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

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

2 Minnesota

3

4 Christina Ahlstrom[1], Petra Muellner[1,] Simon EF Spencer[2], Samuel Hong[3], Amy Saupe[4],

5 Albert Rovira[5], Craig Hedberg[6], Andres Perez[3], Ulrich Muellner[1], Julio Alvarez[3]

6

7 [1] Epi-interactive, Wellington, New Zealand

8 [2] University of Warwick, Coventry, United Kingdom

9 [3] Department of Veterinary Population Medicine, College of Veterinary Medicine, University

10 of Minnesota, St Paul, USA

11 [4] Minnesota Department of Health, Saint Paul, USA

12 [5] Veterinary Diagnostic Laboratory, College of Veterinary Medicine, University of

13 Minnesota, St Paul, USA

14 [6] Division of Environmental Health Sciences, School of Public Health, University of

15 Minnesota, Minneapolis, USA

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

17 petra@epi-interactive.com

18

19　# Summary

20　*Salmonella enterica* is a global health concern because of its widespread association with

21　foodborne illness. Bayesian models have been developed to attribute the burden of human

22　salmonellosis to specific sources with the ultimate objective of prioritizing intervention

23　strategies. Important considerations of source attribution models include the evaluation of the

24　quality of input data, assessment of whether attribution results logically reflect the data

25　trends, and identification of patterns within the data that might explain the detailed

26　contribution of different sources to the disease burden. Here, more than 12,000 non-typhoidal

27　*Salmonella* isolates from human, bovine, porcine, chicken, and turkey sources that originated

28　in Minnesota were analyzed. A modified Bayesian source attribution model (available in a

29　dedicated R package), accounting for non-sampled sources of infection, attributed 4,672

30　human cases to sources assessed here. Most (60%) cases were attributed to chicken, though

31　there was a spike in cases attributed to a non-sampled source in the second half of the study

32　period. Molecular epidemiological analysis methods were used to supplement risk modelling

33　and a visual attribution application was developed to facilitate data exploration and

34　comprehension of the large multi-year dataset assessed here. A large amount of within-source

35　diversity and low similarity between sources was observed and visual exploration of data

36　provided clues into variations driving the attribution modelling results. Results from this

37　pillared approach provided first attribution estimates for *Salmonella* in Minnesota and offer

38　an understanding of current data gaps as well as key pathogen population features, such as

39　serotype frequency, similarity and diversity across the sources. Results here will be used to

40　inform policy and management strategies ultimately intended to prevent and control

41　*Salmonella* infection in the state.

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

43 salmonellosis

44

45 **Bullet points:**

46 • We demonstrate the attribution of salmonellosis to different sources can be enhanced

47 through a pillared approach that incorporates data visualization and molecular

48 epidemiology on top of source attribution modelling. In this study, such

49 supplementary analyses supported findings from the attribution model, demonstrating

50 for example high genotype diversity and low similarity between sources.

51 • We developed a modified Bayesian source attribution model that attributed the

52 majority of salmonellosis cases in Minnesota to chickens. Accounting for a known

53 data gap, we were able to demonstrate the potential impact of a non-sampled source

54 on the number of attributed cases.

55 • A visual attribution application enabled users to dynamically explore the occurrence

56 of Salmonella serotypes in different sources over time. The ability to interact and

57 filter the data assisted in the detection of data irregularities, facilitating accurate

58 attribution estimates and the interpretation of results.

59 Introduction

60 Non-typhoidal *Salmonella enterica* is considered one of the leading causes of foodborne

61 illness in the United States (Scallan et al., 2011) and other countries (Kirk et al., 2015),

62 despite large-scale control efforts initiated by industry, government and consumers (Crim et

63 al., 2015). In the United States, *S. enterica* has the highest reported incidence in humans

64 among bacterial pathogens, with 15.3 cases per 100,000 persons in 2014 (CDC, 2016), the

65 majority of which are sporadic with unknown source of origin (Batz et al., 2005; Ebel et al.,

66 2016). In Minnesota, the incidence of culture-confirmed *Salmonella* cases in 2015 was 17.9

67 per 100,000 persons (Minnesota Department of Health, 2015). The most common form of

68 clinical illness associated with *S. enterica* infection is gastroenteritis, which is mostly non-

69 severe and self-limiting; however, *S. enterica* can cause severe disease or complications in

70 some patients (e.g. sepsis). *S. enterica* has multiple reservoirs, including livestock and

71 domestic pets (Kingsley and Bäumler, 2000), impairing efforts to identify infection sources

72 and transmission routes. Whereas some *S. enterica* serotypes have adapted to individual host

73 species, others are frequently isolated from a broad range of species and environments

74 (Uzzau et al., 2000).

75 Source attribution models aim to estimate the proportion of human cases attributable to

76 specific sources. Estimates obtained can subsequently be used to guide risk managers and

77 decision-makers in the design and implementation of effective intervention strategies (Pires

78 and Hald, 2010; Sears et al., 2011). In the United States, different source attribution methods

79 using a variety of data sources have been employed to estimate the burden of human

80 salmonellosis from several food commodities. For example, the relative proportions of

81 domestically acquired, sporadic human *Salmonella* infections between 1998 and 2003 among

82 multiple food sources was investigated using a molecular subtyping approach (Guo et al.,

4

83   2011), whereas other source attribution assessments relied either on expert elicitation

84   (Hoffmann et al., 2007) or outbreak-associated illnesses (Batz et al., 2012; Gould et al., 2013;

85   Painter et al., 2013). Data collected during outbreak investigations are typically more

86   accessible than sporadic illness data; however, foodborne outbreaks account only for a small

87   proportion of total *Salmonella* infections in humans and may not be representative of the

88   majority of human cases of salmonellosis (Ebel et al., 2016; Painter et al., 2013).

