreaks have large influence metric values, the actual impacts of these individual outbreaks on attribution estimates is minimal.

Appendix Figure 11 presents estimates for scenarios in which each of the 5 outbreaks most influential on attribution estimates (from Appendix Table 5) for each pathogen was excluded one at a time. These scenarios are shown alongside the baseline attribution percentages. Appendix Figure 11 shows that for all but the single most influential outbreak (*L. monocytogenes* in cantaloupe), the exclusion of any single outbreak results in negligible differences in attribution estimates, and no differences in the rank order of food categories. We therefore concluded that our model is robust to all but the most extreme outliers, and that only our estimates for *L. monocytogenes* are sensitive to the impact of individual outbreaks.

References

1. Richardson LC, Bazaco MC, Parker CC, Dewey-Mattia D, Golden N, Jones K, et al. An updated scheme for categorizing foods implicated in foodborne disease outbreaks: a tri-agency collaboration. Foodborne Pathog Dis. 2017;14:701–10. PubMed https://doi.org/10.1089/fpd.2017.2324

2. Gould LH, Walsh KA, Vieira AR, Herman K, Williams IT, Hall AJ, et al.; Centers for Disease Control and Prevention. Surveillance for foodborne disease outbreaks—United States, 1998–2008. MMWR Surveill Summ. 2013;62:1–34. PubMed

3. Painter JA, Hoekstra RM, Ayers T, Tauxe RV, Braden CR, Angulo FJ, et al. Attribution of foodborne illnesses, hospitalizations, and deaths to food commodities by using outbreak data, United States, 1998-2008. Emerg Infect Dis. 2013;19:407–15. PubMed https://doi.org/10.3201/eid1903.111866

4. Painter JA, Ayers T, Woodruff R, Blanton E, Perez N, Hoekstra RM, et al. Recipes for foodborne outbreaks: a scheme for categorizing and grouping implicated foods. Foodborne Pathog Dis. 2009;6:1259–64. PubMed https://doi.org/10.1089/fpd.2009.0350

5. Centers for Disease Control and Prevention. Guide to confirming an etiology in foodborne disease outbreak. 2015 [cited 2018 July 12]. https://www.cdc.gov/foodsafety/outbreaks/investigating-outbreaks/confirming_diagnosis.html

6. Batz MB, Hoffmann S, Morris JG Jr. Ranking the disease burden of 14 pathogens in food sources in the United States using attribution data from outbreak investigations and expert elicitation. J Food Prot. 2012;75:1278–91. PubMed https://doi.org/10.4315/0362-028X.JFP-11-418

Appendix Table 1. Number of nontyphoidal *Salmonella* outbreaks and outbreak illnesses due to a single food category — Foodborne Disease Outbreak Surveillance System, United States, 1998–2012*

Food category	Serotype Enteritidis	Other serovars	All serovars
Beef	9 (157)	38 (1,316)	47 (1,473)
Pork	8 (167)	43 (931)	51 (1,098)
Chicken	26 (580)	88 (2,068)	114 (2,648)
Turkey	7 (420)	42 (888)	49 (1,308)
Other meat, poultry	1 (13)	5 (71)	6 (84)
Game	1 (3)	1 (5)	2 (8)
Dairy	1 (39)	23 (754)	24 (793)
Eggs	113 (4,895)	27 (350)	140 (5,245)
Fish	2 (82)	10 (204)	12 (286)
Other seafood	2 (26)	2 (10)	4 (36)
Grains, beans	0 (0)	7 (268)	7 (268)
Oils, sugars	0 (0)	0 (0)	0 (0)
Fruits	5 (240)	41 (2,270)	46 (2,510)
Seeded vegetables	1 (85)	33 (3,916)	34 (4,001)
Sprouts	5 (174)	28 (1,092)	33 (1,266)
Vegetable row crops	1 (14)	9 (398)	10 (412)
Other produce	2 (45)	16 (1,878)	18 (1,923)
Total	184 (6,940)	413 (16,419)	597 (23,359)

*Number of outbreak-associated illnesses in parentheses. Nontyphoidal *Salmonella* is divided into *S. enterica* ser. Enteritidis and other serovars.

Appendix Table 2. Types of food preparation locations as defined in reported outbreak data and aggregated categories and outbreak counts used in statistical models of outbreak size

Preparation locations identified in outbreak line listings	Categories used ANOVA model
Restaurant – "Fast-food" Restaurant – other/unknown type Restaurant – Sit-down dining	Restaurant
Restaurant or deli	Private Home
Private Home Banquet facility Caterer Caterer (food prepared off-site) Fair, festival, other temporary or mobile service Picnic Camp Day care center Hospital Nursing home, assisted living, home care School Commercial product, no further preparation Grocery store Church, temple, or other religious location Prison, jail Other	Other
Contaminated food imported into U.S. Unknown or Undetermined	Unknown
No Data	Multiple

Appendix Table 3. Summary measures for pathogen-specific ANOVA models of outbreak size for *Salmonella*, *E. coli* O157, *L. monocytogenes*, and *Campylobacter*, based on analysis of single pathogen, single food category outbreaks – Foodborne Disease Outbreak Surveillance System, 1998–2012

	Lack-of-fit		R-squared		Significance (P-value)				Explained Variance		
Pathogen	**Degrees of Freedom**	**P-value**	**Max**	**Model**	**Overall Model**	**Food Category**	**Multi-state**	**Prep. Location**	**Food Category**	**Multi-state**	**Prep. Location**
Salmonella											
Enteritidis	21	0.88	0.32	0.20	<0.001	0.40	<0.001	<0.001	0.22	0.35	0.42
Other serotypes	76	0.45	0.47	0.33	<0.001	<0.05	<0.001	<0.001	0.28	0.54	0.18
E. coli O157	24	0.25	0.45	0.32	<0.001	<0.05	<0.001	0.26	0.27	0.58	0.15
L. monocytogenes	4	0.14	0.90	0.65	<0.05	0.09	0.15	0.55	0.55	0.33	0.12
Campylobacter	14	0.23	0.28	0.14	0.05	0.15	0.70	0.09	0.61	0.00	0.39

