Epidemiol. Infect. (2016), **144**, 396–407. © Cambridge University Press 2015
doi:10.1017/S0950268815001193

Time valuation of historical outbreak attribution data

E. D. EBEL[1], M. S. WILLIAMS[1]*, N. J. GOLDEN[1], W. D. SCHLOSSER[1] AND C. TRAVIS[2]

[1] *Risk Assessment and Analytics Staff, Office of Public Health Science, Food Safety and Inspection Service, USDA, Fort Collins, CO, USA*
[2] *Leidos, Incorporated, Reston, VA, USA*

Received 14 July 2014; Final revision 7 April 2015; Accepted 13 May 2015; first published online 22 June 2015

SUMMARY

Human illness attribution is recognized as an important metric for prioritizing and informing food-safety decisions and for monitoring progress towards long-term food-safety goals. Inferences regarding the proportion of illnesses attributed to a specific commodity class are often based on analyses of datasets describing the number of outbreaks in a given year or combination of years. In many countries, the total number of pathogen-related outbreaks reported nationwide for an implicated food source is often fewer than 50 instances in a given year and the number of years for which data are available can be fewer than 10. Therefore, a high degree of uncertainty is associated with the estimated fraction of pathogen-related outbreaks attributed to a general food commodity. Although it is possible to make inferences using only data from the most recent year, this type of estimation strategy ignores the data collected in previous years. Thus, a strong argument exists for an estimator that could 'borrow strength' from data collected in the previous years by combining the current data with the data from previous years. While many estimators exist for combining multiple years of data, most either require more data than is currently available or lack an objective and biologically plausible theoretical basis. This study introduces an estimation strategy that progressively reduces the influence of data collected in past years in accordance with the degree of departure from a Poisson process. The methodology is applied to the estimation of the attribution fraction for *Salmonella* and *Escherichia coli* O157:H7 for common food commodities and the estimates are compared against two alternative estimators.

Key words: Attributable fraction, Bayesian inference, *Salmonella*, time series.

INTRODUCTION

Every year millions of cases of foodborne illness occur globally [1]. Determining the number of pathogen-specific illnesses associated with a particular food commodity requires a measure of attribution. Population attributable fraction is the statistic that characterizes the fraction of illnesses prevented if a risk factor is removed from the population. Given the limited empirical evidence, however, a less formal definition of attributable fraction is typically employed [2], i.e. the ratio of illnesses associated with a food commodity to the total illnesses from all sources for the specific pathogen.

* Author for correspondence: Dr M. S. Williams, Risk Assessment and Analytics Staff, Office of Public Health Science, Food Safety and Inspection Service, USDA, 2150 Centre Avenue, Building D, Fort Collins, CO 80526, USA.
(Email: mike.williams@fsis.usda.gov)

https://doi.org/10.1017/S0950268815001193 Published online by Cambridge University Press

This paper will refer to that ratio as the attribution fraction.

Government food-safety authorities often rely on attribution estimates derived from annual foodborne disease outbreak investigations. Each year, government personnel investigate reported outbreaks of foodborne illness. These investigations can result in a determination of the most probable food source of an outbreak. The outbreaks in a given year attributed to a particular food commodity represent only a fraction of all foodborne illnesses associated with the commodity [3–5]. Therefore, government officials are interested in monitoring attribution fractions across time to determine if food safety efforts influence outbreak occurrences. Attribution estimates are essential for risk assessment applications that estimate the number of illnesses that might be avoided by a change in food safety policy [6].

Epidemiological investigations can provide empirical evidence of the causative link between a particular contaminated food product and human illness. No standardized approach exists for the estimation of an attribution fraction, but most studies use epidemiological data to support estimation. Examples include a microbial subtyping approach [7, 8] and the analysis of outbreak data [3–5].

In the United States, the counts of reported foodborne disease outbreaks nationwide are generally fewer than 100 per pathogen per year; with <50% of these outbreaks being successfully linked to an implicated food source. Therefore, the estimated fraction of outbreaks attributed to a general food commodity using data from a single year is highly uncertain. In addition to the small number of outbreaks per year, many outbreak surveillance systems have been operational and collecting data in a consistent manner for relatively short periods of time, so attribution fraction estimates are often based on between 2 and 10 years of data [4, 8]. To improve the precision of such estimates, analysts consider incorporating data from previous years or incorporating new data into existing estimates. Nevertheless, questions arise about the validity of including historical data in current estimates and the number of years of data to include in such estimates.

While it is possible to make inferences using only data from the most recent year, such an estimation strategy would ignore the data collected in previous years. Thus, an argument exists for an estimator of the current attribution fraction that could ‘borrow strength’ from data collected in the previous years. The origins of this concept date back to the rolling sample designs in the late 1950s that combine data from a set number of the most recent years of data [9].

An equally weighted moving average estimator is one of the simplest solutions, where a certain number of the most recent years of data are combined without consideration of the age of the data. This estimator usually exhibits the smallest mean square error among rolling sample estimators when applied to populations with small trend components [10], but the number of years of data to use remains subjective. More sophisticated estimation methods, such as time series and non-parametric smoothers, could be employed [10], but annual outbreak counts rarely demonstrate autocorrelation and many surveillance systems are relatively new and lack sufficient temporal data for reliable estimation [11].

In economics, the concept of the time-value of money is well established. For example, discounting is used to adjust future cash values to a present value to account for the opportunity to earn on today’s money in the future. Because data collected in the past may be less valuable than data collected more recently, the information supplied by historical surveillance data should also be adjusted for its time value [12, 13]. Nevertheless, the practice of valuing historical data often ranges between extremes – using current data and ignoring past data or accumulating and treating equally all past data. At its roots, the latter approach is Bayesian in that the analyst is sequentially updating data across time; the posterior distribution estimated for time t is proportional to the likelihood at time t multiplied by the prior distribution informed by time $t-1$. This posterior distribution itself becomes a prior for inferences about time $t+1$.