89   Although early risk assessment models estimated the contribution of a single source to the

90   burden of human illness, Bayesian source attribution models for salmonellosis (Hald et al.,

91   2004) were subsequently developed to consider multiple sources simultaneously while

92   allowing for uncertainty around input parameters. Such an approach has been applied

93   internationally and modified over time to accommodate different contexts and pathogens

94   (Mullner et al. 2009a; Little et al. 2010; Mughini-Gras et al. 2014; David et al. 2013; Guo et

95   al. 2011; Glass et al. 2015; Mullner et al. 2009b). Briefly, the model framework compares the

96   distribution of pathogen serotypes in the source populations to the distribution of serotypes

97   observed in humans via a Poisson regression model fitted within a Bayesian framework.

98   While models developed thus far have provided valuable insights into the contribution of

99   different sources to the human disease burden, they generally rely on a specific type of data

100   that may not always be available. Previous work has, for example, explored the use of

101   different types of prevalence data (Mullner et al. 2009a). The work presented her proposes a

102   further extension that may allow the model to be successfully used in the presence of data

103   gaps.

104   Bayesian source attribution models often rely on multi-level data collected in multiple years

105   and sources, requiring significant resources to manually clean and prepare the data. Model

106   outputs are typically one-dimensional and primarily focus on one quantitative figure (i.e. the

107   attributed proportion of human cases to each source). A comprehensive understanding of the

108   characteristics of the input data and results could enhance the way attribution results are

109   understood by, and communicated to, risk managers and other stakeholders. For example,

110   molecular epidemiological analyses, such as diversity statistics and similarity indexes

111   (Muellner et al., 2011), can be incorporated into the analysis to support model outputs.

112   Additionally, the integration of visual data exploration creates opportunities to dynamically

113   interact with layered data and identify trends to help explain attribution results. Such an

114   approach can help make large, complex datasets digestible through interactive graphics that

115   facilitate incremental exploration of the data (Carroll et al., 2014).

116   The objective of the study presented here was to explore the feasibility of using a pillared

117   attribution approach, adding molecular epidemiology and visual data analysis, to a

118   customised Hald Model based on the work by Mullner et al. (2009), to generate preliminary

119   estimates of the source-specific burden of salmonellosis in Minnesota over a ten-year period

120   using state-level data. Results here will help to inform policy and management activities

121   intended to prevent and control the disease in the state. Additionally, methods presented here

122   may serve as a framework that could be applied to the attribution of sources of infection for

123   *Salmonella* and other foodborne pathogens in the United States and other regions.

124

Materials and methods

126   Data sources

127   Information from human salmonellosis cases reported to the Minnesota Department of Health

128   (MDH) from 2005 to 2014 were collected. Data included case serotype (determined by the

129   MDH Public Health Laboratory [PHL]), date of specimen collection, international travel in

130   the seven days prior to illness onset (self-reported by cases during routine exposure

131   interviews at MDH), and whether the case was part of an identified outbreak. Consistent with

132   the approach developed by Hald et al (2004), cases with a history of international travel were

133   excluded, as were cases attributed to an outbreak.

134   Information on *Salmonella* isolates of food animal (cattle, swine, poultry) origin isolated by

135   the Minnesota Veterinary Diagnostic Laboratory (MVDL) from diagnostic submissions

136   between 2006 and 2015 were also collected. A detailed description of the MVDL database

137   and cattle and swine isolates is available elsewhere (Hong et al., 2016). Finally, data on the

138   relative frequency of *Salmonella* serotypes stratified per meat product (chicken, ground

139   turkey, ground beef, pork chop) from various retail locations in Minnesota that were collected

140   as part of the FoodNet/National Antimicrobial Resistance Monitoring System (NARMS)

141   Retail Food Study between 2002 and 2013 were provided by MDH. These food commodity

142   source data represented 4% of isolates derived from non-human sources and were combined

143   with the animal data (e.g. isolates from pork chops were added to animal porcine isolates).

144   The majority (96%) of food-derived isolates came from chicken breasts and ground turkey.

145   Non-human source categories were limited to bovine, porcine, chicken, and turkey, because

146   typed isolates from other sources were considered too scarce (n < 50) to be included in the

147   risk model. Further, no data on eggs were available.

148   The consistency of serotype naming was checked within and between data sources. All

149   serotypes were defined using naming conventions proposed by the Pasteur Institute (Grimont

150   and Weill, 2007). After 2012, the Centers for Disease Control and Prevention (CDC)

151   recommended naming all *S.* Typhimurium var. 5 – (formerly var. Copenhagen) as *S.*

152   Typhimurium (*S.* I 4,[5],12:i:1,2; CDC, 2014). Thus, the MDH did not electronically record

153   variants of *S.* Typhimurium after 2011; however, variants continued to be identified and

154 recorded in paper records. Therefore, *S.* Typhimurium var. 5– (formerly var. Copenhagen)

155 cases were retrospectively recorded electronically in the current dataset.

156 Diversity and similarity statistics

157 The dataset containing all *Salmonella* isolates from bovine, porcine, chicken, turkey, and

158 sporadic and domestic human cases was used to calculate diversity and similarity statistics.

159 Serotype richness and the diversity of *Salmonella* serotypes from different sources was

160 estimated using rarefaction and Simpson's index of diversity. The function RAREFY in the

161 package VEGAN (version 2.2-3) was implemented in R (version 3.2.3; R Core Team 2016)

162 and Simpson's index of diversity was calculated in Past (version 3.10; Hammer et al. 2009),

163 with 9,999 bootstrap replicates. The similarity of serotypes between different sources was

164 investigated by calculating the proportional similarity index (PSI), which measures the area

165 of overlap between two frequency distributions of each serotype between sources. Bootstrap

166 confidence intervals were estimated as described by Mullner et al. (2009).