Appendix Table 4. Proportion of outbreak information in attribution estimates under alternative recency-weighting decay parameters

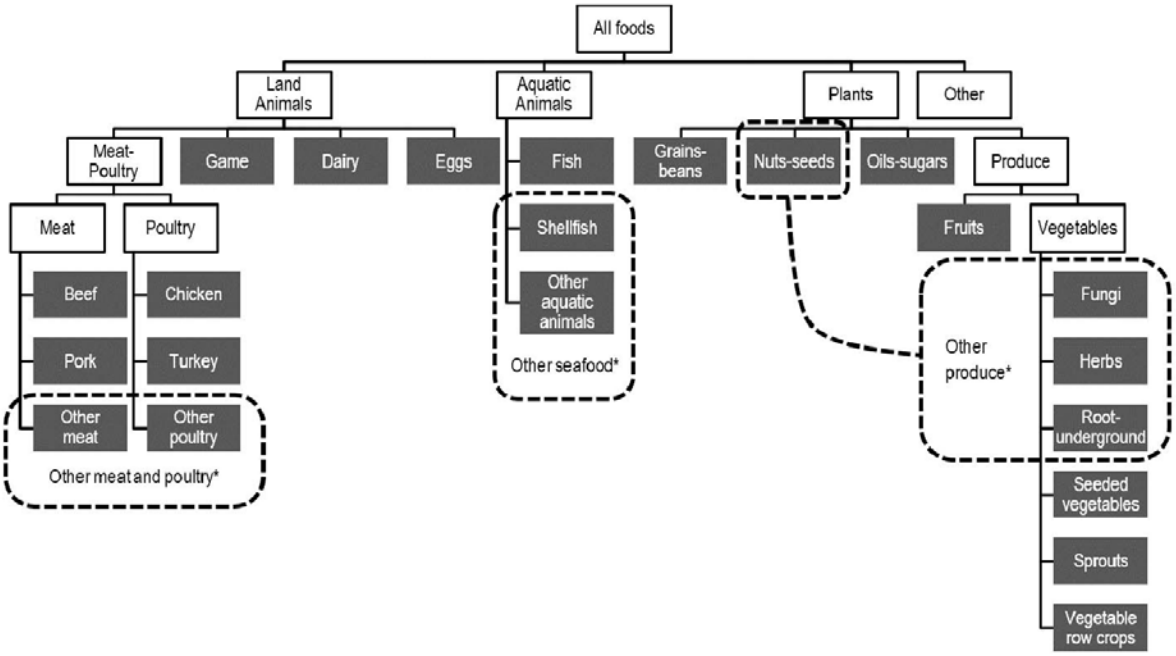
	Decay Parameter			
Years of Data	**0.20**	**0.50**	**0.71***	**0.80**
1998–2002	0%	<1%	5%	10%
2003–2007	5%	16%	28%	31%
2008–2012	95%	83%	67%	58%

* Decay parameter selected for baseline model.

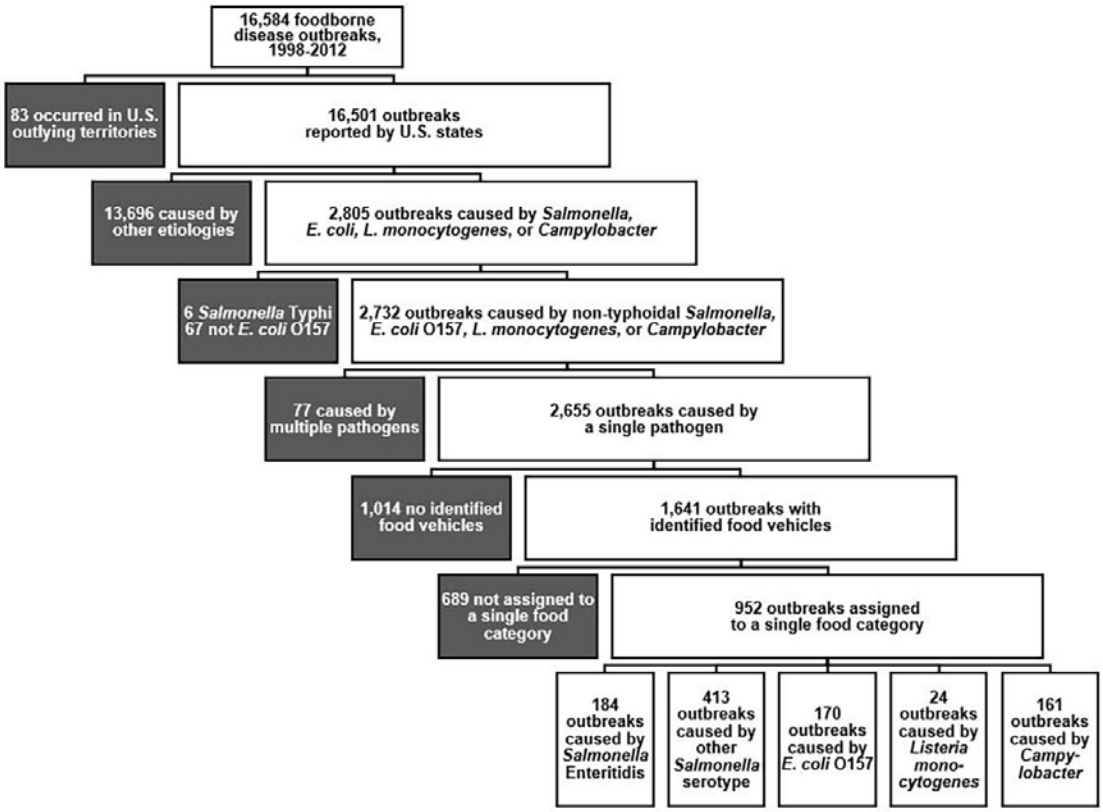
Appendix Table 5. The 5 most influential outbreaks on attribution estimates, for *Salmonella*, *E. coli* O157, *L. monocytogenes*, and *Campylobacter*, based on analysis of single pathogen, single food category outbreaks — Foodborne Disease Outbreak Surveillance System, 1998–2012*

