Inferring source attribution from a multi-year multi-source dataset of Salmonella in Minnesota

Journal:	*Zoonoses and Public Health*
Manuscript ID	ZPH-Oct-16-318.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Ahlstrom, Christina; Epi-interactive, Muellner, P; Epi-interactive Spencer, Simon; University of Warwick Hong, Samuel; University of Minnesota, Veterinary Population Medicine Saupe, Amy; Minnesota Department of Health Rovira, Albert; University of Minnesota, Veterinary Diagnostic Laboratory Hedberg, Craig; University of Minnesota, Division of Environmental Health Sciences Perez, Andres; University of Minnesota, Veterinary Population Medicine Muellner, Ulrich; Epi-interactive Alvarez, Julio; University of Minnesota, Veterinary Population Medicine
Subject Area:	Salmonella spp, Molecular epidemiology, Source attribution, Data visualization

Inferring source attribution from a multi-year multi-source dataset of *Salmonella* in Minnesota

Christina Ahlstrom[1], Petra Muellner[1], Simon EF Spencer[2], Samuel Hong[3], Amy Saupe[4], Albert Rovira[5], Craig Hedberg[6], Andres Perez[3], Ulrich Muellner[1], Julio Alvarez[3]

[1] Epi-interactive, Wellington, New Zealand

[2] University of Warwick, Coventry, United Kingdom

[3] Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota, St Paul, USA

[4] Minnesota Department of Health, Saint Paul, USA

[5] Veterinary Diagnostic Laboratory, College of Veterinary Medicine, University of Minnesota, St Paul, USA

[6] Division of Environmental Health Sciences, School of Public Health, University of Minnesota, Minneapolis, USA

Corresponding author: Petra Muellner, Epi-interactive, Wellington, New Zealand.

petra@epi-interactive.com

Summary

Salmonella enterica is a global health concern because of its widespread association with foodborne illness. Bayesian models have been developed to attribute the burden of human salmonellosis to specific sources with the ultimate objective of prioritizing intervention strategies. Important considerations of source attribution models include the evaluation of the quality of input data, assessment of whether attribution results logically reflect the data trends, and identification of patterns within the data that might explain the detailed contribution of different sources to the disease burden. Here, more than 12,000 non-typhoidal *Salmonella* isolates from human, bovine, porcine, chicken, and turkey sources that originated in Minnesota were analyzed. A modified Bayesian source attribution model (available in a dedicated R package), accounting for non-sampled sources of infection, attributed 4,672 human cases to sources assessed here. Most (60%) cases were attributed to chicken, though there was a spike in cases attributed to a non-sampled source in the second half of the study period. Molecular epidemiological analysis methods were used to supplement risk modelling and a visual attribution application was developed to facilitate data exploration and comprehension of the large multi-year dataset assessed here. A large amount of within-source diversity and low similarity between sources was observed and visual exploration of data provided clues into variations driving the attribution modelling results. Results from this pillared approach provided first attribution estimates for *Salmonella* in Minnesota and offer an understanding of current data gaps as well as key pathogen population features, such as serotype frequency, similarity and diversity across the sources. Results here will be used to inform policy and management strategies ultimately intended to prevent and control *Salmonella* infection in the state.

Keywords: Source attribution, molecular epidemiology, data visualization, *Salmonella*, salmonellosis

Bullet points:

- We demonstrate the attribution of salmonellosis to different sources can be enhanced through a pillared approach that incorporates data visualization and molecular epidemiology on top of source attribution modelling. In this study, such supplementary analyses supported findings from the attribution model, demonstrating for example high genotype diversity and low similarity between sources.
- We developed a modified Bayesian source attribution model that attributed the majority of salmonellosis cases in Minnesota to chickens. Accounting for a known data gap, we were able to demonstrate the potential impact of a non-sampled source on the number of attributed cases.
- A visual attribution application enabled users to dynamically explore the occurrence of Salmonella serotypes in different sources over time. The ability to interact and filter the data assisted in the detection of data irregularities, facilitating accurate attribution estimates and the interpretation of results.

Introduction

Non-typhoidal *Salmonella enterica* is considered one of the leading causes of foodborne illness in the United States (Scallan et al., 2011) and other countries (Kirk et al., 2015), despite large-scale control efforts initiated by industry, government and consumers (Crim et al., 2015). In the United States, *S. enterica* has the highest reported incidence in humans among bacterial pathogens, with 15.3 cases per 100,000 persons in 2014 (CDC, 2016), the majority of which are sporadic with unknown source of origin (Batz et al., 2005; Ebel et al., 2016). In Minnesota, the incidence of culture-confirmed *Salmonella* cases in 2015 was 17.9 per 100,000 persons (Minnesota Department of Health, 2015). The most common form of clinical illness associated with *S. enterica* infection is gastroenteritis, which is mostly non-severe and self-limiting; however, *S. enterica* can cause severe disease or complications in some patients (e.g. sepsis). *S. enterica* has multiple reservoirs, including livestock and domestic pets (Kingsley and Bäumler, 2000), impairing efforts to identify infection sources and transmission routes. Whereas some *S. enterica* serotypes have adapted to individual host species, others are frequently isolated from a broad range of species and environments (Uzzau et al., 2000).

Source attribution models aim to estimate the proportion of human cases attributable to specific sources. Estimates obtained can subsequently be used to guide risk managers and decision-makers in the design and implementation of effective intervention strategies (Pires and Hald, 2010; Sears et al., 2011). In the United States, different source attribution methods using a variety of data sources have been employed to estimate the burden of human salmonellosis from several food commodities. For example, the relative proportions of domestically acquired, sporadic human *Salmonella* infections between 1998 and 2003 among multiple food sources was investigated using a molecular subtyping approach (Guo et al.,

2011), whereas other source attribution assessments relied either on expert elicitation (Hoffmann et al., 2007) or outbreak-associated illnesses (Batz et al., 2012; Gould et al., 2013; Painter et al., 2013). Data collected during outbreak investigations are typically more accessible than sporadic illness data; however, foodborne outbreaks account only for a small proportion of total *Salmonella* infections in humans and may not be representative of the majority of human cases of salmonellosis (Ebel et al., 2016; Painter et al., 2013).

Although early risk assessment models estimated the contribution of a single source to the burden of human illness, Bayesian source attribution models for salmonellosis (Hald et al., 2004) were subsequently developed to consider multiple sources simultaneously while allowing for uncertainty around input parameters. Such an approach has been applied internationally and modified over time to accommodate different contexts and pathogens (Mullner et al. 2009a; Little et al. 2010; Mughini-Gras et al. 2014; David et al. 2013; Guo et al. 2011; Glass et al. 2015; Mullner et al. 2009b). Briefly, the model framework compares the distribution of pathogen serotypes in the source populations to the distribution of serotypes observed in humans via a Poisson regression model fitted within a Bayesian framework. While models developed thus far have provided valuable insights into the contribution of different sources to the human disease burden, they generally rely on a specific type of data that may not always be available. Previous work has, for example, explored the use of different types of prevalence data (Mullner et al. 2009a). The work presented her proposes a further extension that may allow the model to be successfully used in the presence of data gaps.

