# Salmonella source attribution based on microbial subtyping: Does including data on food consumption matter?

Lapo Mughini-Gras *, Wilfrid van Pelt

*National Institute for Public Health and the Environment (RIVM), Centre for Infectious Disease Control (CIb), Epidemiology and Surveillance Unit, Bilthoven, The Netherlands*

## ARTICLE INFO

## ABSTRACT

Source attribution based on microbial subtyping is being performed in many countries to ascertain the main reservoirs of human salmonellosis and to assess the impact of food safety interventions. To account for differences in exposure, the amount of food available for consumption within a country is often included in *Salmonella* source attribution models along with the level of contamination. However, not all foods have an equal probability of serving as vehicles for salmonellas, as some foods are more likely to be consumed raw/undercooked than others, posing a relatively higher risk. Using *Salmonella* data from the Netherlands in 2001–2004, this study aims at elucidating whether and how the incorporation of food consumption data in two source attribution models – the (modified) Dutch and Hald models – affects their attributions. We also propose the incorporation of an additional parameter to weight the amount of food consumed by its likelihood to be consumed raw/undercooked by the population. Incorporating the amount of food consumed caused a drastic change in the ranking of the top reservoirs in the Dutch model, but not in the Hald model, which proved to be insensitive to additional weightings given that its source-dependent factor can account for both food consumption and the ability for foods to serve as vehicles for salmonellas. Compared to attributions without food consumption, the Dutch model including the amount of food consumed showed an increase in the percentage of cases attributable to pigs and a decrease in that of layers/eggs, which became the second reservoir, after pigs. This was not consistent with established knowledge indicating that layers/eggs, rather than pigs, were the main reservoir of human salmonellosis in that period. By incorporating the additional weight reflecting the likelihood for different foods to be consumed raw/undercooked, the attributions of the Dutch model were effectively adjusted, both in terms of ranking and percent contributions of the different reservoirs. We concluded that incorporating food consumption data in the Dutch model can significantly affect the results. Therefore, such data should be either excluded from this model or used together with an additional weight able to adjust the amount of food consumed by its likelihood to be consumed insufficiently cooked. This may help identifying the correct reservoirs, allowing attributions to more closely reflect the real chance for a given food to serve as a vehicle for salmonellas. Conversely, the Hald model works properly irrespective of inclusion of food consumption data.

© 2014 Elsevier B.V. All rights reserved.

## 1. Introduction

Over 100,000 human cases of salmonellosis are notified each year in the 27 European Union (EU) Member States (~502 million population) (EFSA, 2011), yet it is widely accepted that notified cases are subject to considerable underreporting. By using disease risks of returning Swedish travelers for the period 2005–2009 anchored to a Dutch population-based study on gastroenteritis, Havelaar et al. (2013) have estimated that more than 6.2 million human cases of salmonellosis occur annually in the 27 EU Member States. This poses salmonellosis as the second most important cause of human bacterial gastroenteritis, after campylobacteriosis.

Contaminated food is by far the primary source of human salmonellosis, with 86–95% of cases estimated to be foodborne (Majowicz et al., 2010). To ascertain the main reservoirs of salmonellosis and to assess the impact of food safety interventions, source attribution based on microbial subtyping is being performed in many countries within (David et al., 2013a,b; Hald et al., 2004, 2007; Mughini-Gras et al., 2014a,b; Pires et al., 2011; Pires and Hald, 2010; van Pelt et al., 1999; Wahlström et al., 2011) and outside (Guo et al., 2011; Mullner et al., 2009a) the EU. The microbial subtyping approach, based on the comparison of frequency distributions of pathogen subtypes isolated from humans with those isolated from putative sources of infection (Barco et al., 2013; Pires et al., 2009), has received considerable attention since the development of the Hald model for *Salmonella* source attribution in Denmark (Hald et al., 2004). The Hald model uses a

---

* Corresponding author at: National Institute of Public Health and the Environment (RIVM), Centre for Infectious Disease Control (CIb), Epidemiology and Surveillance Unit, PO Box 1-3720 BA Bilthoven, The Netherlands. Tel.: +31 30 274 2519; fax: +31 30 274 4409.

*E-mail addresses:* lapo.mughini.gras@rivm.nl, lapomugras@libero.it (L. Mughini-Gras).

Bayesian approach and has been developed separately from the frequentist Dutch model (van Pelt et al., 1999) to attribute stochastically human cases to putative sources of infection while accounting for differences in *Salmonella* subtypes and sources to cause human infection (Hald et al., 2004). To further improve identifiability and to handle uncertainty in data of poorer quality, a modified Hald model has also been proposed (Mullner et al., 2009a). More recently, a modified Dutch model has been developed and its attributions do not seem to differ from those of the modified Hald model (Mughini-Gras et al., 2014b).

The amount of food consumed or available for consumption within a country (expressed as either total tons of food consumed or average food intake per capita) is often included in *Salmonella* source attribution models to account for differences in the overall exposure to the different sources. For this purpose a variety of food consumption data have been used, including data from national food (agricultural) supply statistics (Guo et al., 2011; Hald et al., 2004, 2007; Mughini-Gras et al., 2014a,b; Pires et al., 2011; Pires and Hald, 2010) and individual dietary surveys (David et al., 2013a,b). Food supply statistics are usually readily available for many years in most countries. However, even when these data are adjusted for losses and wastage of inedible portions during food processing, distribution, storing and preparation (at either retail or household level), their use requires the assumption that all food available for consumption is effectively consumed (Guo et al., 2011). Conversely, individual dietary surveys are costly and labor-intensive, but they may provide more accurate data on the actual food intake and even on the practices of preparation. Yet, individual dietary surveys are hampered by a plethora of limitations linked to misreporting (mainly recall and prevarication biases), non-coverage and/or non-response of certain groups of the population, as well as seasonality and trends in food consumption patterns (Hallström and Börjesson, 2013). Therefore, to be functional to source attribution, these surveys need to be performed on a regular basis or at least be updated by combining their figures with food supply data. This was done by David et al. (2013b) for attributing salmonellosis cases that occurred in France in 2005 using individual food consumption data obtained from a French survey conducted in 1999, when the BSE crisis in Europe had almost reached its highest levels and had already had serious repercussions in the food market.