Overall Influence Rank	Food Category	Year	Food Item(s) Implicated	Preparation setting	Multistate	No. Illnesses	Log (No. Illnesses)	Influence Metric
Salmonella								
17	Eggs	2010	shell egg, other (egg)	Multiple	Yes	1939	7.6	7.4
22	Seeded veg.	2008	jalapeno/serrano peppers, tomato	Unknown	Yes	1500	7.3	5.7
30	Other Produce	2008	peanut butter, peanut paste	Unknown	Yes	714	6.6	4.7
32	Seeded veg.	2009	ground pepper (in salami)	Other	Yes	272	5.6	4.1
40	Other Produce	2006	peanut butter	Other	Yes	715	6.6	2.9
E. coli O157								
12	Veg. Row Crops	2006	spinach	Multiple	Yes	238	5.5	9.7
14	Veg. Row Crops	2008	iceberg lettuce	Unknown	Yes	74	4.3	9.2
16	Veg. Row Crops	2011	romaine lettuce	Multiple	Yes	60	4.1	8.2
18	Veg. Row Crops	2012	romaine lettuce	Unknown	Yes	52	4.0	7.2
20	Fruits	2000	watermelon	Restaurant	No	736	6.6	6.8
L. monocytogenes								
1	Fruits	2011	cantaloupe	Private Home	Yes	147	5.0	3560.4
2	Dairy	2012	ricotta salata cheese	Multiple	Yes	23	3.1	51.4
3	Sprouts	2008	sprouts	Multiple	Yes	20	3.0	41.0
5	Dairy	2009	Mexican-Style Cheese	Private Home	Yes	18	2.9	25.8
7	Dairy	2011	blue-veined cheese, unpasteurized	Other	Yes	15	2.7	21.3
Campylobacter								
4	Seeded veg.	2008	green peas	Other	No	104	4.6	32.7
6	Other Seafood	2008	raw and steamed clams	Multiple	No	268	5.6	22.3
10	Pork	2008	pork	Other	No	27	3.3	11.7
11	Other Seafood	2010	raw clams	Other	No	68	4.2	11.4
15	Other Seafood	1998	oysters	Private Home	No	2	0.7	8.6

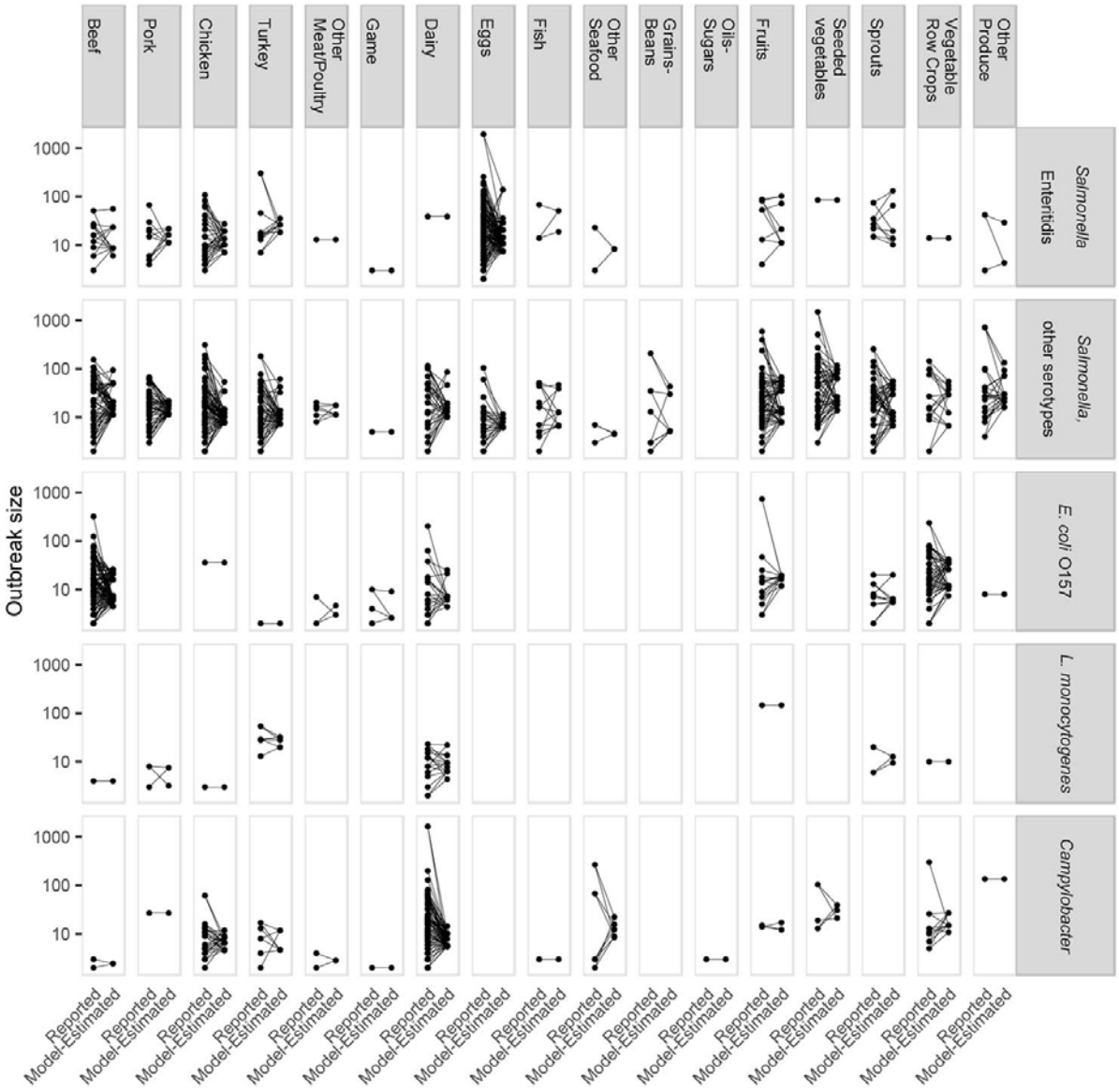
*Overall influence rank is based on the rank order of outbreaks when sorted by the influence metric, shown in the last column. The influence metric is defined as the sum of mean differences squared across all pathogen-food category pairs between the baseline estimate and an attribution estimate with that outbreak excluded.