Bayesian source attribution models often rely on multi-level data collected in multiple years and sources, requiring significant resources to manually clean and prepare the data. Model outputs are typically one-dimensional and primarily focus on one quantitative figure (i.e. the

attributed proportion of human cases to each source). A comprehensive understanding of the characteristics of the input data and results could enhance the way attribution results are understood by, and communicated to, risk managers and other stakeholders. For example, molecular epidemiological analyses, such as diversity statistics and similarity indexes (Muellner et al., 2011), can be incorporated into the analysis to support model outputs. Additionally, the integration of visual data exploration creates opportunities to dynamically interact with layered data and identify trends to help explain attribution results. Such an approach can help make large, complex datasets digestible through interactive graphics that facilitate incremental exploration of the data (Carroll et al., 2014).

The objective of the study presented here was to explore the feasibility of using a pillared attribution approach, adding molecular epidemiology and visual data analysis, to a customised Hald Model based on the work by Mullner et al. (2009), to generate preliminary estimates of the source-specific burden of salmonellosis in Minnesota over a ten-year period using state-level data. Results here will help to inform policy and management activities intended to prevent and control the disease in the state. Additionally, methods presented here may serve as a framework that could be applied to the attribution of sources of infection for *Salmonella* and other foodborne pathogens in the United States and other regions.

Materials and methods

Data sources

Information from human salmonellosis cases reported to the Minnesota Department of Health (MDH) from 2005 to 2014 were collected. Data included case serotype (determined by the MDH Public Health Laboratory [PHL]), date of specimen collection, international travel in

the seven days prior to illness onset (self-reported by cases during routine exposure interviews at MDH), and whether the case was part of an identified outbreak. Consistent with the approach developed by Hald et al (2004), cases with a history of international travel were excluded, as were cases attributed to an outbreak.

Information on *Salmonella* isolates of food animal (cattle, swine, poultry) origin isolated by the Minnesota Veterinary Diagnostic Laboratory (MVDL) from diagnostic submissions between 2006 and 2015 were also collected. A detailed description of the MVDL database and cattle and swine isolates is available elsewhere (Hong et al., 2016). Finally, data on the relative frequency of *Salmonella* serotypes stratified per meat product (chicken, ground turkey, ground beef, pork chop) from various retail locations in Minnesota that were collected as part of the FoodNet/National Antimicrobial Resistance Monitoring System (NARMS) Retail Food Study between 2002 and 2013 were provided by MDH. These food commodity source data represented 4% of isolates derived from non-human sources and were combined with the animal data (e.g. isolates from pork chops were added to animal porcine isolates). The majority (96%) of food-derived isolates came from chicken breasts and ground turkey. Non-human source categories were limited to bovine, porcine, chicken, and turkey, because typed isolates from other sources were considered too scarce ($n < 50$) to be included in the risk model. Further, no data on eggs were available.

The consistency of serotype naming was checked within and between data sources. All serotypes were defined using naming conventions proposed by the Pasteur Institute (Grimont and Weill, 2007). After 2012, the Centers for Disease Control and Prevention (CDC) recommended naming all *S.* Typhimurium var. 5 – (formerly var. Copenhagen) as *S.* Typhimurium (*S.* I 4,[5],12:i:1,2; CDC, 2014). Thus, the MDH did not electronically record variants of *S.* Typhimurium after 2011; however, variants continued to be identified and

recorded in paper records. Therefore, *S.* Typhimurium var. 5– (formerly var. Copenhagen) cases were retrospectively recorded electronically in the current dataset.

Diversity and similarity statistics

The dataset containing all *Salmonella* isolates from bovine, porcine, chicken, turkey, and sporadic and domestic human cases was used to calculate diversity and similarity statistics. Serotype richness and the diversity of *Salmonella* serotypes from different sources was estimated using rarefaction and Simpson's index of diversity. The function RAREFY in the package VEGAN (version 2.2-3) was implemented in R (version 3.2.3; R Core Team 2016) and Simpson's index of diversity was calculated in Past (version 3.10; Hammer et al. 2009), with 9,999 bootstrap replicates. The similarity of serotypes between different sources was investigated by calculating the proportional similarity index (PSI), which measures the area of overlap between two frequency distributions of each serotype between sources. Bootstrap confidence intervals were estimated as described by Mullner et al. (2009).

Source attribution modelling

Data preparation

Salmonella isolates originating from non-human sources that did not match a serotype found in Minnesota human cases during the study period were excluded. Similarly, *Salmonella* isolates originating from humans that did not match a serotype found in non-human sources were excluded. Serotypes that included fewer than five human cases were combined into a serotype designated as "other", similar to the approach taken in previous studies (Hald et al., 2004; Mullner et al., 2009a; David et al., 2012). Sampling effort of non-human sources varied across the study period; thus, isolates from non-human sources were not segregated into years. Human isolates were summarized as total cases per serotype per year between 2005-2014.

Attribution model

The source attribution model described by Hald et al. (2004) and modified by Muellner et al. (2009) was further modified here to simplify and improve applicability and interpretability of results. The approach proposed here may be used in a wider variety of circumstances compared to earlier methods, in particular, without knowledge of the absolute prevalence in each source.

The Poisson regression structure was retained (adapted slightly to include time) $Y_{it} \sim Pois(\sum_{j=1}^{S} \lambda_{ij})$, where Y_{it} is the number of serotype i human cases in year $t = 1,...,T$. The equation defining the expected number of cases of serotype i (for $i=1,\ldots,I$) from source j (for $j=1,\ldots,S$) was modified as

$$\lambda_{ij} = r_{ij} q_i a_j.$$

Model parameters are described in Table 1. Note that this equation does not require the absolute source prevalences p_{ij} or π_j (Mullner et al., 2009a) and instead uses just the relative prevalences r_{ij}, plus the strain i- and the source j-specific factors q_i and a_j, respectively. This model can therefore be fitted when source prevalence data are not available. Let X_{ij} denote the number of serotype i isolates observed in source j. The following prior distributions were assumed:

$(r_{1j}, r_{2j}, ..., r_{Ij}) \sim$ Dirichlet($\gamma_1 + X_{1j}, \gamma_2 + X_{2j}, ..., \gamma_I + X_{Ij}$)

$q_i \sim$ Gamma(θ, θ), with $\theta \sim$ Gamma($\alpha^{(\theta)}$, $\beta^{(\theta)}$)

$a_j \sim$ Exp(α_j)

The hyperparameter θ represents the precision and also the shape parameter of the

distribution of the random effects describing the strain-specific differences q_i. In previous versions of the source attribution model, the prior for the source-specific parameter a_j was difficult to specify. Here, because the prior mean of q_i is one and the relative prevalences sum up to one, the prior mean of a_j (which equals α_j^{-1}) is the expected number of cases from source *j*. This interpretation allows informed priors for the source-specific factors to be easily defined.

