Journal of Infection 91 (2025) 106632

Contents lists available at ScienceDirect

Journal of Infection

journal homepage: www.elsevier.com/locate/jinf

Bacteria and Bacterial Diseases

Attributable sources of the five most prevalent non-typhoidal *Salmonella* serovars across ten European countries

Gijs Teunis [a], Timothy J. Dallman [b], Magdalena Zając [c], Magdalena Skarżyńska [c], Liljana Petrovska [d], Angela Pista [e], Leonor Silveira [e], Lurdes Clemente [f], Amandine Thépault [g], Laetitia Bonifait [g], Annaëlle Kerouanton [g], Marianne Chemaly [g], Julio Alvarez [h], Robert Söderlund [i], Eva Møller Nielsen [j], Marie Chattaway [k,l], Kaye Burgess [m], William Byrne [n], Aldert L. Zomer [b], Maaike van den Beld [a], Antoni P.A. Hendrickx [a], Eelco Franz [a], Sara Pires [o], Tine Hald [o], Lapo Mughini-Gras [a,b,*]

[a] *Centre of Infectious Disease Control (Cib), National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands*
[b] *Faculty of Veterinary Medicine, Utrecht University, Utrecht, the Netherlands*
[c] *National Veterinary Research Institute, Państwowy Instytut Weterynaryjny (PIWet), Puławy, Poland*
[d] *Animal and Plant Health Agency (APHA), Surrey, United Kingdom*
[e] *Department of Infectious Diseases, National Institute of Health Doutor Ricardo Jorge (INSA), Lisbon, Portugal*
[f] *National Institute of Agrarian and Veterinary Research (INIAV), Oeiras, Portugal*
[g] *Unit of Hygiene and Quality of Poultry and Pork Products, Laboratory of Ploufragan-Plouzané-Niort, French Agency for Food Environmental and Occupational Health and Safety (ANSES), Ploufragan, France*
[h] *VISAVET Health Surveillance Center, Universidad Complutense, Madrid, Spain*
[i] *Swedish Veterinary Agency (SVA), Uppsala, Sweden*
[j] *Statens Serum Institut (SSI), Copenhagen, Denmark*
[k] *Gastrointestinal Bacteria Reference Unit, United Kingdom Health Security Agency (UKHSA), London, England, United Kingdom*
[l] *Genomic and Enabling Data Health Protection Research Unit, University of Warwick, Coventry, United Kingdom*
[m] *Teagasc Food Research Centre, Dublin, Ireland*
[n] *Department of Agriculture, Food and the Marine, Dublin, Ireland*
[o] *National Food Institute, Technical University of Denmark (DTU), Kgs. Lyngby, Denmark*

ARTICLE INFO

Article history:
Accepted 10 October 2025
Available online 13 October 2025

Keywords:
Salmonellosis
Source attribution
Modelling
Whole-genome sequencing

SUMMARY

Non-typhoidal *Salmonella* is the second most frequently reported zoonotic pathogen in the European Union and European Economic Area. Most human infections are caused by serovars Enteritidis and Typhimurium. Genomic characterisation of *Salmonella* isolates from humans and animals has become a routine public health surveillance tool in many countries. In this study, the relative contributions of several potential sources of human infection of the five frequently reported *Salmonella* serovars were estimated using machine-learning methods based on a large, cross-sectional collection of genomes from human cases, and animal and environmental sources, across ten European countries. To define the population structure, core-genome Multilocus Sequence Typing was performed. A supervised machine-learning approach was applied for source attribution in the form of a Random Forest classifier. The source and country attribution models achieved moderate accuracy (F1=0.6–0.9), which is lower than in previous studies using machine-learning on Whole Genome Sequencing data. However, attributions of human clinical isolates to different sources were generally in line with previous findings for these five serovars. While the lack of clonality in some sources hindered their prediction, it is also likely that certain sources (e.g., pets) do not serve as major contributors to human infection. Therefore, in most cases attributing these sources to the livestock species they are typically associated with, is likely appropriate. Country attributions showed that substantial human cases are attributable to countries other than their own, indicating geographical interrelatedness of sources. This highlights the value of internationally harmonised *Salmonella*-control policies in the food production chain.