167 Source attribution modelling

168 Data preparation

169 *Salmonella* isolates originating from non-human sources that did not match a serotype found

170 in Minnesota human cases during the study period were excluded. Similarly, *Salmonella*

171 isolates originating from humans that did not match a serotype found in non-human sources

172 were excluded. Serotypes that included fewer than five human cases were combined into a

173 serotype designated as "other", similar to the approach taken in previous studies (Hald et al.,

174 2004; Mullner et al., 2009a; David et al., 2012). Sampling effort of non-human sources varied

175 across the study period; thus, isolates from non-human sources were not segregated into

176 years. Human isolates were summarized as total cases per serotype per year between 2005-

177 2014.

8

178   Attribution model

179   The source attribution model described by Hald et al. (2004) and modified by Muellner et al.

180   (2009) was further modified here to simplify and improve applicability and interpretability of

181   results. The approach proposed here may be used in a wider variety of circumstances

182   compared to earlier methods, in particular, without knowledge of the absolute prevalence in

183   each source.

184   The Poisson regression structure was retained (adapted slightly to include time)

185    $Y_{it} \sim Pois(\sum_{j=1}^{S} \lambda_{ij})$, where Y_{it} is the number of serotype i human cases in year $t = 1,...,T$.

186   The equation defining the expected number of cases of serotype i (for $i=1,...,I$) from source j

187   (for $j=1,...,S$) was modified as

188
$$\lambda_{ij} = r_{ij}q_i a_j.$$

189   Model parameters are described in Table 1. Note that this equation does not require the

190   absolute source prevalences p_{ij} or π_j (Mullner et al., 2009a) and instead uses just the relative

191   prevalences r_{ij}, plus the strain i- and the source j-specific factors q_i and a_j, respectively. This

192   model can therefore be fitted when source prevalence data are not available. Let X_{ij} denote

193   the number of serotype i isolates observed in source j. The following prior distributions were

194   assumed:

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

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

197    $a_j \sim$ Exp(α_j)

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

9

199 distribution of the random effects describing the strain-specific differences q_i. In previous

200 versions of the source attribution model, the prior for the source-specific parameter a_j was

201 difficult to specify. Here, because the prior mean of q_i is one and the relative prevalences sum

202 up to one, the prior mean of a_j (which equals α_j^{-1}) is the expected number of cases from

203 source j. This interpretation allows informed priors for the source-specific factors to be easily

204 defined.

205 In the current implementation, we chose $\alpha_j^{-1} = \sum_{i=1}^{I} \sum_{t=1}^{T} Y_{it}/TS$ for all $j = 1,..., S$, which

206 implies the prior belief that an equal number of cases comes from each source and that cases

207 appear at the average rate observed in the data. That procedure was followed to overcome

208 concerns regarding the choice of uniform priors with fixed boundaries as described by Hald

209 et al. (2004), as the inferences have been shown in some cases to be sensitive to the choice of

210 boundaries (Glass et al., 2015). For the remaining hyperparameters we chose $\gamma_i = 1, \alpha^{(\theta)} =$

211 $\beta^{(\theta)} = 1$ to reflect weak prior knowledge.

212 The distribution of random effects was changed from a log-normal distribution with a mean

213 of 1 to a *Gamma* distribution with a mean of 1, which allows a Gibbs step to be used to

214 update q_i during the Markov chain Monte Carlo (MCMC) (Denison, 2002). The full

215 conditional distribution was given by

216 $q_i|(\theta, r, a, Y) \sim \text{Gamma} \left(\theta + \sum_{t=1}^{T} Y_{it}, \theta + T \sum_{j=1}^{S} r_{ij} a_j\right)$. The hyperparameter θ can be

217 fixed, but we chose to specify a prior for θ so that the variance of the random effects was

218 estimated from data.

219 The relative prevalence parameters r_{ij} were updated using Metropolis-Hastings updates

220 (Chib and Greenberg, 1995) with proposals from the prior distribution in both large and small

221 blocks. The large blocks consisted of all types $i = 1, \ldots, I$ for a given source and the small

222 blocks consisted of just two types for a given source. The source dependent factors a_j were

223 updated jointly using Metropolis-Hastings random walk proposals on the log scale. To avoid

224 poor mixing when the values of a_j were small, single component Metropolis-Hastings

225 proposals from the prior distribution were applied. The model described above is considered

226 a single attribution model because it assumes that the source attribution is the same for each

227 year in the dataset.

228 In the temporal source attribution model, the source dependent factors depend on time and

229 therefore produce a different source attribution for each year in the study. Given that the

230 source data are sparser than the human data, and they were not collected throughout the

231 whole study period, relative prevalences r_{ij} that do not depend on time were used.

232 The temporal attribution model becomes $Y_{it} \sim Pois(\sum_{j=1}^{S} \lambda_{ijt})$, for year $t = 1, \ldots, T$ and

233 type $i = 1, \ldots, I$. The mean number of cases is decomposed as

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

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

236 equally between sources as before.

237 Convergence was confirmed for all parameters by visual inspection of the trace plots and

238 comparison of the posteriors from chains with randomly chosen starting values. Once

239 convergence was established, any evidence of lack-of-fit in the source attribution model was

240 likely due to human cases that were difficult to attribute to any of the sources in the model.

241 To help investigate that feature, an additional source, referred to as a non-sampled source,

242 was included in the model (for which no source data were observed) to identify the quantity

243 and profile of unattributable cases. To further assess the model performance, data from a

244 previous *Campylobacter* source attribution analysis (Mullner et al., 2009a) was analyzed and

245   outputs were compared.

246   The model was run in R (version 3.2.3; R Core Team, 2016) and to facilitate dissemination,

247   an associated R package is under development. This package containing the updated model

248   includes adaptive proposals so that unreasonable mixing is detected and the proposals are

249   adjusted to combat it. Furthermore, the package includes functions to aid interpretation of the

250   output by non-specialists.