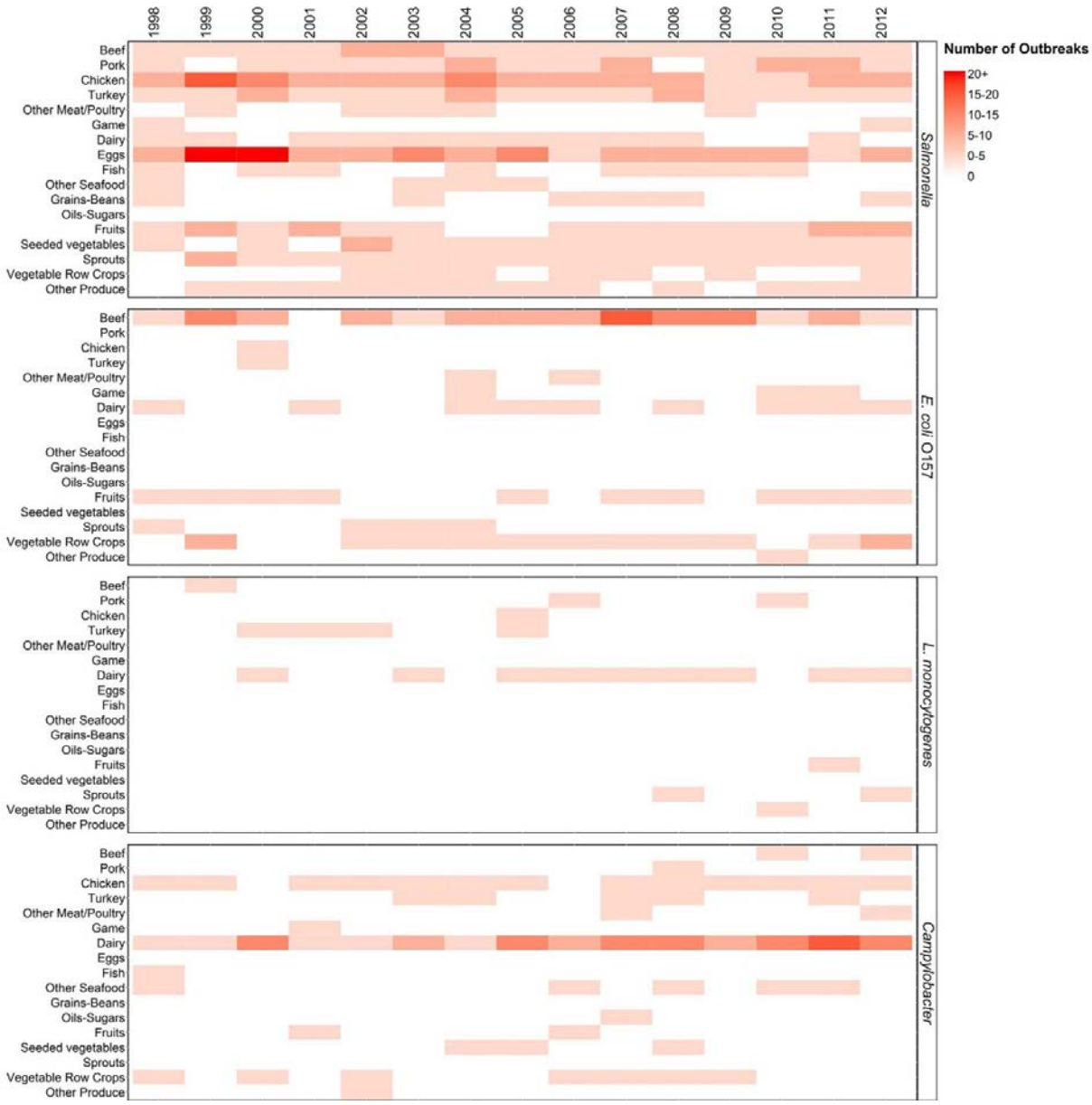
Appendix Figure 1. Hierarchical scheme used to categorize foods implicated in foodborne disease outbreaks. Outbreaks were assigned to one of 22 food categories (dark gray boxes) in the IFSAC food categorization scheme. Due to sparse data, 8 of these food categories were aggregated into 3 combined categories as indicated by the dashed-line boxes, resulting in the 17 food categories used in this analysis. “Other meat and poultry” includes animal species other than beef, pork, chicken and turkey.



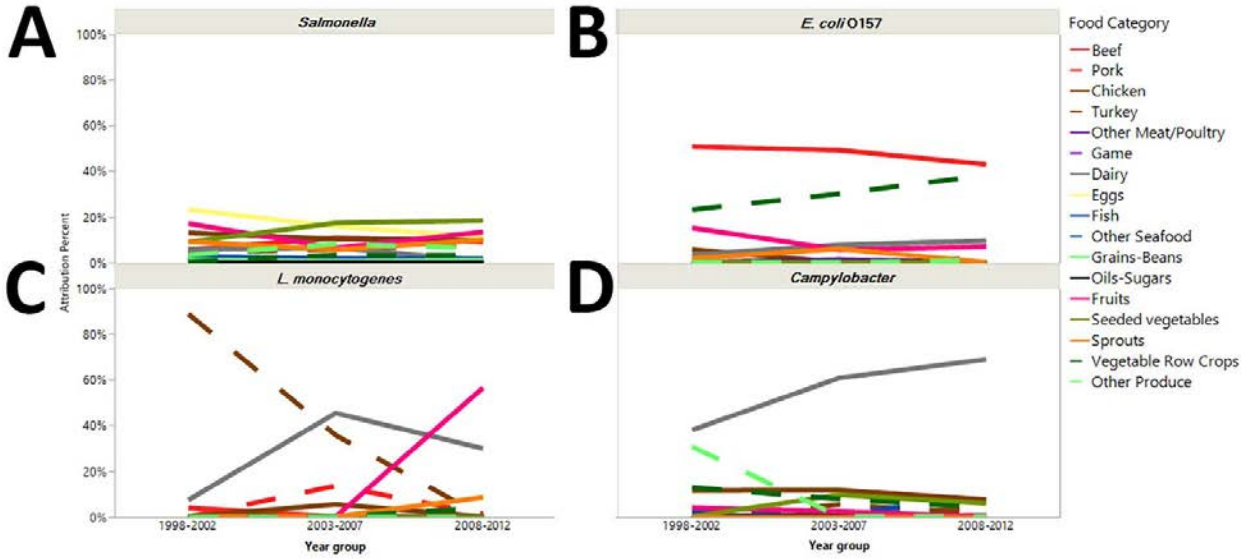
Appendix Figure 2. Data tree showing the number of outbreaks included and excluded in analysis – Foodborne Disease Outbreak Surveillance System, United States, 1998–2012.