In the current implementation, we chose $\alpha_j^{-1} = \sum_{i=1}^{I}\sum_{t=1}^{T} Y_{it}/TS$ for all $j = 1,\ldots, S$, which implies the prior belief that an equal number of cases comes from each source and that cases appear at the average rate observed in the data. That procedure was followed to overcome concerns regarding the choice of uniform priors with fixed boundaries as described by Hald et al. (2004), as the inferences have been shown in some cases to be sensitive to the choice of boundaries (Glass et al., 2015). For the remaining hyperparameters we chose $\gamma_i = 1, \alpha^{(\theta)} = \beta^{(\theta)} = 1$ to reflect weak prior knowledge.

The distribution of random effects was changed from a log-normal distribution with a mean of 1 to a *Gamma* distribution with a mean of 1, which allows a Gibbs step to be used to update q_i during the Markov chain Monte Carlo (MCMC) (Denison, 2002). The full conditional distribution was given by

$q_i|(\theta, r, a, Y) \sim$ Gamma $\left(\theta + \sum_{t=1}^{T} Y_{it}\,, \theta + T\sum_{j=1}^{S} r_{ij}\, a_j\right)$. The hyperparameter θ can be fixed, but we chose to specify a prior for θ so that the variance of the random effects was estimated from data.

The relative prevalence parameters r_{ij} were updated using Metropolis-Hastings updates (Chib and Greenberg, 1995) with proposals from the prior distribution in both large and small blocks. The large blocks consisted of all types $i = 1, \ldots, I$ for a given source and the small

blocks consisted of just two types for a given source. The source dependent factors a_j were updated jointly using Metropolis-Hastings random walk proposals on the log scale. To avoid poor mixing when the values of a_j were small, single component Metropolis-Hastings proposals from the prior distribution were applied. The model described above is considered a single attribution model because it assumes that the source attribution is the same for each year in the dataset.

In the temporal source attribution model, the source dependent factors depend on time and therefore produce a different source attribution for each year in the study. Given that the source data are sparser than the human data, and they were not collected throughout the whole study period, relative prevalences r_{ij} that do not depend on time were used.

The temporal attribution model becomes $Y_{it} \sim Pois(\sum_{j=1}^{S} \lambda_{ijt})$, for year $t = 1, \ldots, T$ and type $i = 1, \ldots, I$. The mean number of cases is decomposed as

$$\lambda_{ijt} = r_{ij} q_i a_{jt}.$$

Similarly, $a_{jt} \sim Exp(\alpha_{jt})$ is assumed, and $\alpha_{jt}^{-1} = \sum_{i=1}^{I} Y_{it}/S$, dividing the prior weight equally between sources as before.

Convergence was confirmed for all parameters by visual inspection of the trace plots and comparison of the posteriors from chains with randomly chosen starting values. Once convergence was established, any evidence of lack-of-fit in the source attribution model was likely due to human cases that were difficult to attribute to any of the sources in the model. To help investigate that feature, an additional source, referred to as a non-sampled source, was included in the model (for which no source data were observed) to identify the quantity and profile of unattributable cases. To further assess the model performance, data from a previous *Campylobacter* source attribution analysis (Mullner et al., 2009a) was analyzed and

outputs were compared.

The model was run in R (version 3.2.3; R Core Team, 2016) and to facilitate dissemination, an associated R package is under development. This package containing the updated model includes adaptive proposals so that unreasonable mixing is detected and the proposals are adjusted to combat it. Furthermore, the package includes functions to aid interpretation of the output by non-specialists.

Visual attribution application development

An application was created to allow users to dynamically investigate the occurrence of *Salmonella* serotypes in different sources and identify patterns in the data. This application, named *Source Explorer*, was developed in an RStudio Shiny (http://shiny.rstudio.com) framework, which enabled the tool to be used locally, as a stand-alone version, or through a web-based interface. The dataset of *Salmonella* serotypes per source was used as input data, including the years in which the isolates were collected. Five main visualization outputs were displayed, including 1) the proportion of the top 10 serotypes of each selected source over time, 2) user-selected serotypes for the selected source over time, 3) a bar chart comparing the top 10 serotypes of the selected source with all other sources, 4) the top 10 human serotypes per year overlayed with non-human sources, and 5) rarefaction results with options to select a subset of sources. The proportion of isolates per source and year was displayed to account for different sampling efforts during the study period.

Results

Salmonella diversity

Table 2 shows the total number of *Salmonella* isolates derived from human, porcine, bovine, turkey and chicken sources in the full dataset and the number of isolates with serotypes found in both human and non-human sources. A total of 991 and 749 international travel and outbreak associated cases were excluded, respectively. An additional 330 cases were further excluded, as they were either *S.* Typi, *S.* Paratyphi, or of an unknown serotype. The number of sporadic and domestically acquired human cases per year in the dataset ranged from a minimum of 405 cases in 2009 to a maximum of 546 cases in 2013. In total, 240 *Salmonella* serotypes were found, 156 (65%) of which were isolated from a single source type. Twelve (5%) serotypes were found in all five sources. Serotype *S.* Typhimurium var. 5- (formerly known as *S*. Copenhagen) had the largest total number of isolates, with 1,617 isolates from human, bovine, and porcine sources.

Rarefaction analysis indicated a larger serotype richness in human *Salmonella* isolates than in those from non-human sources (Figure 1). In addition, a lower level of diversity was found among isolates from chickens than those from other sources. Simpson's index of diversity of isolates from each source is presented in Table 3. *Salmonella* isolates from human, turkey, and porcine sources had the highest serotype diversity, whereas isolates from chicken and bovine sources were less diverse. Overall, the PSI analysis indicated a relatively low level of similarity between the different sources (Table 3). Human isolates were most similar to porcine isolates, however 95% bootstrap confidence intervals overlapped between all sources.

Source attribution

Seventy-five different serotypes were found both in humans and in at least one of the non-human sources, 50 of which included at least five human cases, representing 96% of isolates included in the full dataset. The remaining 25 serotypes included 785 isolates and were designated as "other".

The number of human salmonellosis cases attributed to each source in the single attribution model is shown without (Figure 2A) and with (Figure 2B) inclusion of a non-sampled source. Out of 4,672 human cases, the largest share was attributed to chickens, accounting for 2,790 (60%) and 2,049 (45%) without and with a non-sampled source, respectively.

Results from the temporal attribution model are shown in Figure 3. A marked change in the attribution estimates after 2009 was observed, with an increase in the number of cases attributed to the non-sampled source. The human cases most frequently assigned to the non-sampled source by the model included serotypes *S.* 4,5,12:i:- (8.8%), *S.* Enteritidis (7.6%), *S.* Berta (6.1%), *S.* Infantis (5.9%) and *S.* Heidelberg (4.0%).