251   Visual attribution application development

252   An application was created to allow users to dynamically investigate the occurrence of

253   *Salmonella* serotypes in different sources and identify patterns in the data. This application,

254   named *Source Explorer*, was developed in an RStudio Shiny (http://shiny.rstudio.com)

255   framework, which enabled the tool to be used locally, as a stand-alone version, or through a

256   web-based interface. The dataset of *Salmonella* serotypes per source was used as input data,

257   including the years in which the isolates were collected. Five main visualization outputs were

258   displayed, including 1) the proportion of the top 10 serotypes of each selected source over

259   time, 2) user-selected serotypes for the selected source over time, 3) a bar chart comparing

260   the top 10 serotypes of the selected source with all other sources, 4) the top 10 human

261   serotypes per year overlayed with non-human sources, and 5) rarefaction results with options

262   to select a subset of sources. The proportion of isolates per source and year was displayed to

263   account for different sampling efforts during the study period.

264

265   Results

266   *Salmonella* diversity

267   Table 2 shows the total number of *Salmonella* isolates derived from human, porcine, bovine,

268   turkey and chicken sources in the full dataset and the number of isolates with serotypes found

269   in both human and non-human sources. A total of 991 and 749 international travel and

270   outbreak associated cases were excluded, respectively. An additional 330 cases were further

271   excluded, as they were either *S.* Typi, *S.* Paratyphi, or of an unknown serotype. The number

272   of sporadic and domestically acquired human cases per year in the dataset ranged from a

273   minimum of 405 cases in 2009 to a maximum of 546 cases in 2013. In total, 240 *Salmonella*

274   serotypes were found, 156 (65%) of which were isolated from a single source type. Twelve

275   (5%) serotypes were found in all five sources. Serotype *S.* Typhimurium var. 5- (formerly

276   known as *S.* Copenhagen) had the largest total number of isolates, with 1,617 isolates from

277   human, bovine, and porcine sources.

278   Rarefaction analysis indicated a larger serotype richness in human *Salmonella* isolates than in

279   those from non-human sources (Figure 1). In addition, a lower level of diversity was found

280   among isolates from chickens than those from other sources. Simpson's index of diversity of

281   isolates from each source is presented in Table 3. *Salmonella* isolates from human, turkey,

282   and porcine sources had the highest serotype diversity, whereas isolates from chicken and

283   bovine sources were less diverse. Overall, the PSI analysis indicated a relatively low level of

284   similarity between the different sources (Table 3). Human isolates were most similar to

285   porcine isolates, however 95% bootstrap confidence intervals overlapped between all

286   sources.

287

288   Source attribution

289   Seventy-five different serotypes were found both in humans and in at least one of the non-

290   human sources, 50 of which included at least five human cases, representing 96% of isolates

291   included in the full dataset. The remaining 25 serotypes included 785 isolates and were

292   designated as "other".

293   The number of human salmonellosis cases attributed to each source in the single attribution

294   model is shown without (Figure 2A) and with (Figure 2B) inclusion of a non-sampled source.

295   Out of 4,672 human cases, the largest share was attributed to chickens, accounting for 2,790

296   (60%) and 2,049 (45%) without and with a non-sampled source, respectively.

297   Results from the temporal attribution model are shown in Figure 3. A marked change in the

298   attribution estimates after 2009 was observed, with an increase in the number of cases

299   attributed to the non-sampled source. The human cases most frequently assigned to the non-

300   sampled source by the model included serotypes *S.* 4,5,12:i:- (8.8%), *S.* Enteritidis (7.6%), *S.*

301   Berta (6.1%), *S.* Infantis (5.9%) and *S.* Heidelberg (4.0%).

302   The temporal model generally showed sufficient fit, aligned with previous validations of the

303   model where expected and observed cases were compared to test the validity of the model

304   outputs (e.g. Hald et al. 2004). A more stringent model assessment was also performed in

305   which the prior and the posterior distributions were compared for each of the relative

306   prevalence parameters. The prior for these parameters is based on the source data whilst the

307   posterior incorporates both the source and human data, and so if there is disagreement

308   between these two distributions then some of the human data cannot be aligned with any of

309   the sources. A selection of these plots is presented in Supporting Figure 1. Outputs from the

310   Mullner et al. (2009a) data using the current model (data not shown) were consistent with the

311   results presented in the original paper, further supporting the performance of the model

312   presented here.

313

Visual attribution

315   Multiple filters and visualization options facilitated extensive data exploration in *Source*

316   *Explorer*. Outputs were visualized in charts and data tables and were instantaneously updated

317   as search variables were modified. Individual serotypes or the top 10 serotypes for each

318   source could be selected and visualized over time. For example, the proportion of human

319   isolates belonging to each of the top 10 serotypes from humans was overlayed with the

320   proportion of isolates from each source (Figure 4). This example demonstrates the high

321   proportion of porcine *S.* Typhimurium var. 5- isolates throughout the study period, compared

322   to other sources.

323   *Source Explorer* enabled us to further investigate the source attribution results and potential

324   reasons for the spike in cases attributed to a non-sampled source after 2009. Out of the top 10

325   most common human serotypes in our dataset, *S.* Enteritidis and *S.* 4,5,12:i:- increased

326   relative to the other serotypes isolated after 2009 (Figure 5). Further, these two serotypes

327   were not frequently isolated from non-human sources throughout the study period,

328   representing 0.5 and 2.5 percent of serotypes isolated from non-human sources, respectively.

329

Discussion

331   Source attribution models continue to be developed and modified to accommodate different

332   environments, pathogens, and data sources (David et al., 2013; Glass et al., 2015; Guo et al.,

333   2011; Hald et al., 2004; Little et al., 2010; Mughini-Gras et al., 2014; Mullner et al., 2009a).