If data represent a stochastic process that is stationary (i.e. the rate at which events occur across time is constant), then Bayesian thinking is correct. If data are accumulated across years to estimate a rate parameter that is fundamentally unchanging, then the current inference about the parameter should be more certain than past inferences based on less data. This is similar conceptually to the confidence we gain about the fairness of a coin as we observe an increasing number of flips of the coin with the fraction of heads among flips stabilizing at 0·50.

In the case of a stationary Poisson process, no justification exists for discounting past data, because those data are just as relevant to our estimate as current data. This treatment is also theoretically

https://doi.org/10.1017/S0950268815001193 Published online by Cambridge University Press

appropriate because the sum of independent Poisson random variables is itself a Poisson random variable, where the rate parameter is the sum of the previous observations. Therefore, the observed count of outbreaks can be averaged to model the annual rate.

Epidemiological data are sometimes stationary but often dynamic. If historical surveillance data refer to a period of time during which the underlying infection process is markedly different from the current period, then the historical data may have minimal relevance to estimating current disease occurrence. For example, substantial past surveillance evidence for disease freedom within a population may have no relevance to current inferences if the infectious agent only recently entered the population. Alternatively, successful intentional efforts to control an infectious agent would make past evidence about occurrence less relevant to an inference about its occurrence today.

This paper outlines an approach to estimating attribution fractions when the available data are limited. The method assumes a simple point process model to discount the historical data according to objective statistical principles. This approach is illustrated using *Salmonella* and *Escherichia coli* O157:H7 outbreak data from the United States that are attributed to common food commodities.

Data description

State and local health departments report foodborne disease outbreaks to the United States Centers for Disease Control and Prevention (CDC) through the Foodborne Disease Outbreak Surveillance System (FDOSS) [14]. Reports include, when available, the number of persons ill, the outbreak aetiology, a description of the implicated food vehicle(s), lists of ingredients, and identification of the contaminated ingredient(s). Each outbreak is assigned to one of CDC's food commodity classes [3]. This study only includes outbreak data beginning in 1998, because 1998 was the first year when detailed information on food ingredients was available. The last year for which data are available was 2011. For this analysis, all outbreaks were included where a single aetiological agent and implicated food vehicle were identified. The number of outbreak counts per year attributed to *Salmonella* and *E. coli* O157:H7 for the CDC food commodity classes are summarized in Tables 1 and 2.

The outbreak counts for each commodity represent time-series data, and illness and outbreak counts for both of these pathogens demonstrate strong seasonal patterns [15, 16]. A summarization of temporal patterns is beneficial for motivating the chosen estimation strategy. Significant temporal patterns in annual outbreak counts would suggest that time-series methods could be used in the estimation of attribution fractions. An assessment of temporal patterns was performed using ARMA(*p,q*) models, where (*p,q*) are the autoregressive and moving average orders, respectively [17]. The best-fitting ARMA model, based on Akaike's Information Criterion, was determined for each commodity to demonstrate the general lack of annual patterns in the data (Tables 1 and 2). For 27 out of the 29 pathogen–commodity pairings, the best-fitting model is an ARMA($p = 0$, $q = 0$). The only two commodities to demonstrate significant temporal patterns were the eggs and grains/beans commodities for *Salmonella*, where the best-fitting model is an ARMA(0,1). The mean and variance of the outbreak counts for each commodity–pathogen pair are given in Table 3. The pathogen–commodity classes of *Salmonella*–eggs, *Salmonella*–poultry, and *E. coli* O157:H7–beef represent roughly half of all outbreaks for the two pathogens.

METHODS

The proposed method is based on the observation that if the number of outbreaks were a completely random process, the annual number of outbreaks would follow a Poisson distribution. In this situation, the appropriate estimator of the annual rate parameter is the sum of annual outbreaks divided by the total number of years, because the sum of a Poisson-distributed random variable is itself a Poisson random variable [18]. This implies that the data from previous years are just as relevant as the current year's data and that the data from all years should be averaged to estimate the attribution fraction. If the number of outbreaks does not follow a Poisson distribution, the data from the previous years are less relevant, and their influence on the current year's attribution fraction should be reduced. This discounting is accomplished by reducing the effective number of outbreaks as a function of the number of years since their occurrence. For example, if seven outbreaks occurred at time $t - 2$, the estimator would reduce this number to something less than 7 as a function of the degree of departure of the counts from a Poisson random variable.

In the first step of our analysis, we assess annual counts of outbreaks associated with a particular

https://doi.org/10.1017/S0950268815001193 Published online by Cambridge University Press

Table 1. *Salmonella* outbreak counts from 1998 to 2011 for those outbreaks that identified a food commodity *[14]*

Commodity	ARMA	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Beef	(0,0)	1	3	2	3	6	9	4	4	3	3	3	4	1	2
Crustacean	(0,0)	1	0	0	0	0	1	1	1	0	0	0	0	1	0
Dairy	(0,0)	3	2	0	3	3	1	2	3	3	4	1	0	0	1
Eggs	(0,1)	8	24	22	9	7	12	9	12	3	5	7	7	7	2
Finfish	(0,0)	1	0	5	1	0	0	2	0	0	1	2	2	1	0
Fruits/nuts	(0,0)	1	7	4	6	3	4	0	1	4	1	5	2	3	8
Fungus	(0,0)	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Game	(0,0)	1	1	0	0	2	1	0	0	0	0	0	1	0	0
Grains/beans	(0,1)	1	1	0	0	0	1	1	0	0	1	1	0	0	0
Leafy	(0,0)	0	1	0	1	2	1	2	1	1	2	0	2	1	0
Mollusc	(0,0)	0	0	0	0	0	1	0	1	1	0	0	0	0	0
Oil/sugar	(0,0)	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Pork	(0,0)	3	2	6	6	3	3	6	3	4	7	4	6	6	11
Poultry	(0,0)	13	21	17	14	14	12	24	12	10	7	11	4	8	12
Root	(0,0)	0	0	2	0	0	0	2	2	1	0	0	0	1	0
Sprout	(0,0)	1	6	2	3	1	2	1	1	1	3	1	4	4	3
Vine	(0,0)	1	0	1	0	5	1	2	3	4	2	3	4	4	3
Grand total	(0,0)	35	68	61	46	47	49	56	44	35	37	38	36	37	42

