Trends and sources in human salmonellosis 2020

By Tine Hald (tiha@food.dtu.dk), Patrick Njage, Channie Kahl Petersen and Eva Litrup

In 2020, core genome multilocus sequence typing (cgMLST) profiles were generated for 562 isolates from human cases out of 614 reported infections. The 562 human isolates were attributed to food sources applying a machine learning source attribution model on cgMLST profiles of food isolates from two consecutive years, namely 2019 and 2020. Human *Salmonella* cases from 2020 were predicted by the model and attributed to eight different food and animal sources. The main source was Danish produced pork followed by imported pork and imported duck meat. This chapter describes the attribution to the human cases in more detail followed by a description of the food data used as model input, the method and results.

1 Isolates from human *Salmonella* cases included in the model

Of the 562 human *Salmonella* isolates, 462 cases were sporadic and 100 cases were associated with 10 outbreaks (of which 25 cases were associated with an international outbreak). The source of all ten outbreaks remained unknown. The sporadic cases included 111 travel related cases, 220 domestic cases and 141 with unknown travel history. The source attribution model allocated sporadic cases with no or unknown travel history and the index cases for the 10 outbreaks (361 cases in total). Reported travel-related cases were allocated directly to travel without a modelling step.

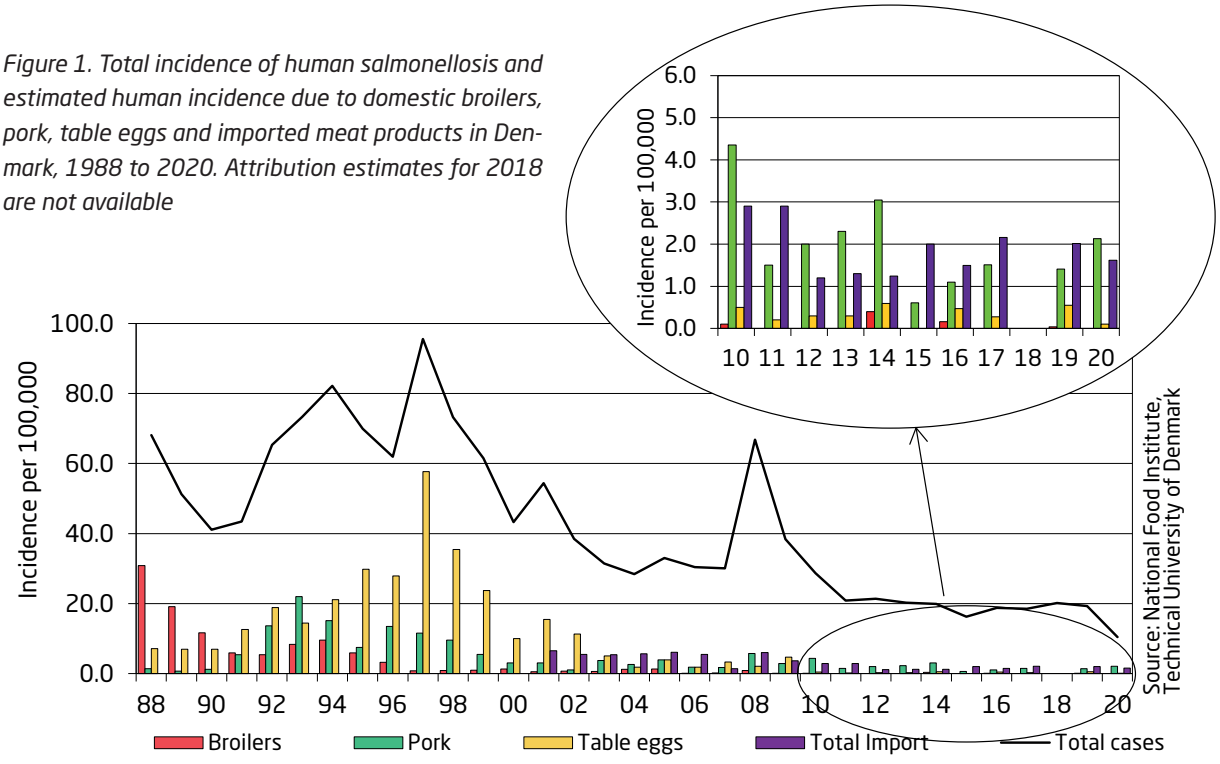
2 Isolates from food and animal included in the model