The temporal model generally showed sufficient fit, aligned with previous validations of the model where expected and observed cases were compared to test the validity of the model outputs (e.g. Hald et al. 2004). A more stringent model assessment was also performed in which the prior and the posterior distributions were compared for each of the relative prevalence parameters. The prior for these parameters is based on the source data whilst the posterior incorporates both the source and human data, and so if there is disagreement between these two distributions then some of the human data cannot be aligned with any of the sources. A selection of these plots is presented in Supporting Figure 1. Outputs from the Mullner et al. (2009a) data using the current model (data not shown) were consistent with the results presented in the original paper, further supporting the performance of the model

presented here.

Visual attribution

Multiple filters and visualization options facilitated extensive data exploration in *Source Explorer*. Outputs were visualized in charts and data tables and were instantaneously updated as search variables were modified. Individual serotypes or the top 10 serotypes for each source could be selected and visualized over time. For example, the proportion of human isolates belonging to each of the top 10 serotypes from humans was overlayed with the proportion of isolates from each source (Figure 4). This example demonstrates the high proportion of porcine *S.* Typhimurium var. 5- isolates throughout the study period, compared to other sources.

Source Explorer enabled us to further investigate the source attribution results and potential reasons for the spike in cases attributed to a non-sampled source after 2009. Out of the top 10 most common human serotypes in our dataset, *S.* Enteritidis and *S.* 4,5,12:i:- increased relative to the other serotypes isolated after 2009 (Figure 5). Further, these two serotypes were not frequently isolated from non-human sources throughout the study period, representing 0.5 and 2.5 percent of serotypes isolated from non-human sources, respectively.

Discussion

Source attribution models continue to be developed and modified to accommodate different environments, pathogens, and data sources (David et al., 2013; Glass et al., 2015; Guo et al., 2011; Hald et al., 2004; Little et al., 2010; Mughini-Gras et al., 2014; Mullner et al., 2009a). A detailed exploration of the data that are inputted into such models can help drive their

evolution to best incorporate and model observed data. Therefore, a multi-pillared approach was used here to explore the burden of sporadic, domestically-acquired human *Salmonella* infections in Minnesota from different sources over a ten-year period. Serotype diversity and similarity in human, bovine, porcine, chicken, and turkey sources was assessed and a visual attribution tool was developed to facilitate data exploration and validation. A modified Bayesian source attribution model was subsequently used to estimate the relative contribution of *Salmonella* isolates from alternative sources.

Source Explorer enabled users to visually assess and interact with the large datasets commonly included in source attribution analyses. Attribution results here were not only clarified through data exploration, but this visualization tool also helped to identify odd trends in the data, such as drastic changes in the proportion of common serotypes over time. For example, a lack of human *S.* Typhimurim var. 5- isolates after 2011, whereas that serotype previously accounted for an average of 30% of all human *Salmonella* serotypes, suggested a change in 2012. Such a finding led to further investigation into serotype naming convention changes by the CDC and the state public health laboratory (CDC, 2014) and ultimately led to the retrospective electronic coding of human *S.* Typhimurim var. 5- cases after 2011. Quality of input data is critical to provide accurate attribution results that best explain the burden of human disease. A thorough description of data and whether trends accurately reflect attribution estimates is an important step that should be included in any attribution analysis. That step further ensures a comprehensive and transparent assessment that can be more easily communicated to various stakeholder groups to build awareness and engagement and ultimately support the development of control strategies (Carroll et al., 2014).

The highest number of sporadic and domestic cases were attributed to chickens in the single attribution analyses, both with and without the inclusion of a non-sampled source. In the temporal analysis, the non-sampled source had the most attributed cases, followed by chickens. That finding supports previous attribution studies in the United States, where poultry was the leading source of human salmonellosis (Batz et al., 2012; Gould et al., 2013; Guo et al., 2011; Hoffmann et al., 2007). The temporal attribution results showed a surprising amount of smoothness over time in the early part of the study, particularly considering that no smoothing factor was incorporated into the model. That time period is also when most of the data originated from some sources and so it may well be the case that changes in attribution seen after 2010 may be the result of the source data becoming out of date rather than such substantial changes in the number of cases coming from each source.

The wide credible intervals in Figure 2b show that there is a large amount of uncertainty in the estimates regarding the dominant source in the single attribution model, with poultry and non-sampled being the two largest estimates. In the temporal attribution model, there is a period where poultry appears to make the largest contribution and a period where non-sampled appears to make the largest contribution. Given that the single attribution model cannot incorporate temporal changes, except as Poisson variation in the number of cases, in this study, we conclude that the single attribution model does not fit as well as the temporal attribution model, which results in the differences observed. This was further supported by the observed vs. expected plot (Supporting Figure 2), which shows more variation in the single attribution model between years than predicted by the Poisson distribution. Better agreement can be observed between observed and expected cases in the temporal model. Remaining deviations might indicate remaining lack of fit, but could also be explained by random variation in source sampling.

As only four non-human sources had sufficient numbers of isolates to be included here, there was likely an overestimation of the burden of illness from these sources. The notable exclusion of eggs as a source may have skewed the attribution estimates, as eggs are a known source of *Salmonella*, particularly *S.* Enteritidis (Hedberg et al., 1993; Wright et al., 2016). Indeed, *S.* Enteritidis did not show a good fit between observed and posterior cases (data not shown) in the current study. A European Union Food Safety Authority report ((BIOHAZ) EFSA Panel on Biological Hazards 2013) recently highlighted the need to include isolates from all potential major non-human sources and that the use of surrogate data or the exclusion of relevant sources may seriously bias attribution results. The inclusion of a non-sampled source in this study was a novel approach in source attribution analyses to deal with the common problem of missing source data; however, it was previously performed as an exploratory technique in other contexts and further work is necessary to confirm the validity of such an approach (Pella and Masuda, 2001). The consistent results obtained using data from a previous source attribution analysis (Mullner et al., 2009a) encourages the use of this model in additional settings. The availability of the model in a dedicated R package will facilitate this.

The supplementary (molecular) epidemiological analysis supported findings from the attribution model. Diversity and similarity statistics highlighted a high degree of diversity and low similarity of *Salmonella* serotypes isolated from all sources. Rarefaction curves for each of the sources did not appear to reach a plateau, indicating that the serotype richness was not fully captured by the current dataset and increased sampling effort in all sources could improve the model fit. Serotype richness was greatest in humans, even when accounting for sampling effort, and the PSI results indicated that overall similarity between sources was relatively low. Low similarity between non-human sources can be advantageous in an attribution analysis, as it supports the attribution of specific serotypes to those sources in

which it is uniquely found (Barco et al., 2013). In a previous study evaluating *Campylobacter* spp. in New Zealand (Mullner et al., 2009b), a higher similarity among sources and humans was found, with similarity estimates ranging from 0.18 to 0.58.