Table 2. *E. coli* O157:H7 outbreak counts from 1998 to 2011 for those outbreaks that identified a food commodity *[14]*

Commodity	ARMA	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Beef	(0,0)	1	14	9	0	8	1	7	5	6	15	12	11	4	5
Dairy	(0,0)	2	0	0	1	0	1	1	1	3	1	3	1	7	2
Fruits/nuts	(0,0)	1	1	3	1	0	0	1	1	1	1	1	0	2	2
Game	(0,0)	0	0	0	0	0	0	2	0	1	0	1	0	2	2
Grains/beans	(0,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Leafy	(0,0)	1	5	0	1	2	2	1	1	4	2	4	3	1	3
Mollusc	(0,0)	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Poultry	(0,0)	0	0	2	0	0	0	0	0	0	0	0	1	0	0
Sprout	(0,0)	1	0	0	0	1	3	1	0	0	0	0	0	0	2
Vine	(0,0)	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Grand total	(0,0)	7	20	14	3	11	7	14	8	15	19	21	16	16	17

commodity to determine if these follow a random Poisson process. A χ^2 goodness-of-fit test was completed to determine if a commodity's annual outbreak counts followed a Poisson process [19]. The Poisson test statistic is

$$\frac{\sum_{t=1}^{T}(s_t-\bar{s})^2}{\bar{s}},$$

where s_t is the number of outbreaks in year t and $\bar{s}$ is the mean number of outbreaks per year. The statistic is assumed to be distributed according to a $\chi^2(T-1)$, where T is the number of years of data considered (e.g. 14 in this study).

Commodities for which the χ^2 test failed to reject the hypothesis that outbreak counts follow a Poisson process were ignored for the remainder of this analysis, because no discounting of that data was necessary. Commodities for which the χ^2 test rejected the null hypothesis were assumed to be overdispersed. Overdispersion relative to a Poisson distribution suggests that the variability in annual counts is larger than that explained by a Poisson random variable.

The second step in our analysis applies to commodities with overdispersed outbreak counts. This step estimates discount rates for the historical data that inform the attribution fractions for those commodities. We do this by interpreting the parameters of fitted beta distributions as measures of implied sample sizes (or sample values)

https://doi.org/10.1017/S0950268815001193 Published online by Cambridge University Press

Table 3. *Summary statistics regarding outbreak counts. The Poisson test P values for significant departure from a Poisson stationary process are shown*

	Salmonella			*E. coli* O157:H7		
	Outbreak counts per year			Outbreak counts per year		
Commodity	Mean	Variance	*P* value	Mean	Variance	*P* value
Beef	3·4	4·3	0·24	7·0	22·9	0·00
Crustacean	0·4	0·2	0·77			
Dairy	1·9	1·8	0·47	1·6	3·3	0·02
Eggs	9·6	40·4	0·00			
Finfish	1·1	1·9	0·04			
Fruit/ nuts	3·5	5·8	0·06	1·1	0·7	0·82
Fungus	0·1	0·1	0·45			
Game	0·4	0·4	0·47	0·6	0·7	0·22
Grains/beans	0·4	0·3	0·84	0·1	0·1	0·45
Leafy	1·0	0·6	0·84	2·1	2·1	0·45
Mollusk	0·2	0·2	0·61	0·1	0·1	0·45
Oil/sugar	0·1	0·1	0·45			
Pork	5·0	5·5	0·35			
Poultry	12·8	27·7	0·01	0·2	0·3	0·09
Root	0·6	0·7	0·22			
Sprout	2·4	2·4	0·43	0·6	0·9	0·10
Vine	2·4	2·6	0·37	0·1	0·1	0·45
Grand total	45·1	107·3	0·00	13·4	29·8	0·01

and evaluating the ratio of the sample value for the overdispersed case to the sample value for a Poisson-distributed case. Because the sample value in the overdispersed case is always less than the Poisson-distributed case, this generates a discount rate of <1.

Previous examples for discounting historical information have assumed the specific process by which overdispersion might occur [12, 13]. Conceptually, historical data that measures the occurrence of a phenomenon in the past should have its contribution to a present-day estimate depreciated by its proportional reduction in sample value. Therefore, a sample collected at time $t-1$, of size n_{t-1}, would have its sample value reduced to $(1-d) \times n_{t-1} = n_{t|t-1}$, where d is a discount factor and $n_{t|t-1}$ is the value of the $t-1$ sample at time t. This current approach does not explain the process by which overdispersion might occur. We simply estimate $d = 1 - (n_{t|t-1}/n_{t-1})$ from available data by assuming that the overdispersed sample value $(n_{t|t-1})$ reflects the annual reduction in sample value relative to its value if it were generated by a stationary Poisson process (n_{t-1}).

This approach relies on a common interpretation of the beta(a,b) distribution's parameters as $a = s + 1$ and $b = n - s + 1$ where s represents the counts of some characteristic among a sample of size n. This parameterization is a common application of Bayes Theorem [20, 21] such that this posterior beta distribution results from a binomial distribution likelihood function and a conjugate prior beta(1,1) distribution.