Salmonella isolated from animal and food were collected as part of the Danish National *Salmonella* surveillance programmes for animals and food and the source attribution model was based on associated cgMLST-profiles. From 2019, 184 isolates were included and 170 were included from 2020. The isolates originated from eight different food and animal sources (Figure 2).

3 Method

In 2017, serotyping and Multiple Locus Variable Tandem Repeat Analysis (MLVA) were replaced by whole genome

Figure 1. Total incidence of human salmonellosis and estimated human incidence due to domestic broilers, pork, table eggs and imported meat products in Denmark, 1988 to 2020. Attribution estimates for 2018 are not available



Source: National Food Institute, Technical University of Denmark

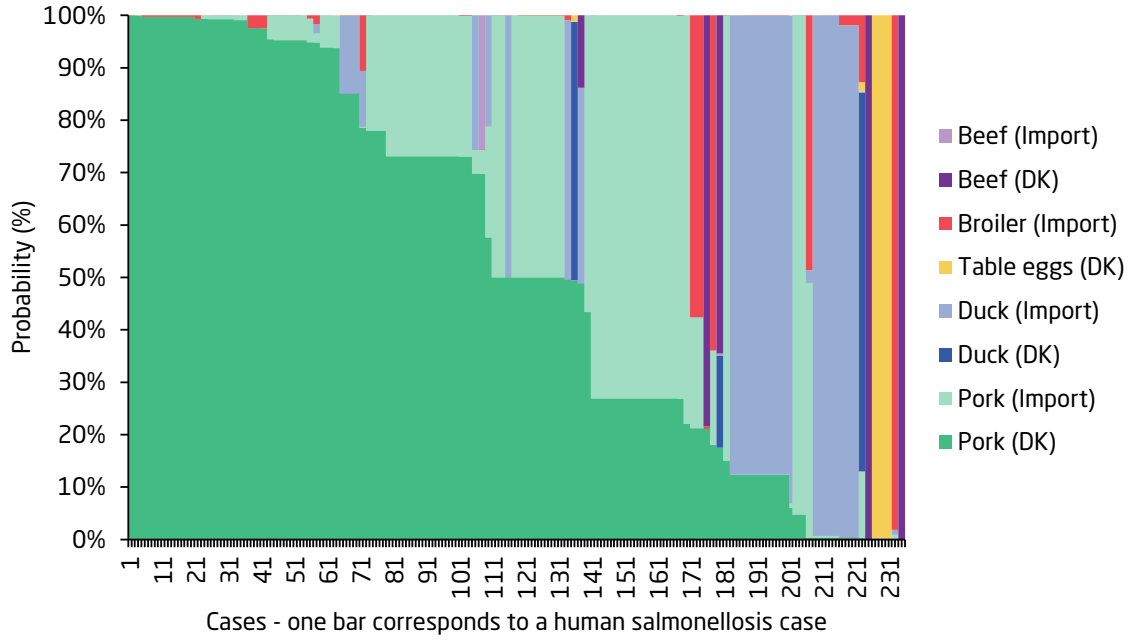
sequencing (WGS) of all isolates found as part of the National *Salmonella* surveillance programmes for animals and food, and the National surveillance of human *Salmonella* infection. Consequently, the Bayesian source attribution model was replaced by a machine learning model, developed for the purpose [1]. Machine learning (ML) is a collective name for mathematical models that learn from data and improves with experience/more data [2]. The models are defined by algorithms capable of recognizing patterns in large and complex datasets making the method applicable for analysing DNA sequence data [2]. The method identifies relevant features in the dataset enabling the ability to make strong allocations.