The most appropriate subtyping method for source attribution is one that provides an appropriate level of discrimination to define subtypes associated with specific sources (EFSA Panel on Biological Hazards, 2013). Serotyping is a common method for differentiation of *Salmonella* types, though molecular typing methods (e.g. multiple-locus variable-number tandem repeat analysis, pulsed-field gel electrophoresis) have also been used (Barco et al., 2013). In some attribution studies, authors adjusted the serotype nomenclature, minimizing the total number by combining variants of the same serotype. For example, Guo et al. (2011) combined variants into the non-variant serotypes. This was not performed here because serotype variants were assumed to not change during the course of transmission and sampling. Further supporting this decision, *S.* Typhimurium and Typhimurium var. 5- were the two most commonly isolated *Salmonella* serotypes, yet were differentially distributed among sources in this study, with var 5- known to show some host association to swine reservoirs; hence, that feature provided critical anchor points for the attribution model, which relies on host association of subtypes. In consequence, the differentiation of the large number of *S.* Typhimurium isolates into two distinct serotypes in this study improved the model fit.

Minnesota data were exclusively used in the source attribution analysis, despite much of the food consumed in Minnesota likely originated from outside the state. Nevertheless, there is an increasing trend in the consumption of locally produced foods, as consumers seek direct farm to retail options (Low et al., 2015). Local production, import, and export data have also been included in an attribution analysis to account for the flow of food commodities across borders (De Knegt et al., 2015), revealing that individual countries within the European

Union had different attribution estimates. Further efforts in expanding data collection at a state level in the United States could similarly help to elucidate spatial patterns in attribution. Inclusion of data on geographical origin of isolates from retail foods, such as where the foods were processed and purchased, could also potentially refine the current analysis. An expanded data collection in the United States should ideally include a large number of samples originating directly from food sources or food processing environments, as opposed to animal reservoirs, which is where the majority of non-human isolates in this study were derived. Nevertheless, given the lack of data from food sources, such data from animal sources are commonly used in source attribution analyses (Mughini-Gras and van Pelt, 2014). As with any analysis that relies on reported cases, underreporting of salmonellosis may have introduced some bias into this study.

In summary, results here demonstrated an enhanced approach to source attribution that encourages data exploration through diversity statistics and visual attribution both prior to and after the use of a Bayesian source attribution model. Results here will help to inform preventive and control strategies for *Salmonella* infection in Minnesota.

Acknowledgements

This study was funded by the Global Food Venture MnDrive initiative.

References

Barco, L., Barrucci, F., Olsen, J.E., Ricci, A., 2013. *Salmonella* source attribution based on microbial subtyping. Int. J. Food. Microbiol. 163, 193–203.

Batz, M.B., Doyle, M.P., Morris Jr, G., Painter, J., Singh, R., Tauxe, R. V, Taylor, M.R., Lo Fo Wong, D.M., 2005. Attributing illness to food. Emerg. Infect. Dis. 11, 993–9.

Batz, M.B., Hoffmann, S., Morris Jr, J.G., 2012. 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. 75, 1278–91.

Carroll, L.N., Au, A.P., Detwiler, L.T., Fu, T., Painter, I.S., Abernethy, N.F., 2014. Visualization and analytics tools for infectious disease epidemiology: a systematic review. J. Biomed. Inform. 51, 287–98.

CDC, 2014. National *Salmonella* surveillance annual report, 2012. US Dep. Heal. Hum. Serv. CDC Atlanta, United States.

CDC, 2016. Foodborne Diseases Active Surveillance Network (FoodNet): FoodNet Surveillance Report for 2014. CDC Atlanta, United States.

Chib, S., Greenberg, E., 1995. Understanding the metropolis-hastings algorithm. Am. Stat. 49, 327–35.

Crim, S.M., Griffin, P.M., Tauxe, R., Marder, E.P., Gilliss, D., Cronquist, A.B., Cartter, M., Tobin-D'Angelo, M., Blythe, D., Smith, K., 2015. Preliminary incidence and trends of infection with pathogens transmitted commonly through food—Foodborne Diseases Active Surveillance Network, 10 US sites, 2006–2014. Morb. Mortal. Wkly. Rep. 64, 495–9.

David, J.M., Guillemot, D., Bemrah, N., Thébault, A., Brisabois, A., Chemaly, M., Weill, F.X., Sanders, P., Watier, L., 2013. The Bayesian microbial subtyping attribution model: robustness to prior information and a proposition. Risk Anal. 33, 397–408.

De Knegt, L. V, Pires, S.M., Hald, T., 2015. Attributing foodborne salmonellosis in humans to animal reservoirs in the European Union using a multi-country stochastic model. Epidemiol. Infect. 143, 1175–86.

Denison, D.G.T., 2002. Bayesian methods for nonlinear classification and regression. John Wiley & Sons.

Ebel, E.D., Williams, M.S., Cole, D., Travis, C.C., Klontz, K.C., Golden, N.J., Hoekstra, R.M., 2016. Comparing characteristics of sporadic and outbreak-associated foodborne illnesses, United States, 2004–2011. Emerg. Infect. Dis. 22, 1193.

EFSA Panel on Biological Hazards, 2013. Scientific Opinion on the evaluation of molecular typing methods for major food-borne microbiological hazards and their use for attribution modelling, outbreak investigation and scanning surveillance: Part 1 (evaluation of methods and applications). EFSA J. 11, 3502.

Glass, K., Fearnley, E., Hocking, H., Raupach, J., Veitch, M., Ford, L., Kirk, M.D., 2015.

Bayesian source attribution of salmonellosis in South Australia. Risk Anal. 36, 561-70.

Gould, L.H., Walsh, K.A., Vieira, A.R., Herman, K., Williams, I.T., Hall, A.J., Cole, D., 2013. Surveillance for foodborne disease outbreaks—United States, 1998–2008. MMWR Surveill. Summ. 62, 1–34.

Grimont, P.A.D. and Weill, F.X., 2007. Antigenic formulae of the *Salmonella* serovars. WHO collaborating centre for reference and research on *Salmonella*, 9.

Guo, C., Hoekstra, R.M., Schroeder, C.M., Pires, S.M., Ong, K.L., Hartnett, E., Naugle, A., Harman, J., Bennett, P., Cieslak, P., 2011. Application of Bayesian techniques to model the burden of human salmonellosis attributable to US food commodities at the point of processing: adaptation of a Danish model. Foodborne Pathog. Dis. 8, 509–16.

Hald, T., Vose, D., Wegener, H.C., Koupeev, T., 2004. A Bayesian approach to quantify the contribution of animal-food sources to human salmonellosis. Risk Anal. 24, 255–69.

Hammer, Ø., Harper, D.A.T., Ryan, P.D., 2009. PAST-PAlaeontological STatistics, ver. 1.89. University of Oslo, Oslo 1–31.

Hedberg, C.W., David, M.J., White, K.E., MacDonald, K.L., Osterholm, M.T., 1993. Role of egg consumption in sporadic *Salmonella* enteritidis and *Salmonella* typhimurium infections in Minnesota. J. Infect. Dis. 167, 107–11.

Hoffmann, S., Fischbeck, P., Krupnick, A., Mcwilliams, M., 2007. Using expert elicitation to link foodborne illnesses in the United States to foods. J. Food Prot. 70, 1220–9.