The annual attribution fraction for a particular commodity, α_f (where f indexes commodities), is a random variable that reflects the influence of the year-to-year variability in counts and the uncertainty inherent in an estimate based on sampling evidence. In our analysis, we parameterize beta$(s_t + 1, n_t - s_t + 1)$ to simulate the sampling uncertainty about α_f, where s_t is a number of outbreaks attributed to commodity f for a given year, and n_t is the total count of outbreaks in a given year. As discussed next, there are two approaches to deriving values of s_t and n_t, but both require Monte Carlo simulations that generate random variables as outcomes. Because these simulated outcomes are generated from a beta distribution, we assume the mean and variance of these outcome variables can themselves be fitted to a beta(a,b) distribution. For the sake of convenience, the method-of-moments approach is used where:

$$a = E[\alpha_f]\left(\frac{E[\alpha_f](1 - E[\alpha_f])}{V[\alpha_f]} - 1\right)$$

$$b = (1 - E[\alpha_f])\left(\frac{E[\alpha_f](1 - E[\alpha_f])}{V[\alpha_f]} - 1\right),$$

and $E[\alpha_f]$ and $V[\alpha_f]$ are the mean and variance estimates from the Monte Carlo simulation. The parameters of this fitted beta(a,b) are then re-interpreted as a measure of the underlying effective sample value because $a + b = s + n - s = n$ [ignoring the 1's contributed by the original beta(1,1) prior].

For the commodities where the assumption of the Poisson process is rejected, there are two alternatives for α_f to consider. The first, $\alpha_{f|\text{Poisson}}$, reflects a distribution that would result if an underlying Poisson process for outbreak counts were true. The second, $\alpha_{f|\text{data}}$, reflects the distribution generated from the actual overdispersed data. We reason that the effective sample value for $\alpha_{f|\text{Poisson}}$ is greater than $\alpha_{f|\text{data}}$ because the variance of $\alpha_{f|\text{Poisson}}$ is less than the variance of $\alpha_{f|\text{data}}$. This is obvious if we use a beta$(s, n - s)$ parameterization and recognize that s is some fraction, r, of n (i.e. $s = r \times n$), where in our case r is the same for either the Poisson or overdispersed case because both random variables will have roughly the same expected value. It can be shown that the resulting variance of the beta distribution is just $r - r^2/n$ (i.e. if $\alpha \sim$ beta$(s,n - s)$ then

$$\text{Variance }(\alpha) \approx \frac{s(n-s)}{n^3} = \frac{rn(n-rn)}{n^3} = \frac{r-r^2}{n}.$$

Therefore, for a fixed value of r, the variance of the beta distribution is inversely proportional to the sample size n.

To estimate $\alpha_{f|\text{Poisson}}$, we implement the following Monte Carlo simulation:

(1) Sample $s_t \sim$ Poisson(λ_f) and $n_t \sim$ Poisson(λ_n), where λ_f and λ_n are the average annual number of commodity-specific outbreaks and total outbreaks, respectively, estimated from the 14 years of data. Because s_t and n_t are necessarily dependent we correlate the Poisson random draws of s_t and n_t based on their estimated correlation coefficient [using the RiskCorrel and RiskCorrMat functions in @Risk (Palisade Corp., USA)].

(2) On each iteration, the simulated s_t and n_t are entered as parameters in a beta$(s_t + 1, n_t - s + 1)$ distribution from which a random draw is simulated.

(3) This process is repeated for 100000 iterations to generate an output for $\alpha_{f|\text{Poisson}}$. The mean and variance of this output is back-fit to a beta distribution, and the a and b parameters are interpreted as above.

To estimate $\alpha_{f|\text{data}}$ from the available overdispersed data, the process is the same as above except random pairs of s_i and n_i for a commodity are selected iteratively across the 14 years of data (i.e. the first element above is replaced with simple random sampling of the s_i and n_i pairs from the data). All simulation work was performed in @Risk (Palisade Corp.) and then validated using R [22].

Let $\alpha_{f|\text{data}} \sim$ beta$(a_{\text{data}}, b_{\text{data}})$ be the estimate resulting from simulating s and n from the empirical evidence in Tables 1 or 2 for a particular food commodity. Let $\alpha_{f|\text{Poisson}} \sim$ beta$(a_{\text{Poisson}}, b_{\text{Poisson}})$ be the estimate resulting from simulating s and n as correlated Poisson random variables. If the data follow a Poisson assumption, then

$$\alpha_{f|\text{data}} \approx \alpha_{f|\text{Poisson}} \text{ and } \frac{a_{\text{data}} + b_{\text{data}}}{a_{\text{Poisson}} + b_{\text{Poisson}}} \approx 1$$

(i.e. both distributions result from the same Poisson process and, therefore, must be equivalent). This would be the result if the commodities eliminated in step 1 of our analysis (i.e. those for which the Poisson assumption was not rejected) were analysed in step 2. Nevertheless, for the commodities in which the Poisson assumption was rejected (i.e. the data reflect a process that is overdispersed), then

$$\frac{a_{\text{data}} + b_{\text{data}}}{a_{\text{Poisson}} + b_{\text{Poisson}}} < 1,$$

and the results imply a reduction in sample value. In this latter case, the effective sample size implied by the Poisson assumption will be larger than that of the actual data. The difference in this ratio suggests how annual attribution fractions might vary beyond a stationary process.

Discounting historical data

We assume the annual discount rate applicable to historical data is

$$d = 1 - \frac{a_{\text{data}} + b_{\text{data}}}{a_{\text{Poisson}} + b_{\text{Poisson}}}.$$

This is the magnitude of adjustment applied to data collected 1 year previously. It reflects how much the effective sample size at time $t{-}1$ is reduced at time t, because the process varies beyond what is expected of a Poisson process. For data collected at time $t = 1$ (n_1), its value at time t (n_t) is computed as $n_t = n_1(1{-}d)^{t-1}$ based on standard compounding methods.