334   A detailed exploration of the data that are inputted into such models can help drive their

335   evolution to best incorporate and model observed data. Therefore, a multi-pillared approach

336   was used here to explore the burden of sporadic, domestically-acquired human *Salmonella*

337   infections in Minnesota from different sources over a ten-year period. Serotype diversity and

338   similarity in human, bovine, porcine, chicken, and turkey sources was assessed and a visual

339   attribution tool was developed to facilitate data exploration and validation. A modified

340   Bayesian source attribution model was subsequently used to estimate the relative contribution

341   of *Salmonella* isolates from alternative sources.

342   *Source Explorer* enabled users to visually assess and interact with the large datasets

343   commonly included in source attribution analyses. Attribution results here were not only

344   clarified through data exploration, but this visualization tool also helped to identify odd

345   trends in the data, such as drastic changes in the proportion of common serotypes over time.

346   For example, a lack of human *S.* Typhimurim var. 5- isolates after 2011, whereas that

347   serotype previously accounted for an average of 30% of all human *Salmonella* serotypes,

348   suggested a change in 2012. Such a finding led to further investigation into serotype naming

349   convention changes by the CDC and the state public health laboratory (CDC, 2014) and

350   ultimately led to the retrospective electronic coding of human *S.* Typhimurim var. 5- cases

351   after 2011. Quality of input data is critical to provide accurate attribution results that best

352   explain the burden of human disease. A thorough description of data and whether trends

353   accurately reflect attribution estimates is an important step that should be included in any

354   attribution analysis. That step further ensures a comprehensive and transparent assessment

355   that can be more easily communicated to various stakeholder groups to build awareness and

356   engagement and ultimately support the development of control strategies (Carroll et al.,

357   2014).

358   The highest number of sporadic and domestic cases were attributed to chickens in the single

359   attribution analyses, both with and without the inclusion of a non-sampled source. In the

360   temporal analysis, the non-sampled source had the most attributed cases, followed by

361   chickens. That finding supports previous attribution studies in the United States, where

362   poultry was the leading source of human salmonellosis (Batz et al., 2012; Gould et al., 2013;

363   Guo et al., 2011; Hoffmann et al., 2007). The temporal attribution results showed a surprising

364   amount of smoothness over time in the early part of the study, particularly considering that

365   no smoothing factor was incorporated into the model. That time period is also when most of

366   the data originated from some sources and so it may well be the case that changes in

367   attribution seen after 2010 may be the result of the source data becoming out of date rather

368   than such substantial changes in the number of cases coming from each source.

369   The wide credible intervals in Figure 2b show that there is a large amount of uncertainty in

370   the estimates regarding the dominant source in the single attribution model, with poultry and

371   non-sampled being the two largest estimates. In the temporal attribution model, there is a

372   period where poultry appears to make the largest contribution and a period where non-

373   sampled appears to make the largest contribution. Given that the single attribution model

374   cannot incorporate temporal changes, except as Poisson variation in the number of cases, in

375   this study, we conclude that the single attribution model does not fit as well as the temporal

376   attribution model, which results in the differences observed. This was further supported by

377   the observed vs. expected plot (Supporting Figure 2), which shows more variation in the

378   single attribution model between years than predicted by the Poisson distribution. Better

379   agreement can be observed between observed and expected cases in the temporal model.

380   Remaining deviations might indicate remaining lack of fit, but could also be explained by

381   random variation in source sampling.

382   As only four non-human sources had sufficient numbers of isolates to be included here, there

383   was likely an overestimation of the burden of illness from these sources. The notable

384   exclusion of eggs as a source may have skewed the attribution estimates, as eggs are a known

385   source of *Salmonella*, particularly *S.* Enteritidis (Hedberg et al., 1993; Wright et al., 2016).

386   Indeed, *S.* Enteritidis did not show a good fit between observed and posterior cases (data not

387   shown) in the current study. A European Union Food Safety Authority report ((BIOHAZ)

388   EFSA Panel on Biological Hazards 2013) recently highlighted the need to include isolates

389   from all potential major non-human sources and that the use of surrogate data or the

390   exclusion of relevant sources may seriously bias attribution results. The inclusion of a non-

391   sampled source in this study was a novel approach in source attribution analyses to deal with

392   the common problem of missing source data; however, it was previously performed as an

393   exploratory technique in other contexts and further work is necessary to confirm the validity

394   of such an approach (Pella and Masuda, 2001). The consistent results obtained using data

395   from a previous source attribution analysis (Mullner et al., 2009a) encourages the use of this

396   model in additional settings. The availability of the model in a dedicated R package will

397   facilitate this.

398   The supplementary (molecular) epidemiological analysis supported findings from the

399   attribution model. Diversity and similarity statistics highlighted a high degree of diversity and

400   low similarity of *Salmonella* serotypes isolated from all sources. Rarefaction curves for each

401   of the sources did not appear to reach a plateau, indicating that the serotype richness was not

402   fully captured by the current dataset and increased sampling effort in all sources could

403   improve the model fit. Serotype richness was greatest in humans, even when accounting for

404   sampling effort, and the PSI results indicated that overall similarity between sources was

405   relatively low. Low similarity between non-human sources can be advantageous in an

406   attribution analysis, as it supports the attribution of specific serotypes to those sources in

407 which it is uniquely found (Barco et al., 2013). In a previous study evaluating *Campylobacter*

408 spp. in New Zealand (Mullner et al., 2009b), a higher similarity among sources and humans

409 was found, with similarity estimates ranging from 0.18 to 0.58.

410 The most appropriate subtyping method for source attribution is one that provides an

411 appropriate level of discrimination to define subtypes associated with specific sources (EFSA

412 Panel on Biological Hazards, 2013). Serotyping is a common method for differentiation of

413 *Salmonella* types, though molecular typing methods (e.g. multiple-locus variable-number

414 tandem repeat analysis, pulsed-field gel electrophoresis) have also been used (Barco et al.,

415 2013). In some attribution studies, authors adjusted the serotype nomenclature, minimizing

416 the total number by combining variants of the same serotype. For example, Guo et al. (2011)

417 combined variants into the non-variant serotypes. This was not performed here because

418 serotype variants were assumed to not change during the course of transmission and

419 sampling. Further supporting this decision, *S.* Typhimurium and Typhimurium var. 5- were

420 the two most commonly isolated *Salmonella* serotypes, yet were differentially distributed

421 among sources in this study, with var 5- known to show some host association to swine

422 reservoirs; hence, that feature provided critical anchor points for the attribution model, which

423 relies on host association of subtypes. In consequence, the differentiation of the large number

424 of *S.* Typhimurium isolates into two distinct serotypes in this study improved the model fit.