* Correspondence to: National Institute for Public Health and the Environment (RIVM), Antonie van Leeuwenhoeklaan 9, 3721 MA Bilthoven, the Netherlands.
E-mail addresses: lapo.mughini.gras@rivm.nl, l.mughinigras@uu.nl (L. Mughini-Gras).

https://doi.org/10.1016/j.jinf.2025.106632

Introduction

Non-typhoidal *Salmonella* is the second most frequently reported zoonotic pathogen in the European Union and European Economic Area (EU/EEA).[1] While there are over 2500 serovars of *Salmonella enterica* subspecies *enterica*, most human infections are caused by serovars Enteritidis and Typhimurium.[2] Despite significant reductions in human cases in Europe since the 1990s due to the adoption of regulatory and prophylactic measures in agricultural settings, this downward trend in reported human salmonellosis has shown clear signs of stagnation since 2016.[1] *Salmonella* can be isolated from a wide range of animals and the environments they inhabit, although some serovars display a more restricted host range than others.[3] Humans are most often exposed to *Salmonella* through the consumption of contaminated food, particularly food of animal origin. However, *Salmonella* can also be transmitted through direct contact with animals, the environment, and, to a limited extent, person-to-person transmission. The contributions of different (animal) reservoirs and associated transmission routes are expected to vary across serovars and exposed populations.[4] To inform risk managers and to target effective interventions, quantitative evidence on the main sources of the pathogen is essential.

Genomic characterisation of *Salmonella* isolates from humans and animals has become a routine public health surveillance tool in many countries. Standardised methodologies, such as core genome multi-locus sequence typing (cgMLST), facilitate stable comparisons of the population structure of isolates and have recently been shown to have potential in source attribution studies due to their high discriminatory power.[5] To exploit the high dimensionality offered by whole-genome sequencing (WGS) data for source attribution, machine learning (ML) approaches have been demonstrated to be an appropriate methodology, and supervised ML in particular has been applied to genomic data from Danish *Salmonella* surveillance,[6] as well as other foodborne pathogens.[7–9] Furthermore, previous research has shown promising results when using ML on cgMLST data for source attribution.[10–12]

In this study, the relative contributions of several potential sources of human infection of five frequently reported *Salmonella* serovars, were estimated using ML methods, namely: Typhimurium, Enteritidis, Infantis, Newport, and Derby. We utilized a large, cross-sectional collection of genomes from human cases, as well as animal and environmental sources, across ten European countries. Specifically, we first studied the population structure of the different serovars through a constructed phylogeny and subsequently applied a supervised ML approach to build a classification model based on source type and country of origin. This model was then used to infer the attributable sources of human isolates to provide the most extensive estimates of the relative contributions to human salmonellosis across sources and countries within Europe. In doing so, we aim to demonstrate the potential of ML methods for source attribution and to offer additional insights into the origins of clinical cases of *Salmonella* in Europe.

Materials and methods

Data

A total of 3548 isolates were collected and whole genome sequenced through routine surveillance activities carried out in 12 public health and veterinary institutes: French Agency for Food Environmental and Occupational Health and Safety (ANSES), Animal and Plant Health Agency (APHA), Technical University of Denmark (DTU), Instituto Nacional de Investigação Agrária e Veterinária (INIAV), National Institute of Agrarian and Veterinary Research (INSA), United Kingdom Health Security Agency UKHSA) (Formally known as Public Health England (PHE)), Państwowy Instytut Weterynaryjny (PIWet), National Institute for Public Health and the Environment (RIVM), Statens Serum Institut (SSI), Swedish Veterinary Agency (SVA), Teagasc, VISAVET Health Surveillance Center (VISAVET), across ten European countries (Denmark, England and Wales, France, Ireland, Portugal, Poland, Spain, Sweden, the Netherlands), as part of the research project “DISCOVER” of the One Health European Joint Programme.[13] Genomes originated from non-travel and non-outbreak related clinical human cases and from non-human origin (animal, food, and environmental sources; Table S1). These isolates were collected between 2003 and 2021 representing five serovars: Enteritidis, Typhimurium and its monophasic variant, Derby, Infantis and Newport. The institutes only submitted sequences for serovars found in their surveillance activities (Table 1).