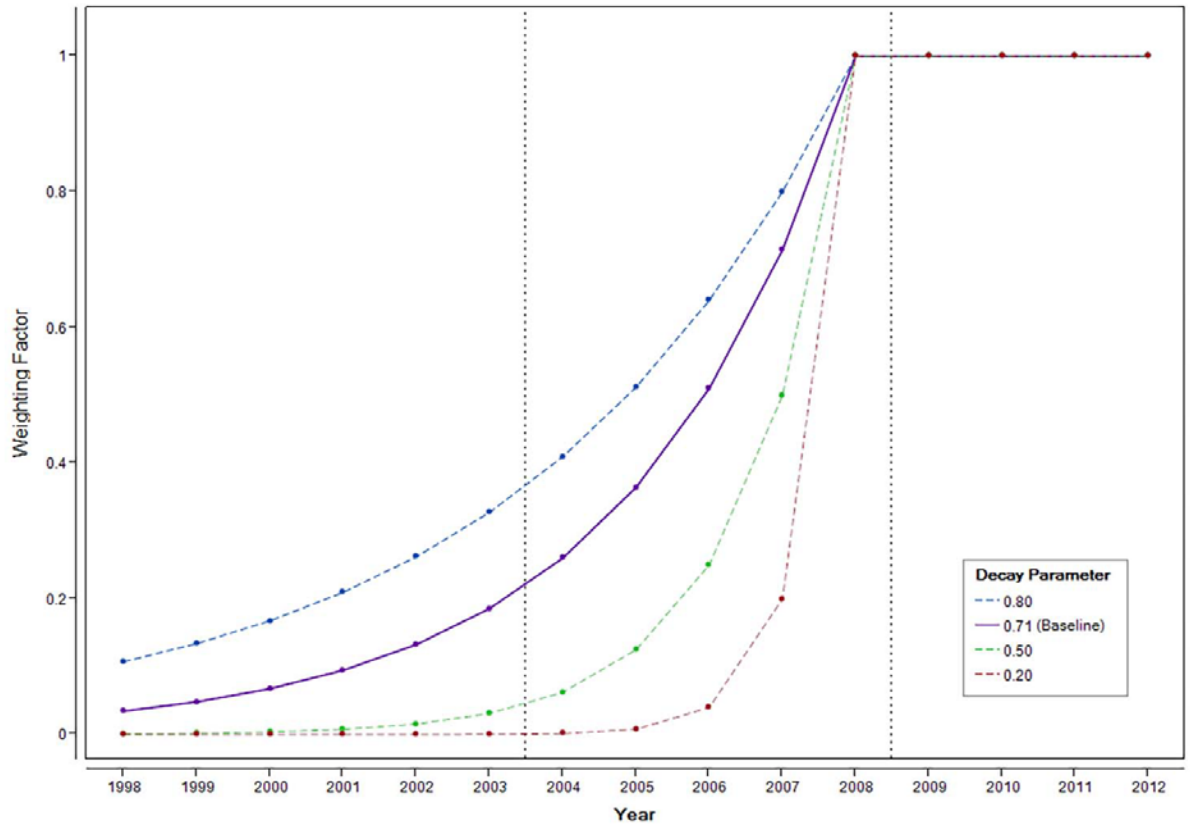
Appendix Figure 3. Comparison of reported and model-estimated number of illnesses per outbreak, by food category, for *Salmonella* Enteriritidis, other *Salmonella* serotypes, *E. coli* O157, *L. monocytogenes*, and *Campylobacter*, based on analysis of single pathogen, single food category outbreaks – Foodborne Disease Outbreak Surveillance System, 1998–2012. Each line in each panel represents a single outbreak, with a line connecting the number of reported illnesses (dot on left) with the number of model-estimated illnesses (dot on right), both presented on the same log-scale. The resulting sideways triangular shape of the combined lines in each panel illustrates the reduced variation achieved by modeling.