425 Minnesota data were exclusively used in the source attribution analysis, despite much of the

426 food consumed in Minnesota likely originated from outside the state. Nevertheless, there is

427 an increasing trend in the consumption of locally produced foods, as consumers seek direct

428 farm to retail options (Low et al., 2015). Local production, import, and export data have also

429 been included in an attribution analysis to account for the flow of food commodities across

430 borders (De Knegt et al., 2015), revealing that individual countries within the European

431 Union had different attribution estimates. Further efforts in expanding data collection at a

432 state level in the United States could similarly help to elucidate spatial patterns in attribution.

433 Inclusion of data on geographical origin of isolates from retail foods, such as where the foods

434 were processed and purchased, could also potentially refine the current analysis. An

435 expanded data collection in the United States should ideally include a large number of

436 samples originating directly from food sources or food processing environments, as opposed

437 to animal reservoirs, which is where the majority of non-human isolates in this study were

438 derived. Nevertheless, given the lack of data from food sources, such data from animal

439 sources are commonly used in source attribution analyses (Mughini-Gras and van Pelt, 2014).

440 As with any analysis that relies on reported cases, underreporting of salmonellosis may have

441 introduced some bias into this study.

442 In summary, results here demonstrated an enhanced approach to source attribution that

443 encourages data exploration through diversity statistics and visual attribution both prior to

444 and after the use of a Bayesian source attribution model. Results here will help to inform

445 preventive and control strategies for *Salmonella* infection in Minnesota.

446

Acknowledgements

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.

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

488   Gould, L.H., Walsh, K.A., Vieira, A.R., Herman, K., Williams, I.T., Hall, A.J., Cole, D.,
489   2013. Surveillance for foodborne disease outbreaks—United States, 1998–2008.
490   MMWR Surveill. Summ. 62, 1–34.

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

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

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

499   Hammer, Ø., Harper, D.A.T., Ryan, P.D., 2009. PAST-PAlaeontological STatistics, ver.
500   1.89. University of Oslo, Oslo 1–31.

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

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

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

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

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

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

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

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

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

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

533   Mughini-Gras, L. and van Pelt, W., 2014. *Salmonella* source attribution based on microbial
534   subtyping: Does including data on food consumption matter? Int. J. Food Microbiol.
535   191, 109–15.

536   Mullner, P., Jones, G., Noble, A., Spencer, S.E.F., Hathaway, S., French, N.P., 2009a. Source
537   attribution of food-borne zoonoses in New Zealand: a modified Hald model. Risk Anal.
538   29, 970–84.

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

543   Painter, J.A., Hoekstra, R.M., Ayers, T., Tauxe, R. V, Braden, C.R., Angulo, F.J., Griffin,
544   P.M., 2013. Attribution of foodborne illnesses, hospitalizations, and deaths to food
545   commodities by using outbreak data, United States, 1998–2008. Emerg. Infect. Dis. 19,
546   407–15.

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

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

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

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

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

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

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

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

566

Supporting Information

568   Supporting Figure 1. Posterior distribution of relative prevalence of *S.* Enteritidis in chicken

569   with informative prior density (based on source typing data) shown in red for four different

570   models – A) Single attribution including only sampled sources, B) Single attribution

571   including non-sampled sources, C) Temporal attribution including only sampled sources, and

572   D) Temporal attribution including non-sampled sources.

573

574   Supporting Figure 2 Plots of the observed and expected cases for individual *Salmonella*

575   serotypes in the A) Single attribution model including only sampled sources, B) Single

576   attribution including non-sampled sources, C) Temporal attribution including only sampled

577   sources, and D) Temporal attribution including non-sampled sources.

578

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.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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
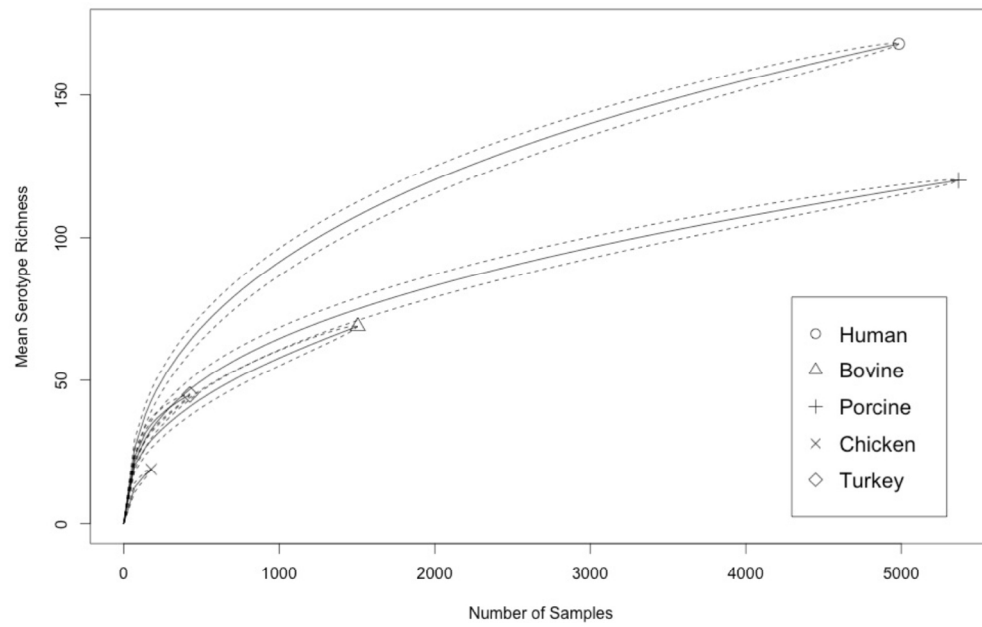
Minnesota.