An aspect that is often overlooked when using food consumption data in source attribution analyses is that not all foods have the same probability of serving as vehicles for salmonellas, as some foods are clearly more likely to be consumed raw or undercooked than others (Mughini-Gras et al., 2014a,b). Such foods are relatively more likely to cause infection directly, and possibly also indirectly through cross-contamination (Hald et al., 2004). The Hald-type models are supposed to account for this by incorporating the so-called source-dependent parameter ($a_j$, see Section 2.3.2), which is an uninformative multiplication factor fitted by the model to arrive to the most likely solution given the observed data (Hald et al., 2004; Mullner et al., 2009a). It has been suggested that values taken by this parameter are somehow biologically interpretable, with high values for sources such as beef being compatible with its higher likelihood of being consumed raw or just slightly cooked (Hald et al., 2004). However, evidence for this notion is rather inconsistent, as other model runs may depict situations that are not directly interpretable with insufficient cooking of certain foods (Pires and Hald, 2010), as the $a_j$ parameter also accounts for other factors (Section 2.3.2).

Using *Salmonella* data from the Netherlands during 2001–2004, this study aims at elucidating whether and how the inclusion of data on food consumption in the (modified) Dutch and Hald models influences their attributions. Next, we propose the incorporation of an additional parameter in the modified Dutch model that weights the amount of food consumed by its likelihood to be consumed raw or undercooked by the population, thereby allowing food consumption data in this source attribution model to reflect more closely the actual chance for a given food to serve as a vehicle for *Salmonella*.

## 2. Material and methods

### 2.1. Data on human salmonellosis cases

A dataset consisting of 3735 sero/phage typed *Salmonella* isolates from human cases that occurred in the Netherlands between January 2002 and December 2003 was obtained from the EU and national reference laboratory for *Salmonella* at the Dutch National Institute for Public Health and the Environment (RIVM). These salmonellosis cases were identified by the Dutch Regional Public Health Laboratories through passive surveillance of people with gastroenteritis. Serotyping and further phage typing of *S.* Enteritidis and *S.* Typhimurium isolates was performed as described by van Duijkeren et al. (2002).

### 2.2. Data on Salmonella in animal reservoirs

*Salmonella* isolates from four food-producing animal reservoirs, i.e. pigs ($n = 1566$), cattle ($n = 714$), broiler chickens ($n = 2894$) and egg laying hens ($n = 780$) were identified by the Dutch Regional Veterinary Services between 2001 and 2004 as part of diagnostic activities and a diversity of surveillance programs on farms, slaughterhouses and at retail. These isolates were sent to the RIVM for serotyping and further phage typing of *S.* Enteritidis and *S.* Typhimurium isolates as for human isolates.

### 2.3. Source attribution analysis

Of the 3735 human salmonellosis cases, 271 (7.3%) and 350 (9.4%) were discarded from the source attribution analysis because they were travel- and outbreak-related, respectively. Another 170 cases (4.6%) were discarded because their infecting sero/phage types were only found in humans and not in any of the considered animal reservoirs. The original and modified Dutch models, as well as the modified Hald model, were then applied to attribute the remaining 2944 sporadic and domestic cases to pigs, cattle, broilers and layers/eggs. The dataset was structured in such a way that there were 128 different *Salmonella* subtypes found in humans and in at least one of the considered reservoirs. These subtypes included 61 serotypes along with 51 phage types of *S.* Typhimurium and 16 phage types of *S.* Enteritidis.

#### 2.3.1. Original and modified Dutch model

The original Dutch model (oDM) and the modified Dutch model (mDM) have been presented in detail elsewhere (Barco et al., 2013; Mughini-Gras et al., 2014a,b; Mullner et al., 2009b; van Pelt et al., 1999). Briefly, in the oDM, the expected number of human infections caused by subtype $i$ originating from reservoir $j$, denoted as $\lambda_{ij}$, is given by

$$\lambda_{ij} = \frac{r_{ij}}{\sum_j r_{ij}} \times e_i$$

where $r_{ij}$ is the relative frequency of serotype $i$ from reservoir $j$, and $e_i$ is the frequency of human salmonellosis cases of subtype $i$. A sum over subtypes gives the total number of cases attributable to reservoir $j$; 95% confidence intervals (CI) are estimated by bootstrapping.

Compared to the oDM, the mDM incorporates modeled pathogen prevalence and food consumption weights, taking the following form

$$\lambda_{ij} = \frac{p_{ij} \times m_j}{\sum_j p_{ij} \times m_j} \times e_i$$

where $p_{ij}$ is the prevalence of subtype $i$ from reservoir $j$, and $m_j$ is the "standard" food consumption weight (kg/person per year) for reservoir $j$. Pathogen prevalence is modeled as $p_{ij} = \pi_j \times r_{ij}$, where $\pi_j$ is the overall prevalence of all *Salmonella* subtypes in reservoir $j$, and $r_{ij}$ is still the