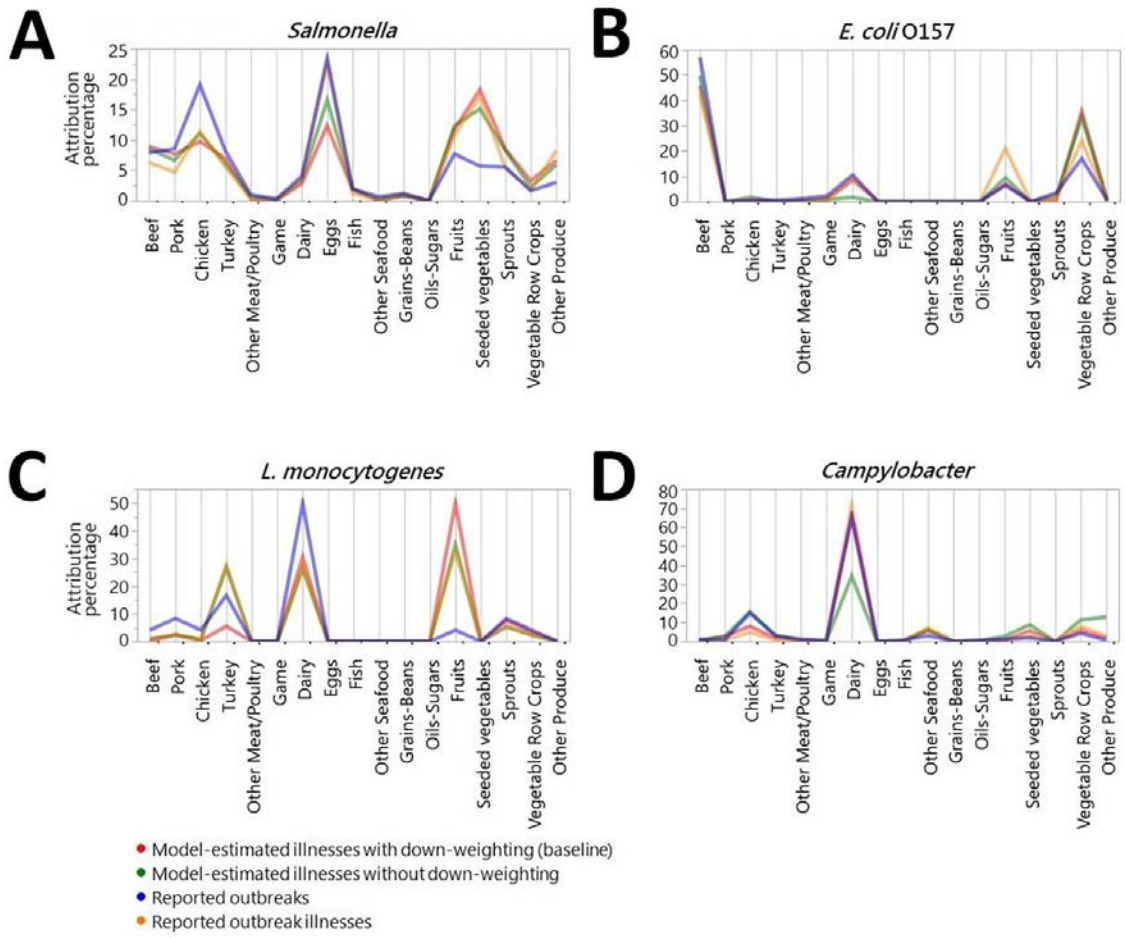
Appendix Figure 4. Number of reported outbreaks caused by a single pathogen and due to a single food category, by food category and year, for *Salmonella*, *E. coli* O157, *L. monocytogenes*, and *Campylobacter* – Foodborne Disease Outbreak Surveillance System, United States, 1998–2012.



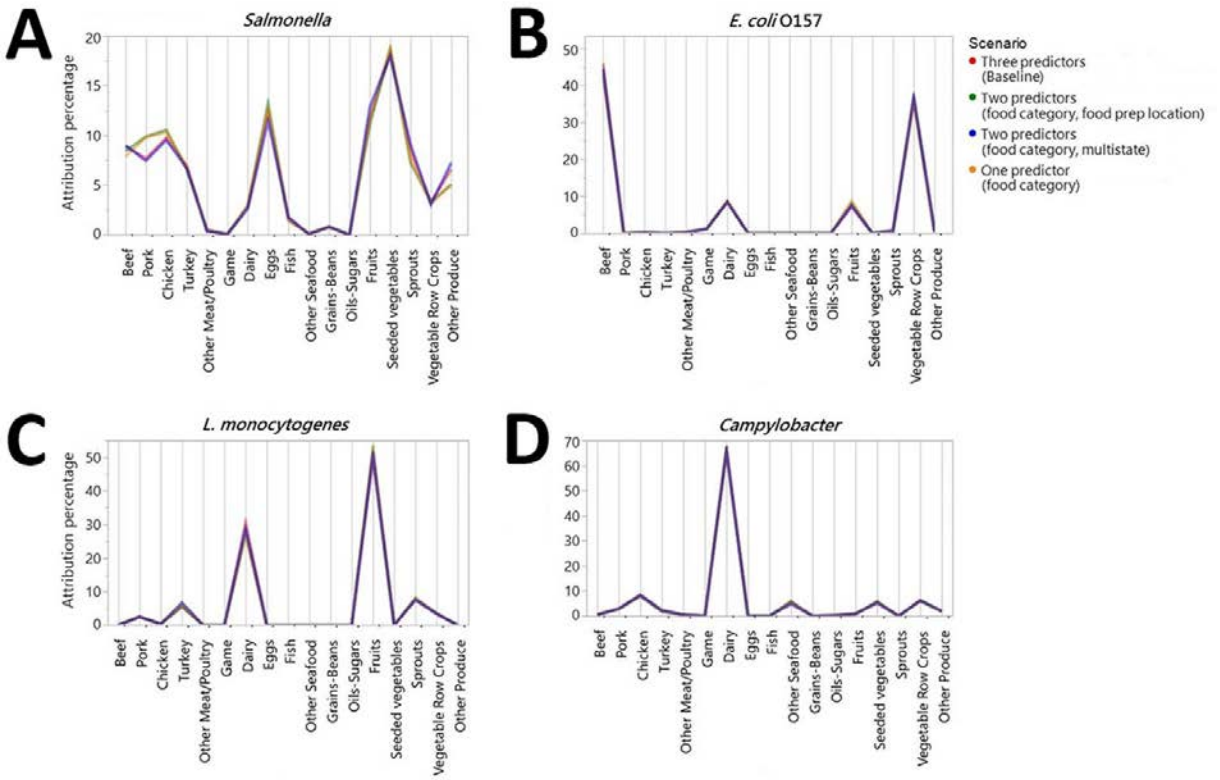
Appendix Figure 5. Estimated percentages of illnesses caused by *Salmonella*, *E. coli* O157, *L. monocytogenes*, and *Campylobacter* attributed to food categories for 5-year windows, based on analysis of single pathogen, single food category outbreaks – Foodborne Disease Outbreak Surveillance System, 1998–2012. Estimates calculated using ANOVA model-estimated outbreak illnesses for single pathogen, single food category outbreaks from 1998–2012, with down-weighting of outbreaks from 1998–2007.



