Trends and sources in human salmonellosis

By Maja Lykke Brinch (malbri@food.dtu.dk), Marianne Sandberg, Patrick Njage, Abbey Olsen and Eva Litrup

In 2021, core genome Multi-locus sequence typing profile (cgMLST) profiles were generated for 639 isolates from human cases out of 694 reported infections. The 639 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 2020 and 2021. Human *Salmonella* cases from 2021 were predicted by the model and attributed to nine different food and animal sources. The main source was Danish produced pork followed by imported pork and imported duck meat. This chapter describes the human cases in more detail followed by a description of the food data used as model input, the method and results.

Isolates from human *Salmonella* cases included in the model

Of the 639 human *Salmonella* isolates, 452 cases were sporadic and 187 cases were from 12 outbreaks (of which 117 cases were associated with an international outbreak). The source of six outbreaks remained unknown. The sporadic cases included 56 travel related cases, 275 domestic cases and 121 with unknown travel history. The source attribution model was used to allocate sporadic cases with no or unknown travel history and the index cases for the 12 outbreaks (408 cases in total). Travel related cases were allocated to travel without using the model.

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 2020, 170 isolates were included and 244 were from 2021. The isolates originated from 9 different food sources (Figure 1).

Figure 1. Total incidence of human salmonellosis and estimated human incidence due to domestic broilers, pork, layers/table eggs and imported meat products in Denmark, 1988 to 2022. No source account for 2018 or 2020-2021 was calculated

Method

In 2017, serotyping and Multiple Locus Variable Tandem Repeat Analysis (MLVA) were replaced by whole genome 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 2021, 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. A 'feature reduction' step identified which 73 loci (of the 3,002) that provided most information about the source-patterns and these were then used in the model while the remaining loci were excluded. The final model was thus constructed from 413 food isolates and associated allelic values of the 73 loci.

We applied a supervised classification ML model. The classification is supervised, because the machine is 'told' 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 differentiate between the sources based on their alleleic 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 models 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.

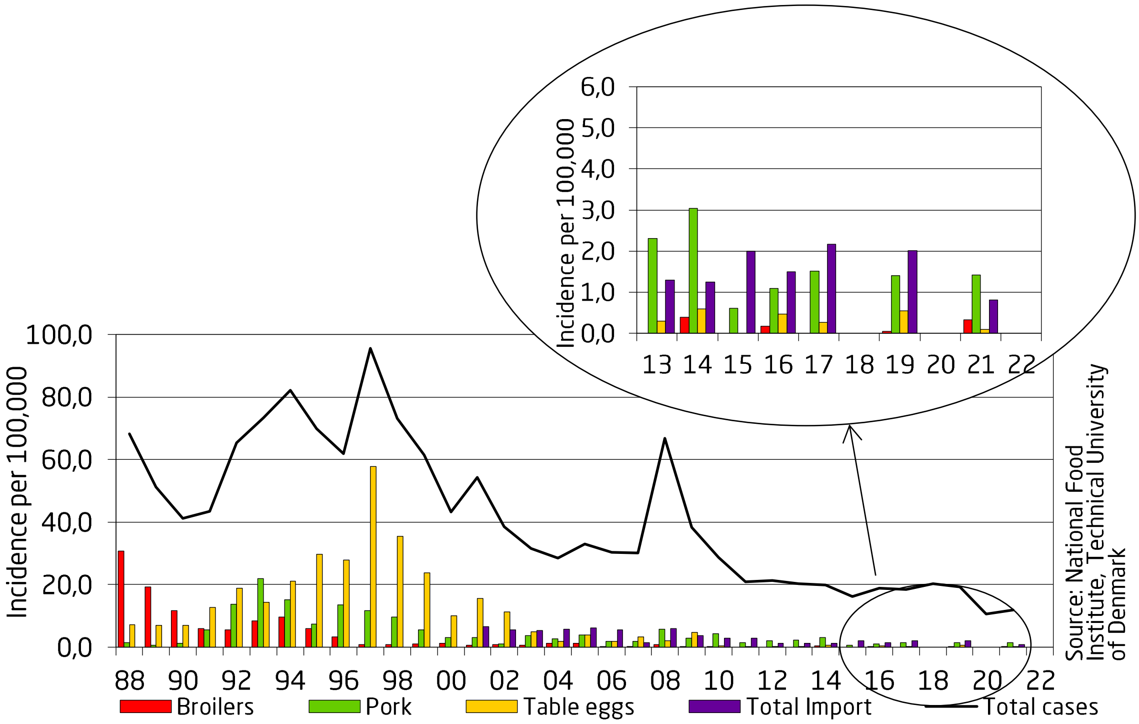
The ML model does not compute uncertainty intervals per se, but takes the uncertainties into account when building the model by repeating the model building 10 times and applying a 7-fold cross validation for each model build. The uncertainty of the results is reflected in Figure 2 where the probability 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.