Phylogenetic analysis

To define the population structure, cgMLST was performed using chewieSnake[14] utilising the cgMLST V2 scheme acquired from Enterobase.[15] Allele profiles created for all sequences were visualised in a minimum spanning tree through GrapeTree.[16] Genomes with > 5% cgMLST loci missing were removed from further analysis. Genomes were typed further using the Achtman 7 Gene MLST scheme and the respective eBURST groups (eBG, Appendix 1).[15,17]

Source classification model

A supervised ML approach was applied for source attribution in the form of a Random Forest (RF) classifier using Python’s library Scikit-learn.[18] A conceptual model for the development of ML approaches for source attribution is shown in Fig. S0 (Appendix 1).

Feature encoding and reduction

The cgMLST allele profiles determined by chewieSnake were used as input features. Loci with > 15% missing data were excluded while

Table 1
Metadata of submitted *Salmonella* isolates for each country, by serovar and institute, before quality control.

Country and institute		Derby	Enteritidis	Infantis	Newport	Typhimurium	Total
Denmark	DTU	0	0	0	0	210	210
	SSI	0	0	0	0	141	141
England and Wales	APHA	71	61	23	119	314	588
	UKHSA	118	99	100	100	200	617
France	Anses	0	75	0	0	97	172
Ireland	Teagasc	0	0	0	0	136	136
Netherlands	RIVM	10	145	40	10	12	217
Poland	PIWet	95	384	121	139	258	997
Portugal	INIAV	0	32	21	0	0	53
	INSA	9	29	14	27	58	137
Spain	VISAVET	14	177	13	0	47	251
Sweden	SVA	2	0	0	0	27	29
Total		319	1002	332	395	1500	3548

Table 2
Source and country classification accuracies of each *Salmonella* serovar.