For the *Salmonella* source attribution 2020, we applied cgMLST, by which all core genes are used in the analysis, and strains are differentiated by their allelic variations. Statens Serum Institut provided the cgMLST profiles for each sequence using the Enterobase scheme [3] in BioNumerics version 7.6 (Applied Maths, Sint-Martens-Latem, Belgium). The core genome of *Salmonella* consist of 3,002 loci with one single locus having several allele variations, thereby providing a high discriminatory power compared to previous methods used. The allelic values from the 3,002 cgMLST loci were used to train the machine learning models using the 354 food and animal isolates.

We applied a supervised classification ML model. The classification is supervised, because the machine is informed about from which of the different animal sources (classes) each of the specific isolates from food and animal originates, and the model then identifies those cgMLST that are able to distinguish between the sources based on their allelic variation. The ML model was constructed from a training dataset consisting of the majority (70%) of the food isolates. The accuracy of the model was then determined from the model's ability to allocate the origin of the remaining part (30%) of the animal and food isolates. As soon as a model with a satisfying accuracy was obtained, a final model using the entire (100%) of the food isolates was constructed. The probability of each human isolate to originate from a specific source was allocated from the final model. The sum of these probabilities within each source equals the total number of human cases attributed per source. Human isolates that could not be allocated to any of the sources, for which data were available, are referred to an unknown source category.

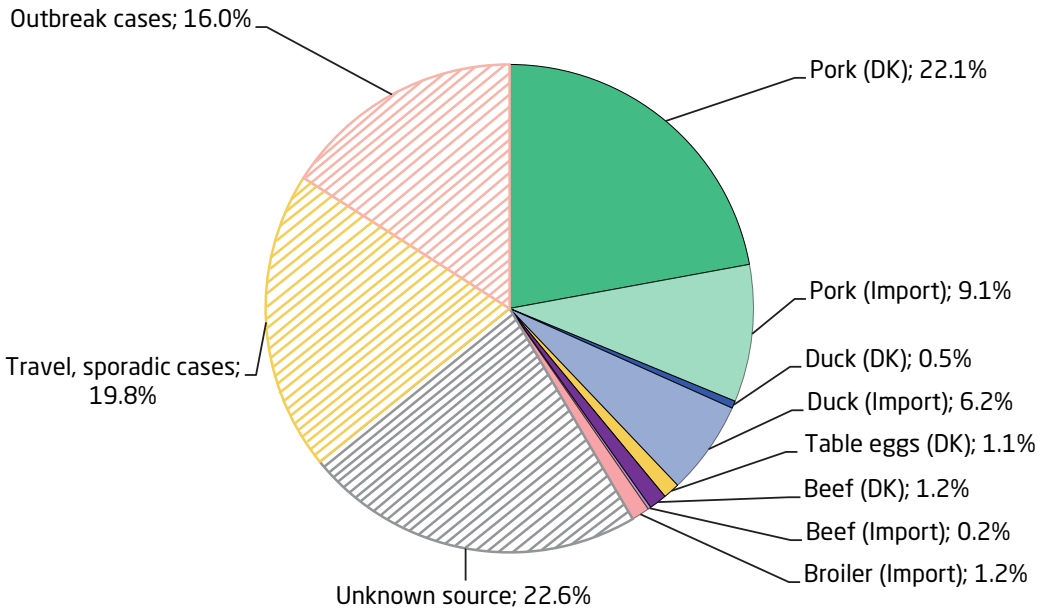
The ML model does not compute uncertainty intervals per se, but takes the uncertainties into account when building the model by repeating the model building 10 times and applying a 7-fold cross validation for each model build. The uncertainty of the results is reflected in

Figure 2. Probability of sources attributed to each sporadic case (incl. outbreak index cases) by ML model. The 234 predicted human cases are lined up along the x-axis and the source specific probabilities for each of the human cases are stacked along the y-axis. Human cases attributed to an unknown source not shown



Source: National Food Institute, Technical University of Denmark

Figure 3. Relative attribution (%) of the 562 human salmonellosis cases in 2020



Note: The grey-striped category consist of cases that could not be allocated a source by the ML model. The yellow- and pink-striped categories were not included in the ML model, but allocated directly from the case information.