The following example illustrates the calculations needed to discount surveillance data to estimate α_f for the current year. Assume we have 3 years of data so that $t = 1,2,3$, where $t = T = 3$ is the current year.

https://doi.org/10.1017/S0950268815001193 Published online by Cambridge University Press

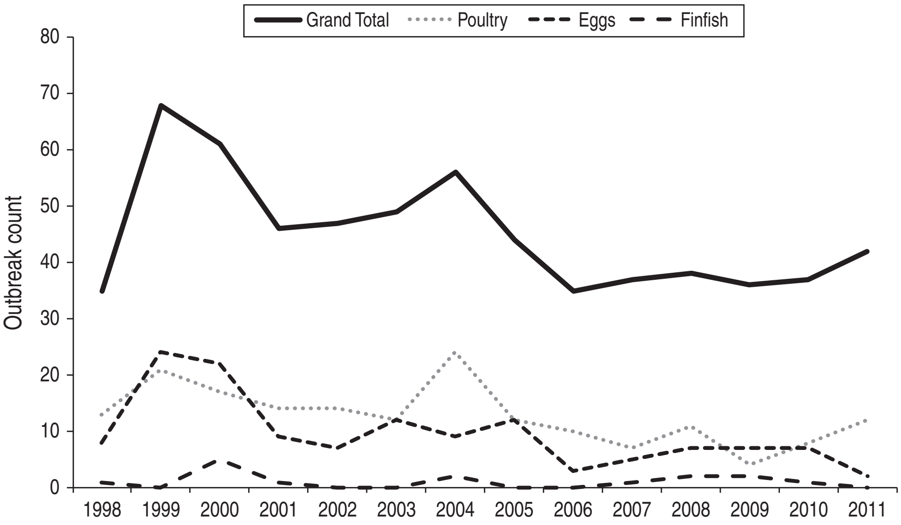

Fig. 1. Time-series data for *Salmonella* outbreak counts.

The estimator $\alpha_{f,t} = 3$, is modelled as:

$$\alpha_{f,t=3} \sim \text{beta}(s^* + 1, n^* - s^* + 1),$$

$$\text{where } s^* = s_{t=1}(1-d)^{3-1} + s_{t=2}(1-d)^{3-2} + s_{t=3}(1-d)^{3-3}$$

and

$$n^* = n_{t=1}(1-d)^{3-1} + n_{t=2}(1-d)^{3-2} + n_{t=3}(1-d)^{3-3}.$$

Using the commodity-specific outbreak counts yield the general formula

$$s^* = \sum_{t=1}^{T} (1-d)^{T-t} s_t.$$

Note that the estimator for s^* and n^* resembles the estimator for an exponential smoother, which can also be used to describe the time-value of historical information [23]. Where the exponential smoother differs is the lack of the normalizing constant d and the necessary selection of an initial value for the smoothing process [11, 24]. While an estimator based on the exponential smoother was tested in this study, the results are not presented because the selection of the initial value adversely affected the performance of the estimator. This effect reflects the limited number of years of data.

Examples

The examples compare the implications of estimating α_f using just the current year of data, aggregating all prior years of data without discounting, and aggregating all prior years of data using the proposed discounting methodology. The methodology is applied to the estimation of attribution fractions for the United States. The selected pathogen–commodity pairs for *Salmonella* and *E. coli* O157:H7 are ones with outbreak counts that exhibit a significant overdispersion relative to the Poisson distribution.

RESULTS

The Poisson test finds that the *Salmonella* outbreak counts for eggs, finfish, and poultry are not Poisson distributed (Table 3). Similarly, this test finds that the *E. coli* O157:H7 outbreak counts for beef and dairy are not Poisson distributed. In these cases, the variance is greater than the mean such that the data appear overdispersed relative to a Poisson random variable.

The *Salmonella* outbreak data demonstrate that poultry and egg outbreak counts are highly correlated with the total outbreak count per year (Fig. 1). In both cases, a Pearson correlation of $r > 0{\cdot}8$ was calculated. By contrast, counts of finfish outbreaks are poorly correlated with total outbreak counts ($r = 0{\cdot}2$). The *E. coli* O157:H7 outbreak data (Fig. 2) demonstrate a high correspondence between beef outbreak and total outbreak counts ($r = 0{\cdot}82$), but little correspondence exists between dairy and total outbreak counts ($r = 0{\cdot}2$).

Table 4 presents the discount rate for those pathogen–commodity pairs whose outbreak counts are significantly overdispersed. The largest discount rate applies for *Salmonella*–eggs and *E. coli* O157:H7–beef. In both cases, the resulting respective d values of 0·5 and 0·6 imply that prior evidence concerning these pathogen–commodity pairs should be sharply discounted such that data much older than 4 years

https://doi.org/10.1017/S0950268815001193 Published online by Cambridge University Press

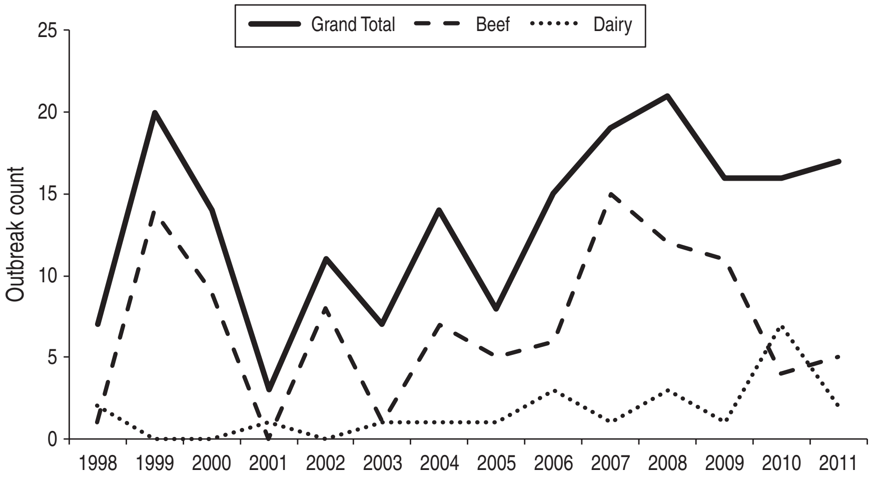

Fig. 2. Time-series data for *E. coli* O157:H7 outbreaks.