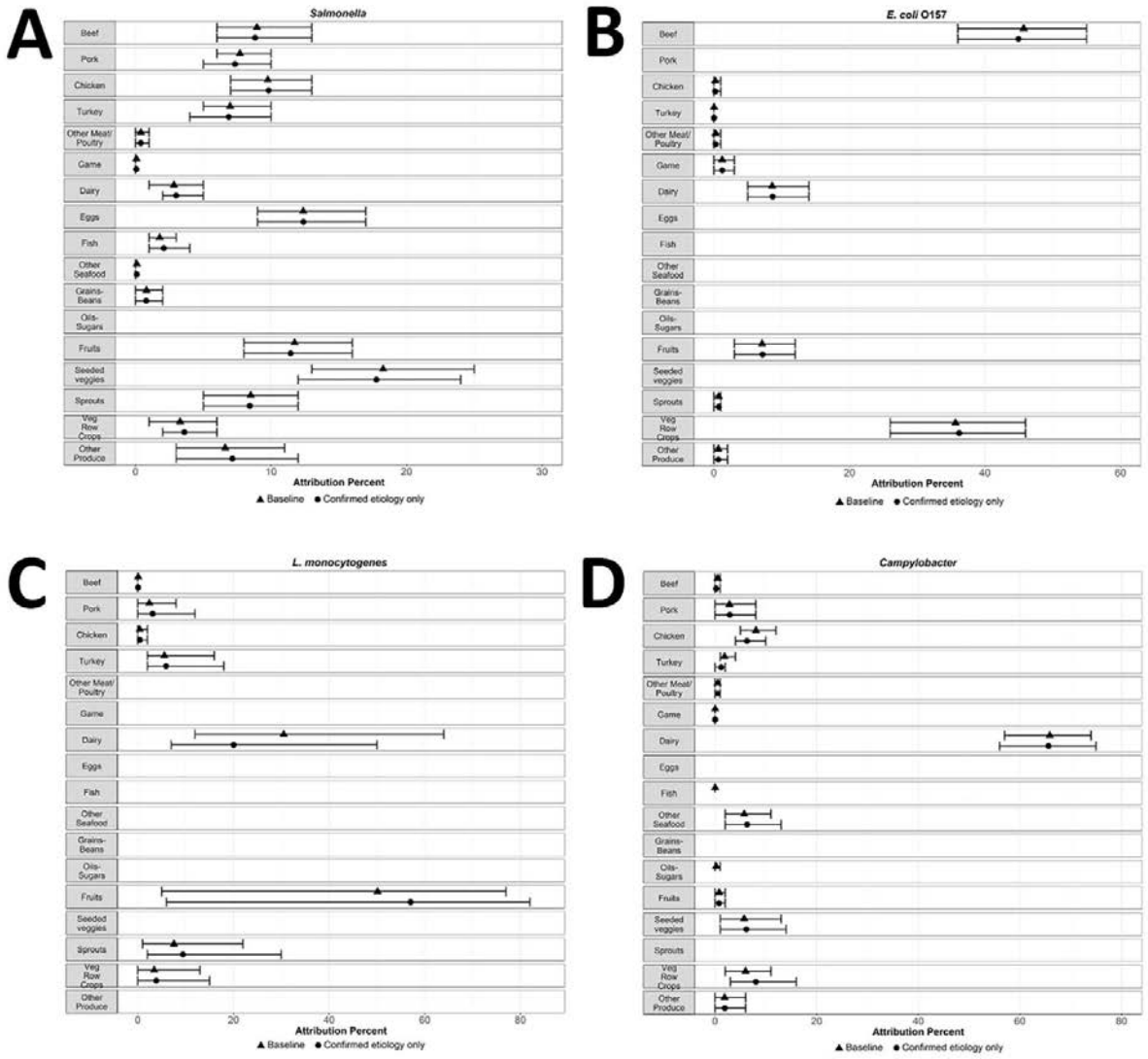
Appendix Figure 6. Comparison of multiplicative weighting factors evaluated to recency-weight model-estimated outbreak illnesses, by year and decay parameter.



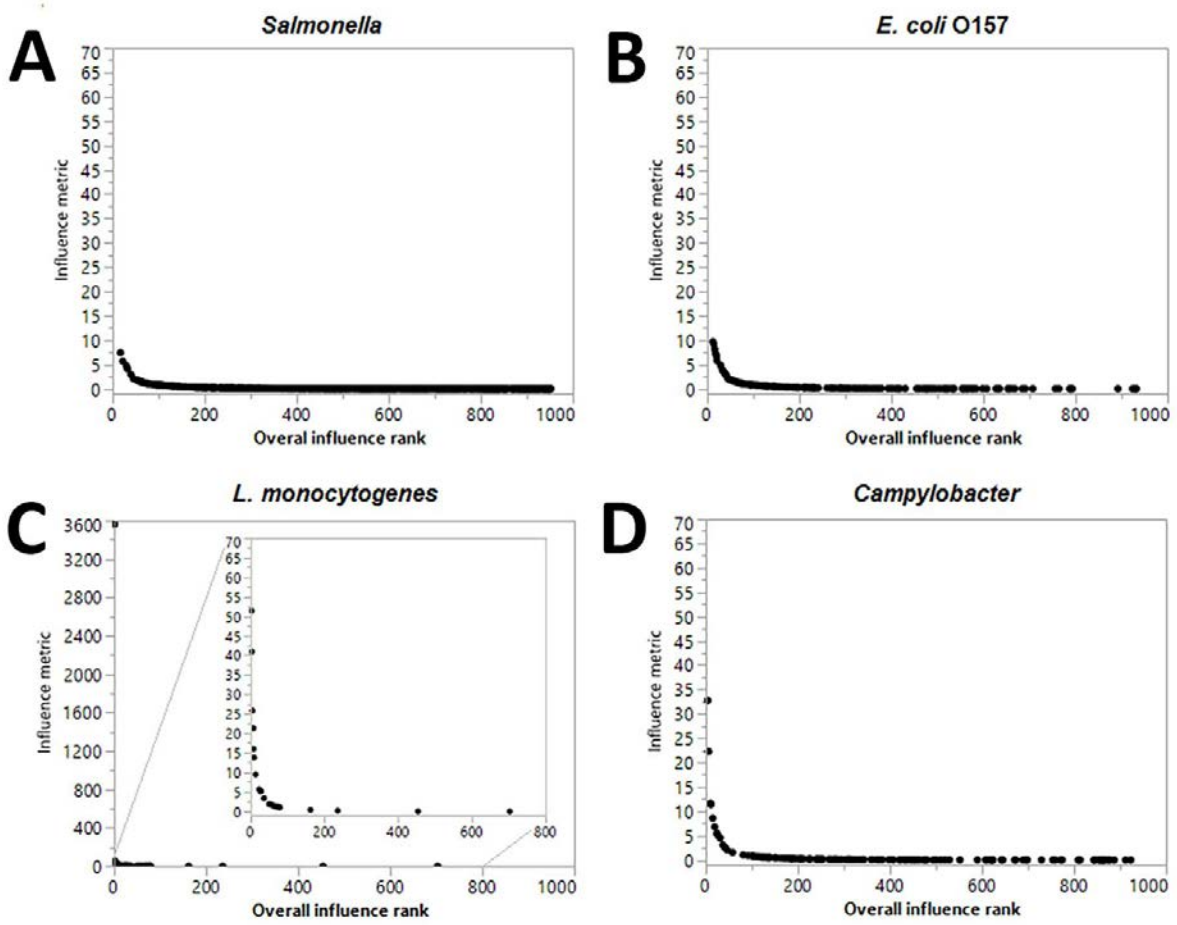
Appendix Figure 7. Comparison of measures used to calculate estimated percentages of illnesses attributed to 17 food categories for *Salmonella*, *E. coli* O157, *L. monocytogenes*, and *Campylobacter*, based on analysis of single pathogen, single food category outbreaks – Foodborne Disease Outbreak Surveillance System, 1998–2012. Each panel shows baseline estimated attribution percentages based on numbers of model-estimated illnesses with down-weighting of older outbreaks (in red). This is compared to attribution percentages based on model-estimated illnesses without down-weighting (green), and to attribution percentages calculated without any statistical modeling – namely, based on the number of reported outbreaks (purple) and number of reported outbreak illnesses (orange).