Hong, S., Rovira, A., Davies, P., Ahlstrom, C., Muellner, P., Rendahl, A., Olsen, K., Bender, J.B., Wells, S., Perez, A., Alvarez, J., 2016. Serotypes and Antimicrobial Resistance in *Salmonella enterica* Recovered from Clinical Samples from Cattle and Swine in Minnesota, 2006 to 2015. PloS ONE, 11(12), p.e0168016.

Kingsley, R.A., Bäumler, A.J., 2000. Host adaptation and the emergence of infectious disease: the *Salmonella* paradigm. Mol. Microbiol. 36, 1006–14.

Kirk, M.D., Pires, S.M., Black, R.E., Caipo, M., Crump, J.A., Devleesschauwer, B., Döpfer, D., Fazil, A., Fischer-Walker, C.L., Hald, T., 2015. World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. PLoS Med 12, e1001921.

Little, C.L., Pires, S.M., Gillespie, I.A., Grant, K., Nichols, G.L., 2010. Attribution of human *Listeria monocytogenes* infections in England and Wales to ready-to-eat food sources placed on the market: adaptation of the Hald *Salmonella* source attribution model. Foodborne Pathog. Dis. 7, 749–56.

Low, S.A., Adalja, A., Beaulieu, E., Key, N., Martinez, S., Melton, A., Perez, A., Ralston, K., Stewart, H., Suttles, S., Vogel, S., Jablonski, B.B.R. 2015. Trends in US local and regional food systems: A report to Congress.

Minnesota Department of Health, 2015. Annual Summary of Communicable Disease Reported to the Minnesota Department of Health, 2015. Disease Control Newsletter, 43.

Muellner, P., Marshall, J.C., Spencer, S.E.F., Noble, A.D., Shadbolt, T., Collins-Emerson, J.M., Midwinter, A.C., Carter, P.E., Pirie, R., Wilson, D.J., 2011. Utilizing a combination of molecular and spatial tools to assess the effect of a public health intervention. Prev. Vet. Med. 102, 242–53.

Mughini-Gras, L., Barrucci, F., Smid, J.H., Graziani, C., Luzzi, I., Ricci, A., Barco, L., Rosmini, R., Havelaar, A.H., Van Pelt, W., 2014. Attribution of human *Salmonella* infections to animal and food sources in Italy (2002–2010): adaptations of the Dutch and modified Hald source attribution models. Epidemiol. Infect. 142, 1070–82.

Mughini-Gras, L. and van Pelt, W., 2014. *Salmonella* source attribution based on microbial subtyping: Does including data on food consumption matter? Int. J. Food Microbiol. 191, 109–15.

Mullner, P., Jones, G., Noble, A., Spencer, S.E.F., Hathaway, S., French, N.P., 2009a. Source attribution of food-borne zoonoses in New Zealand: a modified Hald model. Risk Anal. 29, 970–84.

Mullner, P., Spencer, S.E.F., Wilson, D.J., Jones, G., Noble, A.D., Midwinter, A.C., Collins-Emerson, J.M., Carter, P., Hathaway, S., French, N.P., 2009b. Assigning the source of human campylobacteriosis in New Zealand: a comparative genetic and epidemiological approach. Infect. Genet. Evol. 9, 1311–9.

Painter, J.A., Hoekstra, R.M., Ayers, T., Tauxe, R. V, Braden, C.R., Angulo, F.J., Griffin, P.M., 2013. Attribution of foodborne illnesses, hospitalizations, and deaths to food commodities by using outbreak data, United States, 1998–2008. Emerg. Infect. Dis. 19, 407–15.

Pella, J. and Masuda, M., 2001. Bayesian methods for analysis of stock mixtures from genetic characters. *Fishery Bulletin*, *99*(1), pp.151-167.

Pires, S.M. and Hald, T., 2010. Assessing the differences in public health impact of *Salmonella* subtypes using a Bayesian microbial subtyping approach for source attribution. Foodborne Pathog. Dis. 7, 143–51.

R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Scallan, E., Hoekstra, R.M., Angulo, F.J., Tauxe, R. V, Widdowson, M.-A., Roy, S.L., Jones, J.L., Griffin, P.M., 2011. Foodborne illness acquired in the United States—major pathogens. Emerg. Infect. Dis. 17.

Sears, A., Baker, M.G., Wilson, N., Marshall, J., Muellner, P., Campbell, D.M., Lake, R.J., French, N.P., 2011. Marked campylobacteriosis decline after interventions aimed at poultry, New Zealand. Emerg. Infect. Dis. 17, 1007–15.

Uzzau, S., Brown, D.J., Wallis, T., Rubino, S., Leori, G., Bernard, S., Casadesús, J., Platt, D.J., Olsen, J.E., 2000. Host adapted serotypes of *Salmonella enterica*. Epidemiol. Infect. 125, 229–55.

Wright, A.P., Richardson, L., Mahon, B.E., Rothenberg, R., Cole, D.J., 2016. The rise and decline in *Salmonella enterica* serovar Enteritidis outbreaks attributed to egg-containing

foods in the United States, 1973-2009. Epidemiol. Infect. 144, 810.

Supporting Information

Supporting Figure 1. Posterior distribution of relative prevalence of *S.* Enteritidis in chicken with informative prior density (based on source typing data) shown in red for four different models – A) Single attribution including only sampled sources, B) Single attribution including non-sampled sources, C) Temporal attribution including only sampled sources, and D) Temporal attribution including non-sampled sources.

Supporting Figure 2 Plots of the observed and expected cases for individual *Salmonella* serotypes in the A) Single attribution model including only sampled sources, B) Single attribution including non-sampled sources, C) Temporal attribution including only sampled sources, and D) Temporal attribution including non-sampled sources.

Table 1. Parameter interpretations for the source attribution model.

Parameter	Parameter Interpretation
λ_{ij}	Expected number of human cases of type *i* from source *j*.
q_i	Strain-specific factor for strain *i* (e.g. survivability, pathogenicity to humans, virulence).
a_j	Source-specific factor for source *j* (e.g. risk through typical storage and preparation of food from source *j*).
r_{ij}	The relative prevalence of type *i* in source *j*.

Table 2. Total number of *Salmonella* isolates recovered from different sources and the number of isolates belonging to a serotype found in both human and food/animal sources in Minnesota.