Table 4. *Summary results for determining the discount rates for annual outbreak data are shown for those commodities with overdispersed outbreak count data*

Pathogen (commodity)	Model	Attribution fraction mean	Attribution fraction variance	Effective sample value	Annual discount rate, *d*
Salmonella (poultry)	Poisson	0·29	0·007	30	0·3
	Overdispersed	0·29	0·009	22	
Salmonella (eggs)	Poisson	0·22	0·005	31	0·5
	Overdispersed	0·21	0·010	16	
Salmonella (finfish)	Poisson	0·04	0·001	30	0·1
	Overdispersed	0·04	0·002	27	
E. coli O157:H7 (beef)	Poisson	0·51	0·026	9	0·6
	Overdispersed	0·48	0·052	4	
E. coli O157:H7 (dairy)	Poisson	0·18	0·017	7	0·3
	Overdispersed	0·19	0·026	5	

will have little influence on current estimates. For example, a discount rate of 0·6 implies that data from four years ago will only be worth about 4% of its original value, i.e. $(1 - 0{\cdot}6)^4 \approx 0{\cdot}04$. The lowest discount rate was applicable to *Salmonella*-finfish. This discount rate of 0·1 suggests that decade-old data may still have relevance to current estimates; in this case $(1 - 0{\cdot}1)^{10} \approx 0{\cdot}37$ such that 10-year-old data maintains 37% of its value.

Two applications of the computed discount rates are illustrated in Figures 3 and 4. In the case of *Salmonella*–poultry (Fig. 3), we see that each year's attribution fraction estimate changes somewhat with sharp deviations occurring (e.g. see 2004 and 2009 in Fig. 3). If the outbreak data are aggregated sequentially without any discounting, then the limits of the credible interval about the attribution fraction get progressively tighter across time. By contrast, if the outbreak data are discounted sequentially, then the credible intervals initially become tighter but stabilize and remain wider than the undiscounted illustration. As expected, the credible intervals for the aggregated data (undiscounted or discounted) are narrower than the intervals based on each year's data; although these limits are exactly the same in 1998 because no aggregation has occurred in the first year.

In Figure 3, comparing the results for 2009 illustrates the statistical merit of discounting aggregated data. In that year, the observed fraction of poultry-associated outbreaks was about 0·13. Nevertheless, the credible interval about that estimate overlaps with the credible interval for the aggregated with discount estimate. By contrast, no overlap exists with the credible interval estimated by aggregating without discounting.

https://doi.org/10.1017/S0950268815001193 Published online by Cambridge University Press

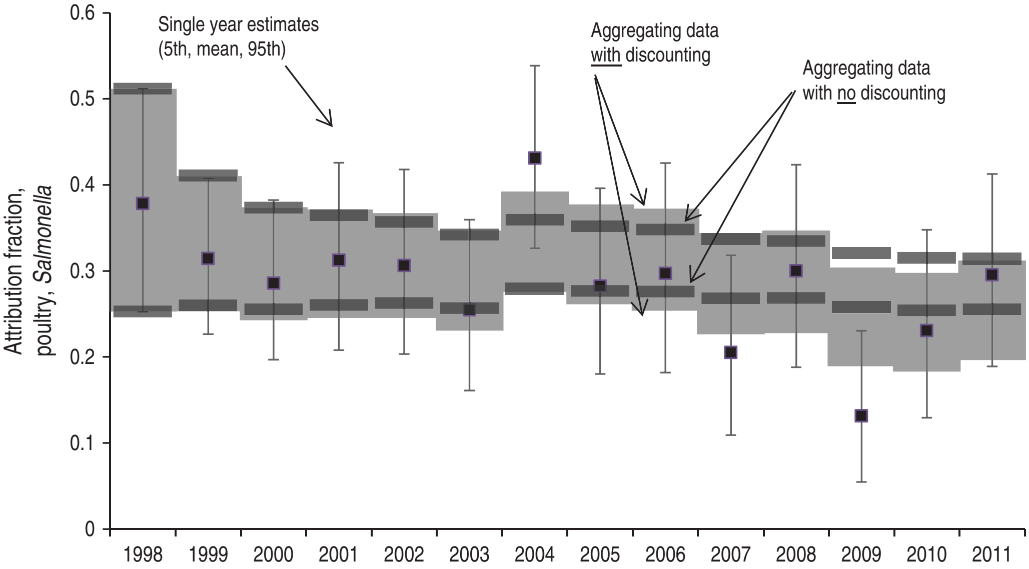

Fig. 3. Illustrative example of the evolution of the estimated attribution fraction from *Salmonella*–poultry outbreaks. Single year estimates are contrasted with aggregated estimates with or without discounting. In all cases, lower and upper credible limits are the 5th and 95th percentiles.