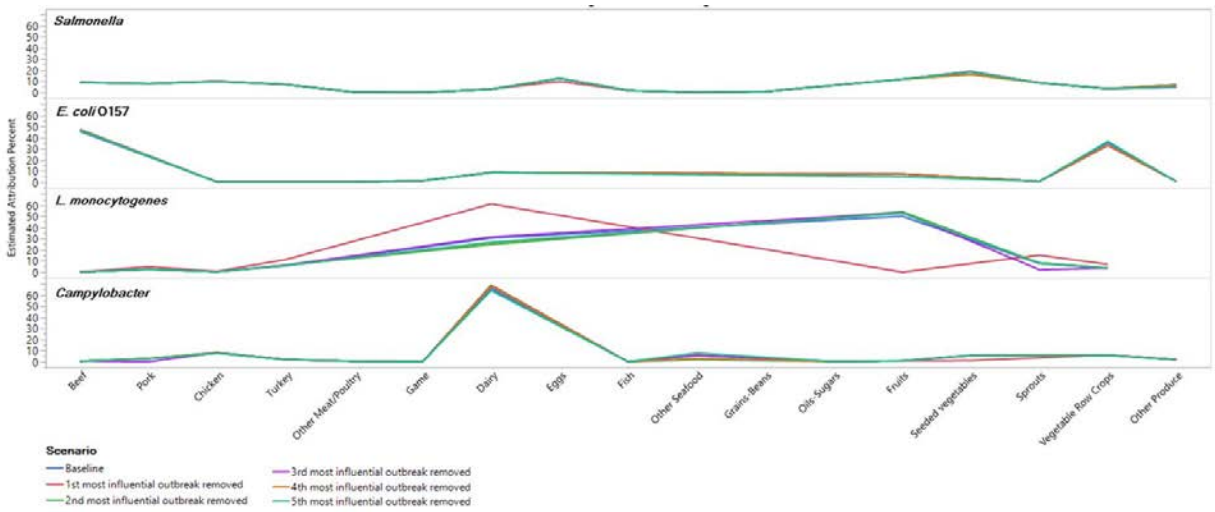
Appendix Figure 8. Estimated percentages of illnesses attributed to food categories under alternative ANOVA modeling scenarios for *Salmonella*, *E. coli* O157, *L. monocytogenes*, and *Campylobacter*, based on analysis of single pathogen, single food category outbreaks – Foodborne Disease Outbreak Surveillance System, 1998–2012. Each panel shows attribution estimates calculated using model-estimated illnesses from 4 different models for each pathogen. The baseline model specification includes food category, the type of location in which food was prepared, and whether an outbreak occurred in a single or multiple states. Credibility intervals are not shown.



Appendix Figure 9. Comparison of estimated attribution percentages (and 90% credibility intervals) for scenarios including or excluding outbreaks where etiology status is indicated as suspected, for *Salmonella*, *E. coli* O157, *L. monocytogenes*, and *Campylobacter*, based on analysis of single pathogen, single food category outbreaks – Foodborne Disease Outbreak Surveillance System, 1998–2012. Baseline estimates are based on including in the analysis all outbreaks for which etiology status is indicated as being either confirmed or suspected; the alternate scenario is based on including only those outbreaks for which etiology status is indicated as confirmed.



Appendix Figure 10. Calculated influence metric for each outbreak, for *Salmonella*, *E. coli* O157, *L. monocytogenes*, and *Campylobacter*, in descending order, based on analysis of single pathogen, single food category outbreaks – Foodborne Disease Outbreak Surveillance System, 1998–2012. The influence metric is defined as the sum of mean differences squared across all pathogen-food category pairs between the baseline set of attribution estimates and a set of attribution estimates with that outbreak excluded. The overall influence rank of each outbreak is based on the rank order of outbreaks when sorted by the influence metric in descending order. Because *L. monocytogenes* had an extreme value, an inset with the same scale as other pathogens is used to show influence metrics for all other outbreaks.



Appendix Figure 11. Impacts of excluding each of the top 5 outbreaks most influential on attribution estimates for *Salmonella*, *E. coli* O157, *L. monocytogenes*, and *Campylobacter*, based on analysis of single pathogen, single food category outbreaks – Foodborne Disease Outbreak Surveillance System, 1998–2012. Each panel shows baseline estimates of attribution percentages compared to estimates for scenarios in which the most influential outbreaks have been excluded. In each scenario, a single outbreak is excluded from the model.