Source: National Food Institute, Technical University of Denmark

Figure 2 where the probability of each human case belonging to one of the included sources is illustrated. This source account attributed sporadic human *Salmonella* cases including index cases from 10 foodborne outbreaks to food sources included in the source attribution model.

4 Results

The first model applied did not allocate any human *S.* Enteritidis cases to specific sources indicating that *S.* Enteritidis strains found in imported duck and chicken, and Danish layers were genetically quite different from the strains recovered from humans. For this reason, *S.* Enteritidis cases were removed from the final model, as this increased the accuracy of the model. The sources of the 68 domestic *S.* Enteritidis cases, therefore, remains unknown.

The final model attributed 234 (64.8%) of the 361 sporadic human cases to a food source (Figure 1). Most of the cases had a high probability (>70%) of originating from a single source, whereas other cases had a more or less equal chance of originating from two or three sources (Figure 2).

Similar to previous years, the most important food source was Danish produced pork (124 cases corresponding to 22.1% of the 562 human cases) followed by imported pork (9.1%), imported duck meat (6.2%), imported chicken (1.2%) and Danish produced beef (1.2%) (Figure 3). Few cases were also attributed to Danish layers/eggs (1.1%), Danish produced duck (0.5%) and imported beef (0.2%). In total, 140 (24.9%) of the 562 cases were attributed to Danish produced food, 94 (16.7%) cases to imported food and 127 (22.6%) cases to the unknown source category. In many ways, 2020 was an unusual year also in relation to human salmonellosis. The total number of reported human cases was almost halved from 1,122 cases in 2019 to 614 cases in 2020. Half of this reduction can be explained by less travel-related cases, which fell from 419 in 2019 to 111 in 2020. The number of sporadic domestic cases including those with unknown travel history also decreased with around 100 cases. Because of the reduced number of travel cases, the relative number of cases allocated to particularly Danish pork increased from 8.0% to 22.1%, whereas the absolute increase was less pronounced.

This year, the source attribution model is based on food data from 2019 and 2020, and allocated human cases from 2020 only. Two years of source data were included to enhance the robustness of the predictive results. A model applying more years of source data, would most likely increase the number of cases that can be attributed to a specific source, as done in previous years. On the other hand, if specific *Salmonella* types are present in given sources occasionally and disappear again, including these in following years could potentially be misleading. Including source data from past years may therefore introduce bias in the

model. It was, therefore, assessed to be more appropriate to include only two years of source data in order to strike a balance between robustness and credibility of the results.

The ML model is an additional tool to investigate the relative importance of sources of human salmonellosis. Like all other methodologies it has uncertainties, but the outputs supplement the other methodologies currently used for similar purposes e.g. outbreak investigations to aid decision-making. Furthermore, the model has potential for further expansion and development to account for factors such as prevalence and human consumption patterns.

5 References

1. Munck NSM, Njage PMK, Leekitcharoenphon P, Litrup E & Hald T (2020). Application of Whole Genome Sequences and Machine Learning in Source Attribution of Salmonella Typhimurium. Risk Analysis 40, 1693-1705. doi:10.1111/risa.13510.
2. Libbrecht MW & Noble WS (2015). Machine learning applications in genetics and genomics. Nat. Rev. Genet. 16, 321–332. doi: 10.1038/nrg3920.
3. Alikhan NF, Zhou Z, Sergeant MJ & Achtman MA (2018). A genomic overview of the population structure of Salmonella. PLoS Genetics 14, e1007261.

6 Computation and Software

R version 3.5.1 (2018-07-02)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 18363)
Specific package versions:
ggpubr_0.4.0, rstatix_0.7.0, knitr_1.30, dplyr_1.0.5, caret_6.0-86, lattice_0.20-35, ggplot2_3.3.0, reshape_0.8.8, readxl_1.3.1 and topicmodels_0.2-11
DeiC National Life Science Supercomputer at DTU