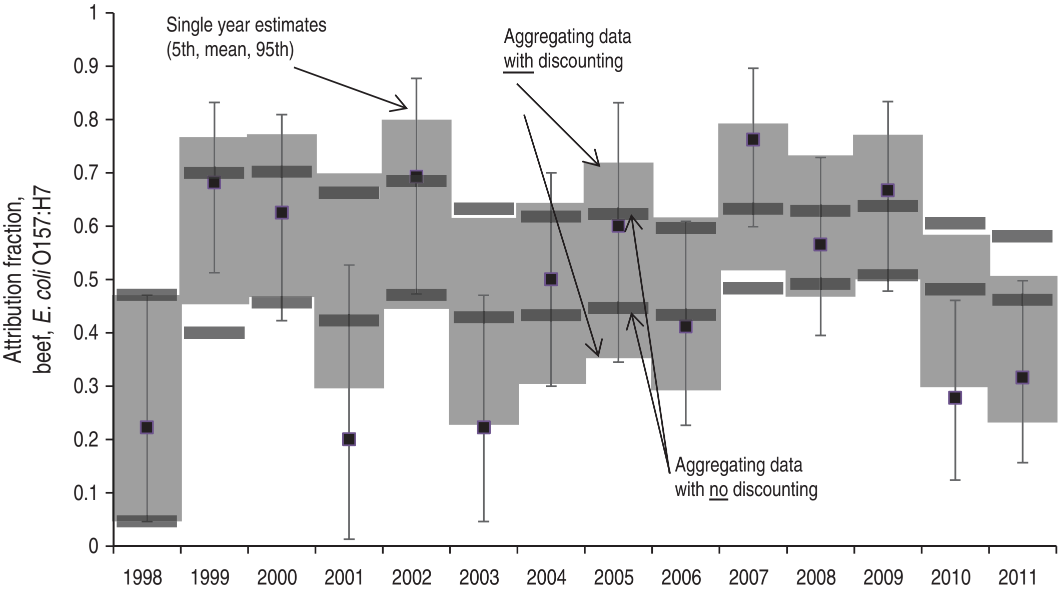

Fig. 4. Illustrative example of the evolution of the estimated attribution fraction from *E. coli* O157:H7–beef outbreaks. Single year estimates are contrasted with aggregated estimates with or without discounting. In all cases, lower and upper credible limits are the 5th and 95th percentiles.

Therefore, discounting historical data reduces the influence of older data relative to new data and generates credible intervals that, at least for these examples, overlap the credible intervals of the new data.

Figure 4 allows similar comparisons for the *E. coli* O157:H7–beef pairing. Individual year attribution fraction estimates fluctuate wildly across time. Aggregating data without discounting implies progressively increased credibility (Fig. 4) while aggregating data with discounting suggests relatively stable credibility across time. Furthermore, the discounted credible intervals overlap estimates for each year better than the undiscounted intervals. In particular, the discounted credible interval overlaps the interval

https://doi.org/10.1017/S0950268815001193 Published online by Cambridge University Press

estimated for the 2010 data while the undiscounted interval does not overlap.

DISCUSSION

This approach to time valuation of historical outbreak data provides epidemiologists and risk analysts with an objective estimate of food source attribution. Governments rely on attribution fractions for prioritizing control strategies to reduce foodborne illnesses. For example, the European Food Safety Authority risk rankings across foods and pathogens [25], the US decision to ban the use of fluoroquinolone antibiotics in poultry [26] and the US decisions to reduce *Salmonella* contamination on poultry carcasses [6] were all based on attribution fraction estimates. Improved methods for estimating food source attribution from available outbreak evidence will increase the accuracy and reliability of these inputs to government decision-making.

For the majority of commodity classes considered here, we found no evidence to suggest that outbreak data should be discounted across time. For example, for 14/17 commodity classes associated with *Salmonella* outbreaks, we determined there was no justification for discounting historical evidence when estimating the current attribution fraction. This was also true for 8/10 commodity classes associated with *E. coli* O157:H7 outbreaks. In their estimation of attribution fractions from foodborne disease outbreaks, Painter *et al.* [4] did not discount outbreak illness data. Our results support this non-discounting approach generally. Nevertheless, our analysis does suggest that discounting of historical data is appropriate for some of the commodity classes for the pathogens we evaluated.

Estimating the attribution fraction for commodity classes for which discounting of historical information is warranted (e.g. *Salmonella*–poultry, *Salmonella*–eggs, *Salmonella*–finfish, *E. coli* O157:H7–beef, *E. coli* O157:H7–dairy) can be accomplished objectively using our methods. In these cases, our approach estimates current attribution fractions that are more certain than an approach that only uses the current year's information; and it is likely more accurate than an approach that aggregates all historical information without any accounting of the time value of those data. Furthermore, in comparison with a non-discounted aggregation of data, our methods generate wider credible intervals that imply less certainty about the true attribution fraction for a food source. When these wider intervals are used to estimate the annual number of a pathogen's illnesses associated with a particular food, the result is a less certain estimate of the illnesses available annually for prevention by a policy. Consequently, a proposed policy for reducing annual illnesses associated with a food might require a more substantive change to achieve some threshold benefit relative to an estimate based on non-discounted attribution estimates. With respect to monitoring illnesses across time, the wider limits of the discount-based attribution fractions imply that changes of a larger magnitude are necessary to conclude there has been a true change in illnesses due to a particular food.

When warranted, we determine the appropriate discount rate for a commodity class's data by deriving beta distributions for a hypothetical stationary process and a non-stationary process based on the observed data. Each distribution reflects how the attribution fraction randomly varies from year to year while also accounting for the uncertainty inherent in the sampling evidence. Each estimated beta distribution implies an underlying sample size that informs the annual attribution fraction. Because it reflects a more variable process, the non-stationary beta distribution implies a smaller sample size than the stationary beta distribution. The ratio of these two sample sizes implies how much to discount annual data from the non-stationary process.

Our method uses the discount rate to reduce the value of historical outbreak attribution data. These data reflect the counts of outbreaks attributed to particular foods. Therefore, counts from past years are discounted as they are aggregated to the present. The older the count data, the more it is discounted. It should be noted that these methods are only applicable to outbreak counts; previous estimates of food source attribution in the United States were based on aggregating the illnesses associated with outbreaks (e.g. [4]) while estimates for other countries have been based on outbreak counts like the approach used here (e.g. [5]). Because the number of illnesses associated with an outbreak varies with the severity of the outbreak (i.e. cases per outbreak), the convolution of outbreak frequency and severity would need to be considered when using illnesses for attribution estimation. The lack of a theoretic argument for what constitutes a stationary pattern for total outbreak-associated illnesses per year suggests that our approach would not be amenable to use of illness counts for determining the appropriate discount rate to apply to those data.

https://doi.org/10.1017/S0950268815001193 Published online by Cambridge University Press

Aggregating data across time is consistent with Bayesian thinking. In Bayesian analysis, the prior distribution reflects information known before collection of new data. The Bayesian updating process estimates a posterior distribution about a parameter by considering both the new data and the prior information. If the prior distribution is very narrow, then the posterior distribution for the parameter is very similar to the prior unless an overwhelming amount of new data is available. A very broad prior distribution can generate a very different posterior distribution that is more reflective of the new data.