eBG1	Typhimurium & Monophasic Typhimurium				Enteritidis				Infantis				Newport				Derby			
	Precision	Recall	F1	N	Precision	Recall	F1	N	Precision	Recall	F1	N	Precision	Recall	F1	N	Precision	Recall	F1	N
Broilers	0.557	0.506	0.531	77	0.554	0.577	0.565	248	0.743	0.872	0.802	86	0.855	0.815	0.835	65	0.667	0.133	0.222	15
Cats	0.120	0.097	0.107	31	0.000	0.000	0.000	6	1.000	1.000	1.000	4	0.000	0.000	0.000	2	0.000	0.000	0.000	1
Cattle	0.370	0.298	0.330	57	0.000	0.000	0.000	6	0.500	0.333	0.400	3	0.545	0.400	0.462	15	0.222	0.667	0.333	3
Dogs	0.472	0.397	0.431	63	0.000	0.000	0.000	9	0.875	0.538	0.667	13	0.692	0.474	0.563	19	0.000	0.000	0.000	6
Ducks	0.636	0.600	0.618	35	0.844	0.794	0.818	34	0.833	0.769	0.800	13	0.000	0.000	0.000	7	0.000	0.000	0.000	1
Environment	0.000	0.000	0.000	4	0.000	0.000	0.000	9	0.000	0.000	0.000	0	0.000	0.000	0.000	1	0.118	0.333	0.174	6
Horses and donkeys	0.422	0.388	0.404	49	0.500	0.385	0.435	13	0.000	0.000	0.000	2	0.471	0.762	0.582	21	0.000	0.000	0.000	0
Layers	0.521	0.543	0.531	70	0.709	0.721	0.715	369	0.500	0.444	0.471	9	0.462	0.400	0.429	15	0.077	0.143	0.100	7
Other pets	0.960	0.857	0.906	28	0.000	0.000	0.000	3	0.000	0.000	0.000	0	1.000	1.000	1.000	4	0.000	0.000	0.000	0
Other poultry	0.538	0.512	0.525	41	0.214	0.188	0.200	16	0.600	0.600	0.600	5	0.667	0.833	0.741	12	0.000	0.000	0.000	1
Pigs	0.813	0.913	0.860	494	0.000	0.000	0.000	2	0.154	0.167	0.160	12	0.800	0.800	0.800	20	0.904	0.717	0.800	92
Reptiles	0.750	0.857	0.800	7	0.571	0.600	0.585	20	0.938	0.882	0.909	17	0.950	0.905	0.927	21	0.000	0.000	0.000	0
Sheep/Goat	0.231	0.158	0.187	19	0.000	0.000	0.000	1	0.000	0.000	0.000	1	0.000	0.000	0.000	1	0.000	0.000	0.000	0
Turkey	0.538	0.452	0.491	62	0.357	0.500	0.417	10	0.400	0.222	0.286	9	0.778	0.778	0.778	36	0.690	0.667	0.678	30
Wild Birds	0.545	0.667	0.600	9	0.444	0.222	0.296	18	0.000	0.000	0.000	1	0.000	0.000	0.000	0	0.000	0.000	0.000	1
Wild Mammals	0.500	0.286	0.364	7	0.667	0.333	0.444	12	0.667	0.667	0.667	3	0.000	0.000	0.000	0	0.000	0.000	0.000	0
Micro Average_Source	0.668	0.668	0.668	1053	0.604	0.604	0.604	776	0.702	0.702	0.702	178	0.699	0.699	0.699	239	0.571	0.571	0.571	163
Macro Average Source	0.469	0.443	0.452	1053	0.286	0.254	0.263	776	0.424	0.382	0.398	178	0.425	0.422	0.418	239	0.157	0.156	0.136	163
Weighted Average Source	0.644	0.668	0.653	1053	0.604	0.604	0.602	776	0.688	0.702	0.688	178	0.710	0.699	0.698	239	0.710	0.571	0.614	163
Denmark	0.703	0.798	0.748	208	0.000	0.000	0.000	0	0.000	0.000	0.000	0	0.000	0.000	0.000	0	0.000	0.000	0.000	0
England and Wales	0.842	0.836	0.839	293	0.741	0.690	0.714	58	0.313	0.250	0.278	20	0.938	0.938	0.938	112	0.953	0.924	0.938	66
France	0.876	0.914	0.895	93	0.789	0.833	0.811	72	0.000	0.000	0.000	0	0.000	0.000	0.000	0	0.000	0.000	0.000	0
Ireland	0.963	0.963	0.963	135	0.000	0.000	0.000	0	0.000	0.000	0.000	0	0.000	0.000	0.000	0	0.000	0.000	0.000	0
Netherlands	0.500	0.333	0.400	6	0.882	0.536	0.667	56	0.524	0.917	0.667	12	0.000	0.000	0.000	0	0.000	0.000	0.000	0
Poland	0.797	0.739	0.767	249	0.886	0.955	0.919	382	0.897	0.905	0.901	116	0.945	0.945	0.945	127	0.916	0.894	0.905	85
Portugal	1000	0.889	0.941	9	0.438	0.656	0.525	32	0.900	0.857	0.878	21	0.000	0.000	0.000	0	0.000	0.000	0.000	0
Spain	0.647	0.579	0.611	38	0.914	0.790	0.848	176	0.500	0.222	0.308	9	0.000	0.000	0.000	0	0.429	0.600	0.500	10
Sweden	0.824	0.636	0.718	22	0.000	0.000	0.000	0	0.000	0.000	0.000	0	0.000	0.000	0.000	0	1.000	1.000	1.000	2
Micro Average Country	0.813	0.813	0.813	1053	0.844	0.844	0.844	776	0.792	0.792	0.792	178	0.941	0.941	0.941	239	0.890	0.890	0.890	163
Macro Average Country	0.715	0.669	0.688	1053	0.465	0.446	0.448	776	0.313	0.315	0.303	178	0.188	0.188	0.188	239	0.330	0.342	0.334	163
Weighted Average Country	0.814	0.813	0.812	1053	0.854	0.844	0.843	776	0.787	0.792	0.783	178	0.941	0.941	0.941	239	0.902	0.890	0.895	163