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 3. Proportional similarity index and Simpson's index of diversity with 95% bootstrap CI given in parentheses of *Salmonella* serotypes in Minnesota.
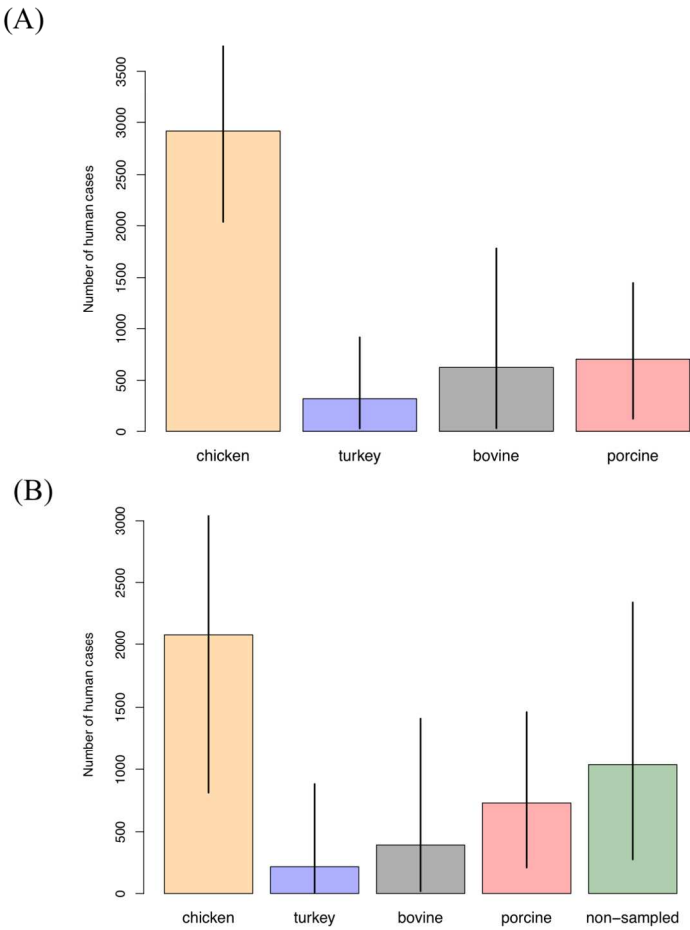
Source	PSI				Simpson's index of diversity
	Human	Bovine	Porcine	Chicken	
Human	-				0.92 (0.92, 0.93)
Bovine	0.28 (0.26, 0.30)	-			0.88 (0.87, 0.90)
Porcine	0.36 (0.34, 0.37)	0.23 (0.21, 0.25)	-		0.90 (0.90, 0.91)
Chicken	0.32 (0.24, 0.37)	0.16 (0.12, 0.19)	0.18 (0.15, 0.20)	-	0.73 (0.68, 0.79)
Turkey	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)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

(A)

(B)

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.
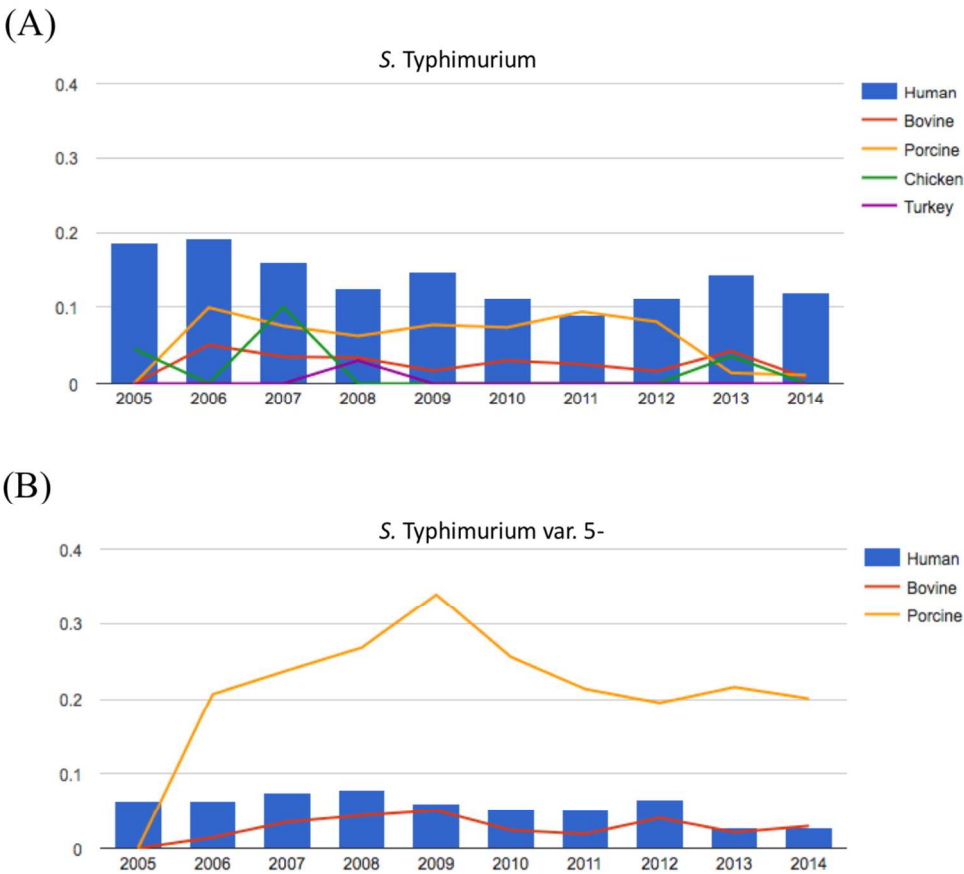
Figure 2