Our methods make prior distributions more broad by reducing the value of historical count data while still maintaining their central tendencies. In this manner, the prior distribution has less influence on the resulting attribution fraction estimate than it would if no discounting occurred. Nevertheless, the discounted prior distribution still generates more informed distributions compared to a completely uninformed prior distribution. Practically, an uninformed prior distribution would be the result of a discount rate that approaches 1.

This method is a more objective relative to alternative approaches that either make assumptions about how to discount historical data or choose to ignore it. Previous research has explained that the time value of epidemiological information depends on the dynamics of infectious processes [12, 13]. Given some historical sampling evidence, it is possible to model how the infectious disease might evolve from its time of collection to the current time of inference. Given the various pathways infection might take, the resulting current time estimate is necessarily more uncertain than the inference made at the time the data were collected.

The proposed method is statistically rigorous in that it relies on goodness of fit methods to examine the appropriateness of a stationary process assumption. If the Poisson test fails to reject the assumed stationary Poisson process for a product–pathogen pair, then it is recommended that historical data not be discounted. If the variability observed among outbreaks is consistent with a Poisson process, then there is no reason to think historical data are less relevant than data that are more current. Nevertheless, if the test rejects the stationary Poisson process, then the attribution fraction is more dynamic, and its changes beyond a Poisson process can result in a degradation of value of historical evidence. The use of the beta distributions in our approach serves to quantify the implied reduction in sample size that attends the increased variability in hyper-Poisson processes.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Scallan E, *et al.*** Foodborne illness acquired in the United States – major pathogens. *Emerging Infectious Diseases* 2011; **17**: 7–15.
2. **Rockhill B, Newman B, Weinberg C.** Use and misuse of population attributable fractions. *American Journal of Public Health* 1998; **88**: 15–19.
3. **Painter JA, *et al.*** Recipes for foodborne outbreaks: a scheme for categorizing and grouping implicated foods. *Foodborne Pathogens and Disease* 2009; **6**: 1259–1264.
4. **Painter JA, *et al.*** Attribution of foodborne illnesses, hospitalizations, and deaths to food commodities, United States, 1998–2008. *Emerging Infectious Diseases* 2013; **19**: 407–415.
5. **Ravel A, *et al.*** Exploring historical Canadian foodborne outbreak data sets for human illness attribution. *Journal of Food Protection* 2009; **72**: 963–1976.
6. **Ebel ED, *et al.*** Simplified framework for predicting changes in public health from performance standards applied in slaughter establishments. *Food Control* 2012; **28**: 250–257.
7. **Guo C, *et al.*** Application of Bayesian techniques to model the burden of human salmonellosis attributable to U.S. food commodities at the point of processing: adaptation of a Danish model. *Foodborne Pathogens and Disease* 2011; **8**: 509–516.
8. **Hald T, *et al.*** A Bayesian approach to quantify the contribution of animal-food sources to human salmonellosis. *Risk Analysis* 2004; **24**: 255–269.
9. **Alexander CH.** Still rolling: Leslie Kish's 'rolling samples' and the American Community Survey. *Survey Methodology* 2002; **28**: 35–41.
10. **Johnson DS, Williams MS, Czaplewski RL.** Comparison of estimators for rolling samples using forest inventory and analysis data. *Forest Science* 2003; **49**: 50–63.
11. **Hyndman R, *et al.*** *Forecasting with Exponential Smoothing: the State Space Approach*. Berlin: Springer Science & Business Media, 2008.
12. **Ebel ED, Williams MS, Tomlinson SM.** Estimating herd prevalence of bovine brucellosis in 46 U.S.A. states using slaughter surveillance. *Preventive Veterinary Medicine* 2008; **85**: 295–316.
13. **Schlosser W, Ebel E.** Use of a Markov-chain Monte Carlo model to evaluate the time value of historical testing information in animal populations. *Preventive Veterinary Medicine* 2001; **48**: 167–175.
14. **CDC.** Foodborne disease outbreak surveillance (http://www.cdc.gov/outbreaknet/surveillance_data.html). Atlanta, 2013.

https://doi.org/10.1017/S0950268815001193 Published online by Cambridge University Press

15. **Williams MS, *et al.*** Temporal patterns in the occurrence of *Salmonella* in raw meat and poultry products and their relationship to human illnesses in the United States. *Food Control* 2014; **35**: 267–273.
16. **Williams MS, *et al.*** Determining relationships between the seasonal occurrence of *Escherichia coli* O157:H7 in live cattle, ground beef, and humans. *Foodborne Pathogens and Disease* 2010; **7**: 1247–1254.
17. **Brockwell PJ, Davis RA.** *Time Series: Theory and Methods*: New York: Springer Science & Business Media, 2009.
18. **Taylor HM, Karlin S.** *An Introduction to Stochastic Modeling*. Boston: Academic Press, 1998.
19. **Snedecor GW, Cochran WG.** *Statistical Methods*. Iowa: Iowa State University Press, 1989.
20. **Vose D.** *Risk Analysis: A Quantitative Guide*, 3rd edn. West Sussex: John Wiley & Sons, 2008.
21. **Gelman A, *et al.*** *Bayesian Data Analysis* (New York: Chapman & Hall/CRC Texts in Statistical Science), 2003.
22. **R Development Core Team.** *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2015.
23. **Frees EW.** *Regression Modeling with Actuarial and Financial Applications*. New York: Cambridge University Press, 2009.
24. **Hyndman RJ, Athanasopoulos G.** *Forecasting: Principles and Practice*. Melbourne: OTexts, 2014.
25. **EFSA.** Scientific opinion on the development of a risk ranking toolbox for the EFSA BIOHAZ panel. *EFSA Journal* 2015; **13**: 3939.
26. **Bartholomew MJ, *et al.*** A linear model for managing the risk of antimicrobial resistance originating in food animals. *Risk Analysis* 2005; **25**: 99–108.

https://doi.org/10.1017/S0950268815001193 Published online by Cambridge University Press