Micro average is the overall global average without considering sample imbalance. Macro average considers all classes equally in the mean calculation. Weighted average computes the metric for each class, then take the average weighted by class support (number of samples).

Fig. 1. Results of source attribution for *Salmonella* Typhimurium. Top-left bar chart shows the proportion of sequences attributed to each source. Below it is the proportion of sequences originating from each country. Top-right shows the confidence level of each prediction and the proportion of RF trees voting for each source. Bottom-left shows a minimum-spanning tree of the Typhimurium isolates and their associated source. Bottom-right then shows the same tree with the source predicted by the RF.

remaining missing values were imputed using *k*-Nearest Neighbours (kNN). This imputed missing data by majority vote of its k^5 closest points, and has been shown to perform well in phylogenetic data.[10] The cgMLST feature data is categorical in nature with each value representing a different allelic sequence. Thus, OneHotEncoder (OHE) was applied, where a single variable with *n* observations and *y* distinct values is transformed to *y* binary variables with *n* observations each[19] converting it a presence/absence table for each allelic variation. However, this created a multitude of features (> 20,000) as there are 3002 loci in the reference genome. There are various reasons for not including too large numbers of features into ML models, such as decreased accuracy, overfitting or other technical limitations.[20] Therefore, a feature reduction method was applied in the form of a Python implementation of Multivariate methods with Unbiased Variable selection in R (MUVR).[21] To prevent data-leakage, clonal isolates originating from the same country, year and source (< 5 cgMLST allelic distance) were deduplicated.

Sample imbalance

The composition for all classes was not even. To minimise this influence on training the algorithm, the training data set was balanced using RandomOverSampler from Imbalanced-learn.[22]

Model construction and validation

For model construction, the non-clinical data set for each serovar was randomly split into 10 folds of equal size. One fold was put aside for testing and the model was trained on the remaining folds, in order to perform cross-validation. This was done for 10 iterations, such that each fold is used for testing once, and the average performance over the 10 iterations was calculated. In each iteration, for the same testing set, two ML classifiers were trained to predict source and country separately. Both were then validated with the same testing sequences. The indicators used to evaluate model performance were valid accuracy (F1-score), precision and recall, calculated from a confusion matrix. F1-score is the ability of the model to correctly classify the country and source of the testing sequences.[23]

Source attribution

A final model was built using a balanced set of all non-clinical data, using the best performing features. This model was applied to the clinical human case data to predict their sources and countries of origin. For each source, the sum of probabilities is equal to the number of human cases attributed to that source.

Fig. 2. Results of source attribution for *Salmonella* Enteritidis. Top-left bar chart shows the proportion of sequences attributed to each source. Below it is the proportion of sequences originating from each country. Top-right shows the confidence level of each prediction and the proportion of RF trees voting for each source. Bottom-left shows a minimum-spanning tree of the Enteritidis isolates and their associated source. Bottom-right then shows the same tree with the source predicted by the RF.