Results

The model attributed 142 (34.8%) of the 408 human cases to a food source (Figure 1). Most of the cases had a high probability (>86%) of originating from a single source, whereas other cases had a more or less equal chance of originating from two or three sources (Figure 3).

Similar to previous years, the most important food source was Danish produced pork (61 cases corresponding to 9.5% of the 639 human cases) followed by imported pork (3.4%), imported duck meat (3.3%), Danish produced broilers (2.3%) and Danish produced beef (1.7%). Few cases were also attributed to Danish layers/eggs (0.8%), imported broilers (0.6%), Danish produced duck (0.3%) and imported beef (0.2%). In total, 94 (23%) cases was attributed to Danish produced food, 48 (11.8%) cases attributed to imported food and 266 (41.6%) cases attributed to an unknown source.

Figure 1. Total incidence of human salmonellosis and estimated human incidence due to domestic broilers, pork, layers/table eggs and imported meat products in Denmark, 1988 to 2022. No source account for 2018 or 2020-2021 was calculated



Source: National Food Institute, Technical University of Denmark

Figure 2. Relative attribution (%) of the 516 human salmonellosis cases in 2021.

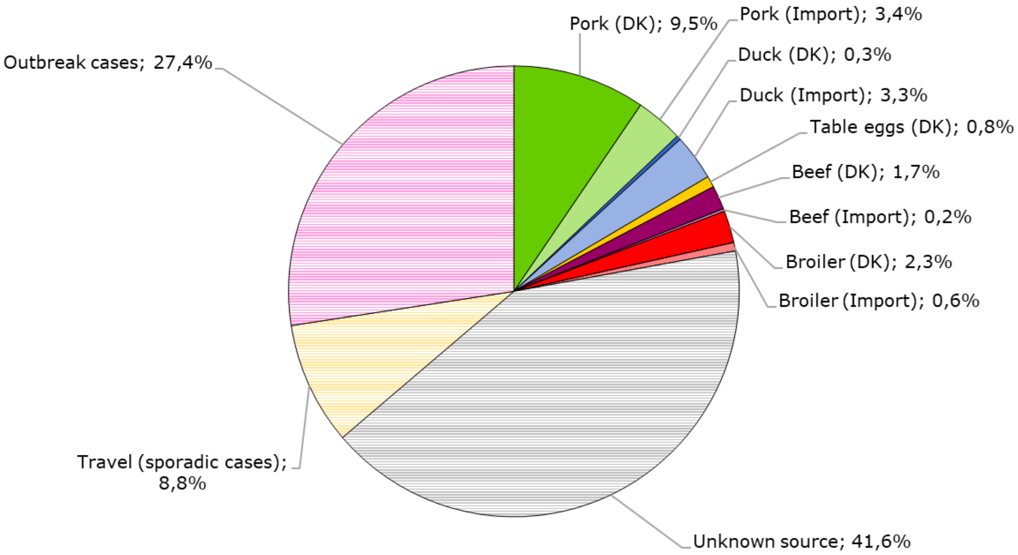
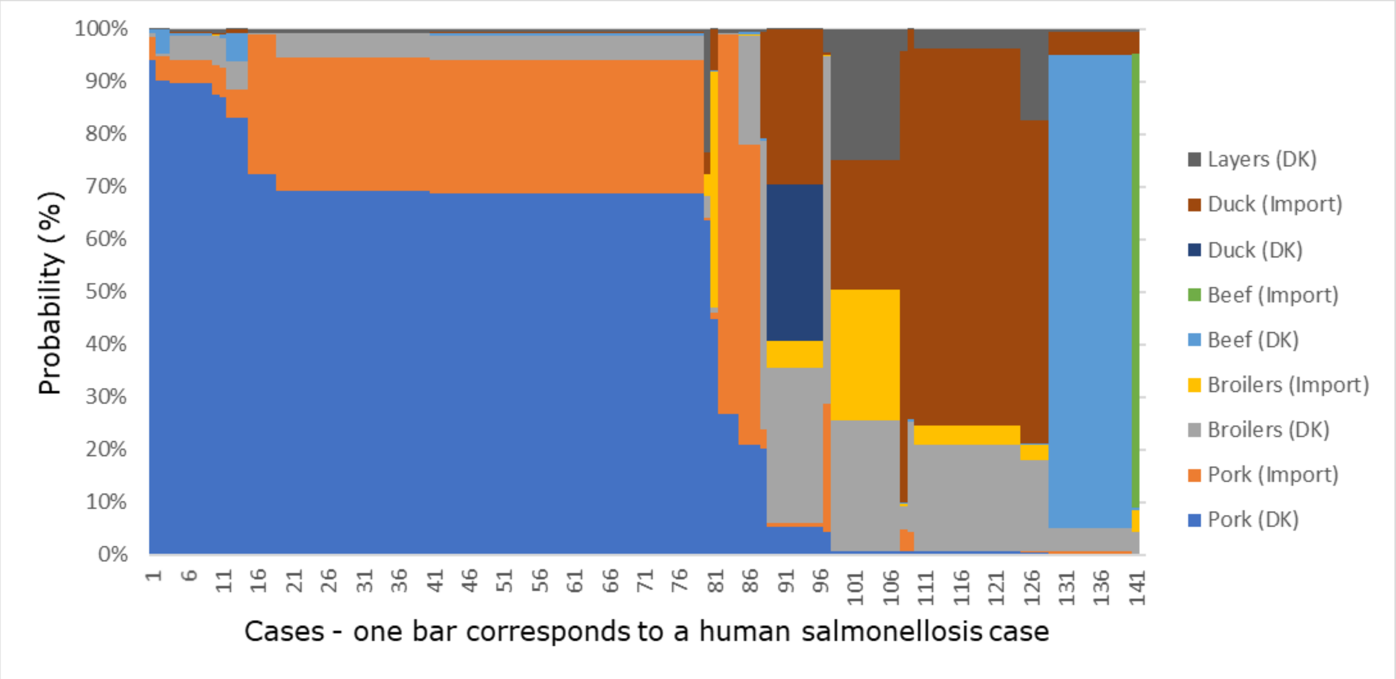


Figure 3. Probability of sources attributed to each sporadic case (incl. outbreak index cases)) by the ML model. The 142 predicted human cases are lined up along the x-axis and the source probabilities for each of the human cases are stacked along the y-axis. Human cases attributed to an unknown source are not shown.



Discussion

The total number of reported human cases was low due to the Covid 19 pandemic with 694 cases in 2021, similar to 2020 where the number was 619. The travel-related cases, in 2021, was also only 66.

This year, the source attribution model is based on food data from 2020 and 2021, and allocated human cases from 2021 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 the 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.

In the future, the usability of new bioinformatical methods will be investigated with the aim of improving the number of predicted cases for the source attribution model for *Salmonella*.

References

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