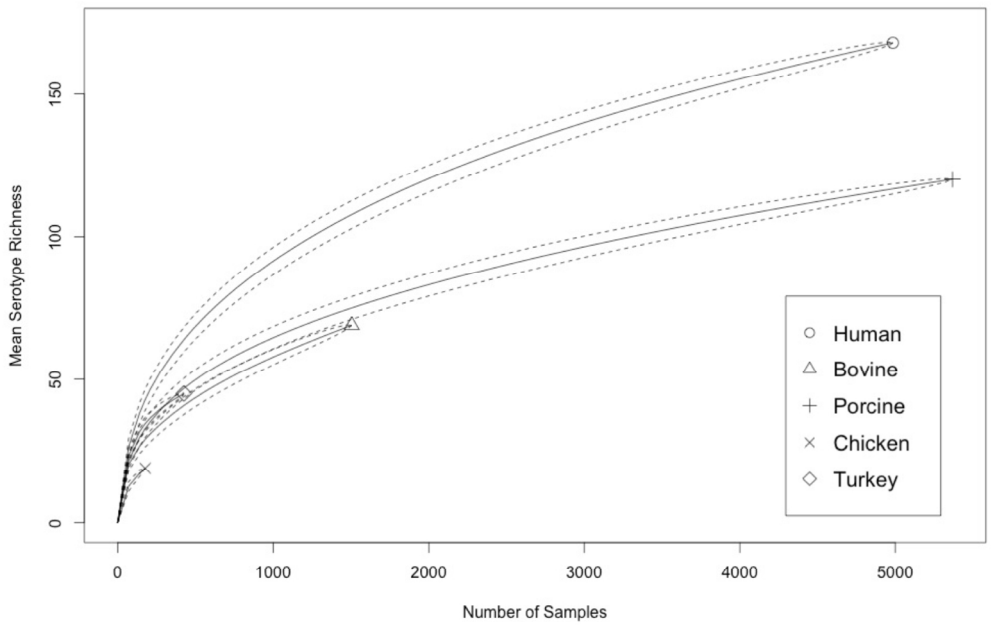
Source	**# isolates in the full dataset**	**# isolates with shared serotypes (% of full dataset)**
Human	4985	4672 (94%) [a]
Porcine	5368	5257 (98%) [b]
Bovine	1505	1484 (99%) [b]
Turkey	426	402 (94%) [b]
Chicken	177	170 (96%) [b]

[a] shared with any other source

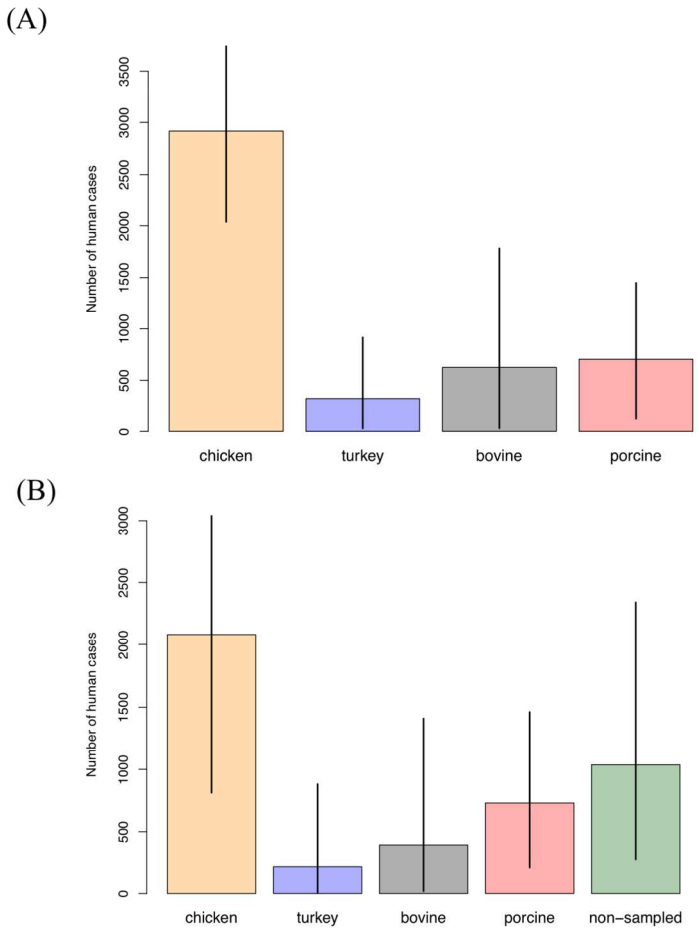
[b] shared with human isolates

Table 3. Proportional similarity index and Simpson's index of diversity with 95% bootstrap CI given in parentheses of *Salmonella* serotypes in Minnesota.

	PSI				Simpson's index of diversity
Source	**Human**	**Bovine**	**Porcine**	**Chicken**	
Human	-				0.92 (0.92, 0.93)
Bovine	0.28 (0.26, 0.30)	-			0.88 (0.87, 0.90)
Porcine	0.36 (0.34, 0.37)	0.23 (0.21, 0.25)	-		0.90 (0.90, 0.91)
Chicken	0.32 (0.24, 0.37)	0.16 (0.12, 0.19)	0.18 (0.15, 0.20)	-	0.73 (0.68, 0.79)
Turkey	0.21 (0.19, 0.24)	0.19 (0.16, 0.22)	0.30 (0.26, 0.32)	0.24 (0.18, 0.27)	0.92 (0.91, 0.93)

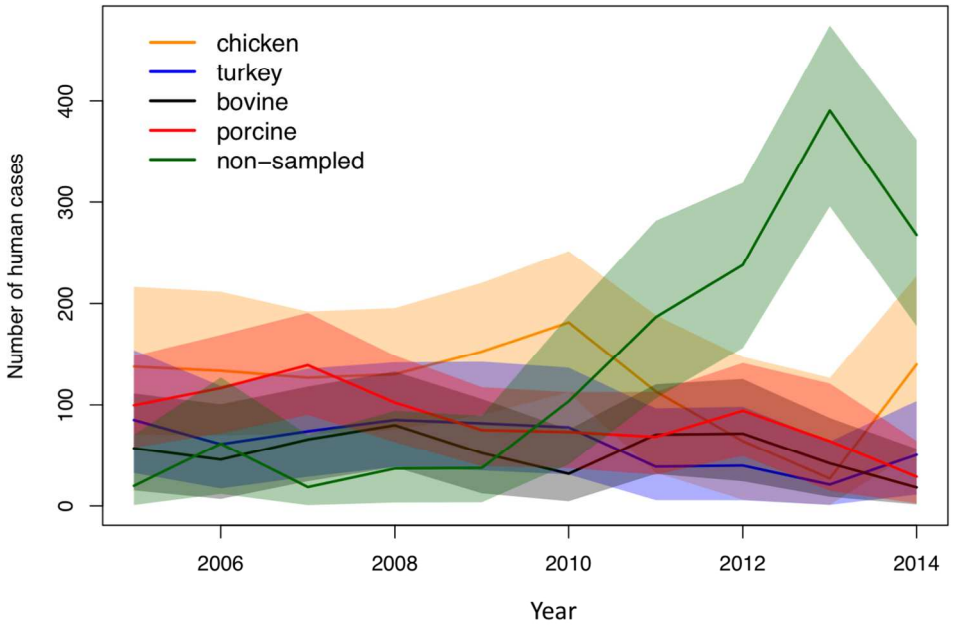


Rarefaction curve indicating the mean serotype richness of Salmonella serotypes from human, bovine, porcine, chicken, and turkey sources in Minnesota.
Figure 1
581x401mm (72 x 72 DPI)



Single attribution results based on 4,672 human salmonellosis cases in Minnesota between 2005 and 2014. The graphs show the number of attributed cases to each source with 95% Bayesian credible intervals without (A) and with (B) the inclusion of a non-sampled source. Year of isolation was not included in this model.