Results

The non-human isolates were categorised into 17 sources (Table S1); 33 sequences were removed from subsequent analyses due to ambiguous categorisation (e.g., unknown livestock). Additionally, all sequences of small clonal complexes (n < =35) were removed as they likely reflect convergent evolution of the same serotype in distinct genetic backgrounds from different ancestors.[17] The number of sequences included post quality control of each source and country is shown in Tables S2–S6. In total, 1418 Typhimurium, 984 Enteritidis, 310 Infantis, 374 Newport, and 295 Derby sequences were included. This includes both clinical and non-clinical isolates.

Model training, validation and self-attribution

The F1 score for each of the source classifiers ranged from 0.6 (Enteritidis) to 0.7 (Newport). For Typhimurium (F1=0.65), two classes achieved an F1 score > 0.7. The best performing classes were the majority class pigs (F1=0.86) and 'other pets' (F1=0.91) in contrast to the worst performing classes environment (F1=0.00) and cats (F1=0.11). For Enteritidis (F1=0.60), two classes achieved an F1 score > 0.7 (Ducks F1=0.82, layers F1=0.71) in contrast to the poor performance of many of the minority classes (cats, cattle, dogs, environment, other pets, pigs, sheep/goats, F1=0.00). For Infantis (F1=0.69), three classes achieved an F1 score > 0.7, broilers (F1=0.80), ducks (F1=0.80) and reptiles (F1=0.91), also in contrast to poor performance in many of the minority classes. For Newport (F1=0.70), the classes broilers (F1=0.84), other poultry (F1=0.74), reptiles (F1=0.93) and turkey (F1=0.78) performed best with limited self-attribution accuracy for the low frequency classes and the pet sources. Finally, for Derby (F1=0.61), only pigs (F1=0.80) and turkeys (F1=0.68) had an adequate self-attribution accuracy.

The F1 score for the models predicting country of origin all outperformed their respective source model (Typhimurium (F1=0.81), Enteritidis (F1=0.84), Infantis (F1=0.78), Newport (F1=0.94) and Derby (F1=0.90)). All country classes had a F1 score > 0.7 apart from Spain (F1=0.61) and the Netherlands (F1=0.40) in the Typhimurium model, Portugal (F1=0.30) and the Netherlands (F1=0.67) in the Enteritidis model, Spain (F1=0.31) and England and Wales (F1=0.28) in the Infantis model. The final number of input features selected for training the model are displayed in Table S7. All models and classes are presented in Table 2 with a breakdown of precision, recall and accuracy.

Source attribution

Attribution estimates to sources are displayed in Figs. 1–5. Attribution estimates to countries are displayed in Figs. S2–S6.

Typhimurium

The majority of human clinical isolates (N=365) were attributed to pigs (63.8%), followed by dogs (7.1%), horses/donkeys (6.0%),