594x792mm (72 x 72 DPI)
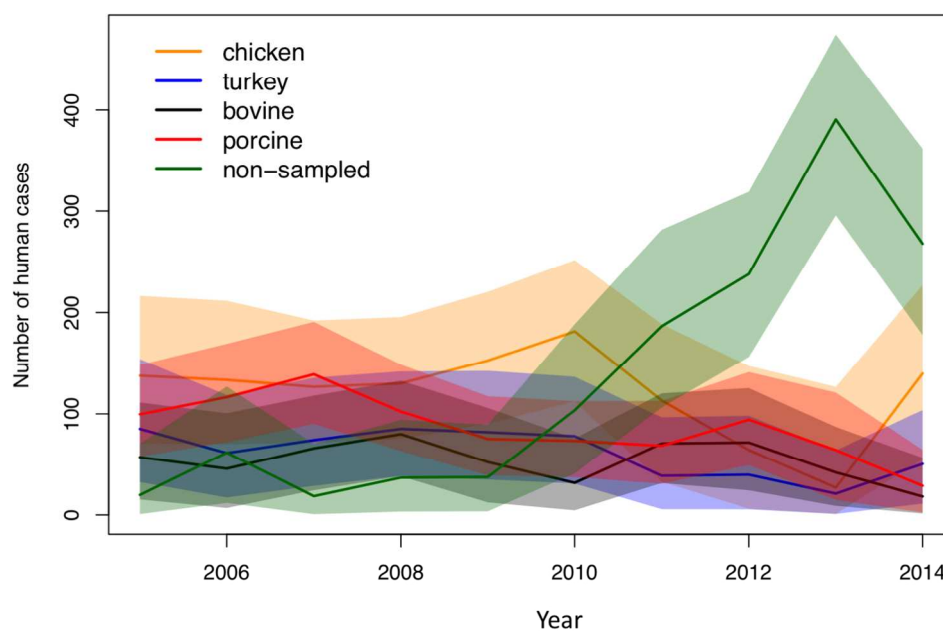
Temporal source attribution results based on 4,672 salmonellosis cases in Minnesota between 2005 and 2014. The graphs show the number of attributed cases to each source with 95% Bayesian credible intervals, incorporating the year in which human isolates were derived.
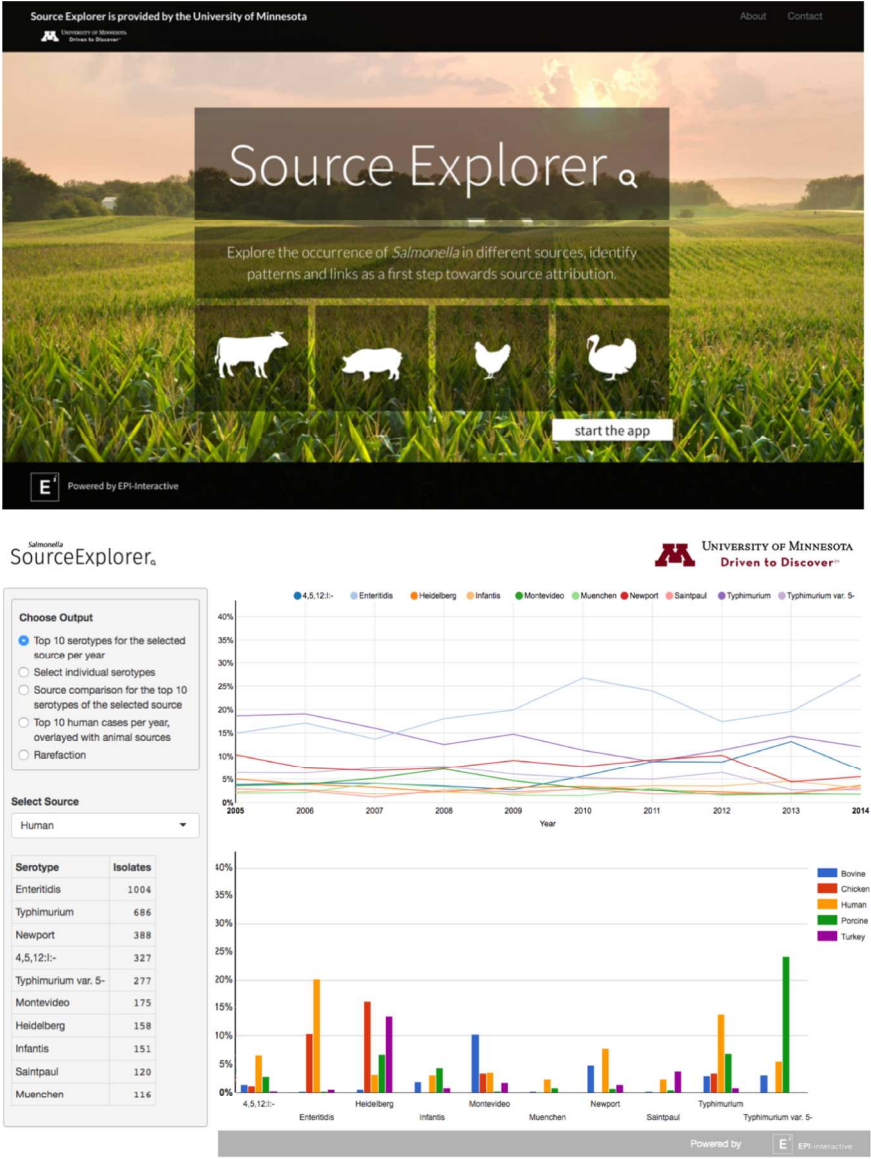
Figure 3

538x358mm (72 x 72 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

(A)

(B)

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)