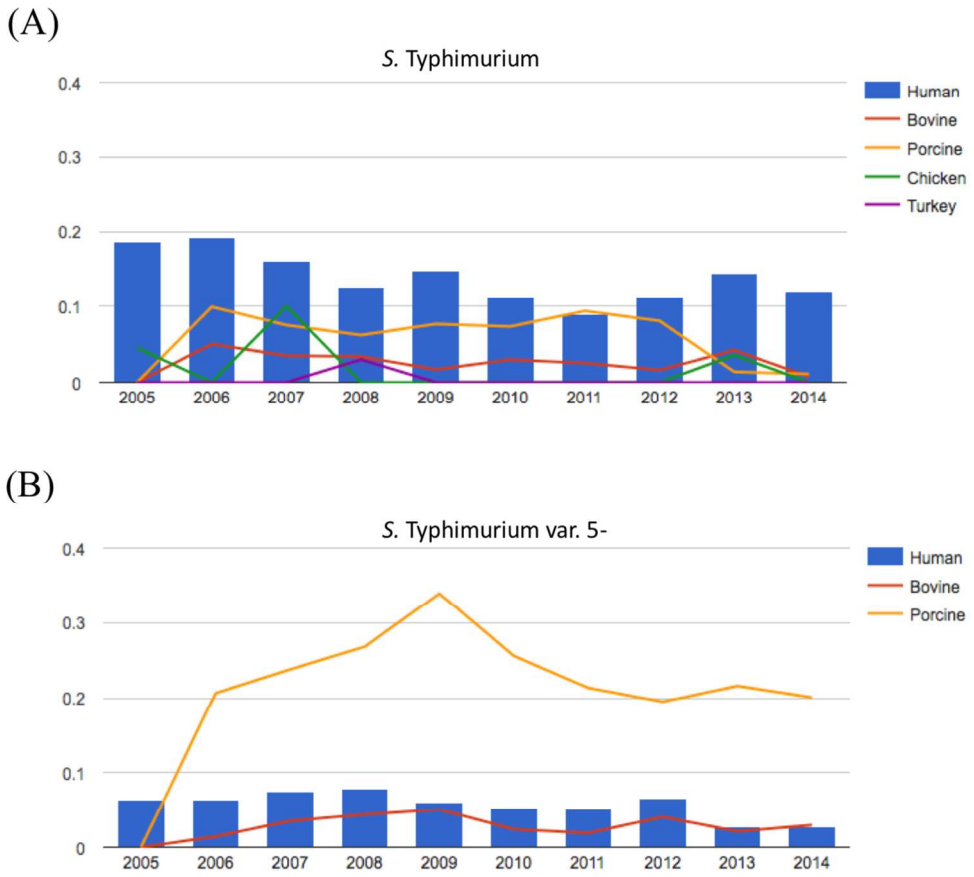
Figure 2
594x792mm (72 x 72 DPI)



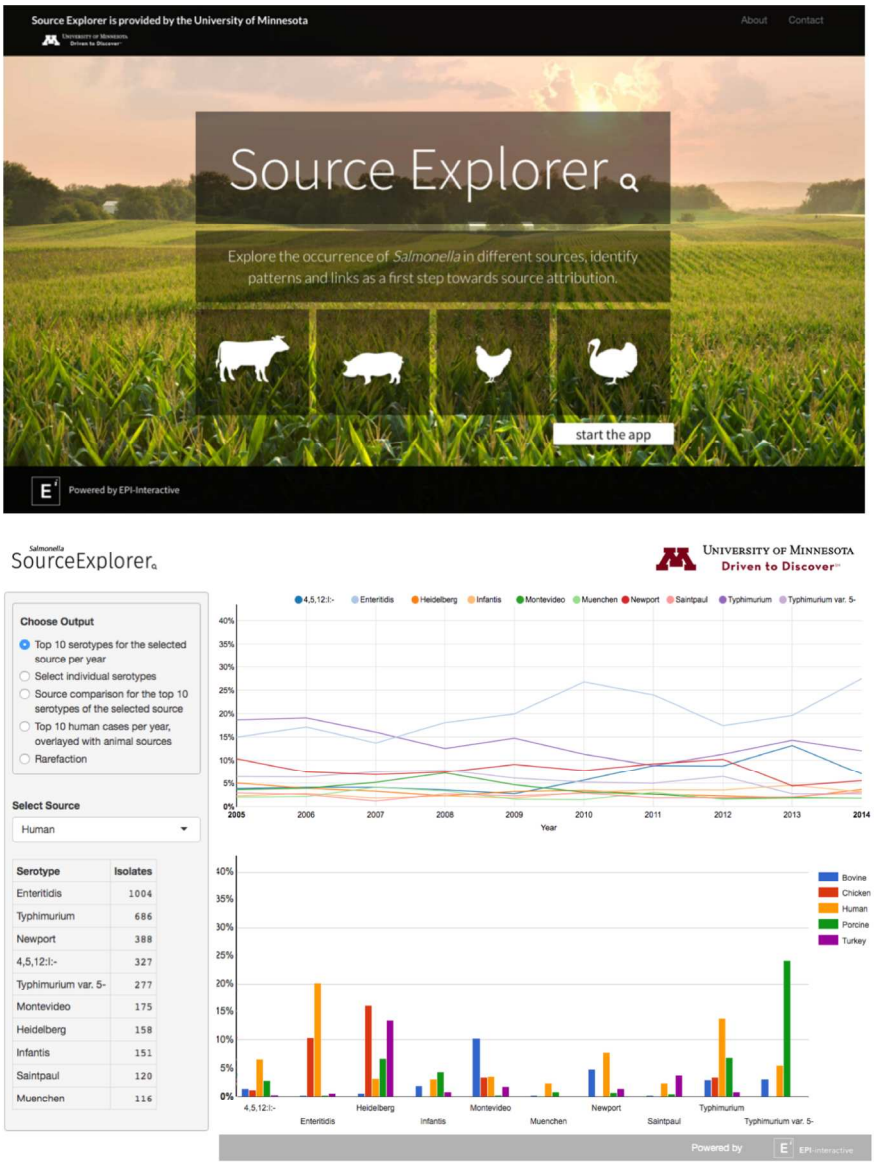
Temporal source attribution results based on 4,672 salmonellosis cases in Minnesota between 2005 and 2014. The graphs show the number of attributed cases to each source with 95% Bayesian credible intervals, incorporating the year in which human isolates were derived.

Figure 3

538x358mm (72 x 72 DPI)



Source Explorer outputs displaying the proportion of S. Typhimurium (A) and S. Typhimurium var. 5- (B) isolates from human and non-human sources in Minnesota between 2005 and 2014.
Figure 4
528x483mm (72 x 72 DPI)



Top panel: Source Explorer landing page. Bottom panel: Screenshot of selected Source Explorer functionalities used to explore source attribution input data and results.

Figure 5

594x792mm (72 x 72 DPI)