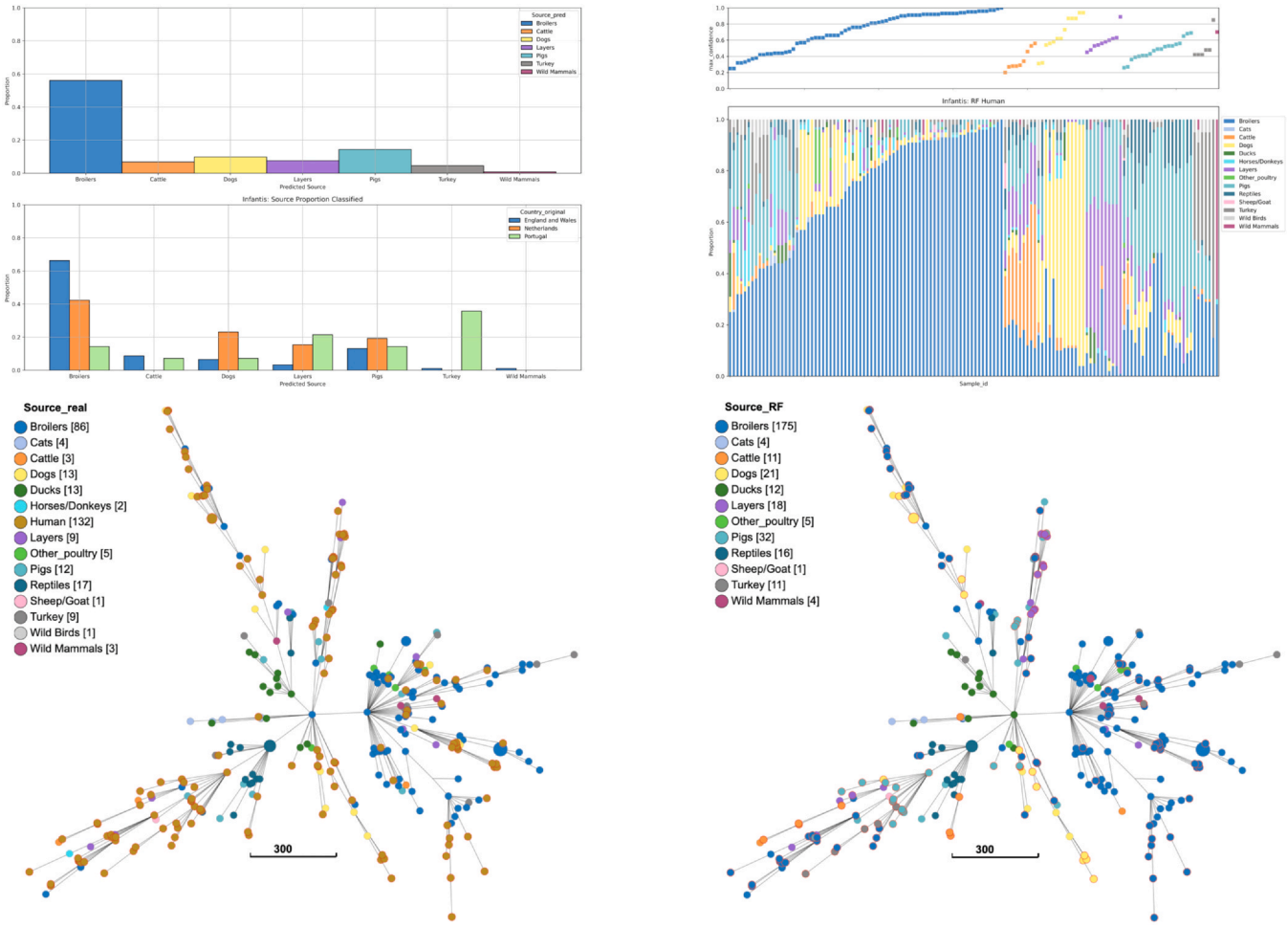

Fig. 3. Results of source attribution for *Salmonella* Infantis. Top-left bar chart shows the proportion of sequences attributed to each source. Below it is the proportion of sequences originating from each country. Top-right shows the confidence level of each prediction and the proportion of RF trees voting for each source. Bottom-left shows a minimum-spanning tree of the Infantis isolates and their associated source. Bottom-right then shows the same tree with the source predicted by the RF.

broilers (4.9%), cats (4.9%), layers (4.4%), cattle (3.3%), other poultry (2.2%), sheep/goat (1.6%), and other pets and reptiles (0.3%) (Fig. 1). For all countries that submitted clinical sequences, pigs were predicted as the most frequent source of human infection (Denmark - 81%, England and Wales - 51.9%, Portugal - 64.1%) apart from the Netherlands where 66.7% (4/6) of isolates were attributed to broilers.

The proportion of human clinical isolates attributed to sources stratified by country was; Denmark (30.1%), England and Wales (29.9%), followed by Poland (18.4%), Spain (9.0%), France (8.8%), Sweden (2.2%), Portugal (0.8%), Ireland (0.5%) and the Netherlands (0.3%). For those countries with both human and non-human sources, 51.1% of Danish and 42.6% of English isolates were estimated to originate from sources consistent with their country of origin. In contrast, 7.7% of Portuguese human infections were predicted to originate within Portuguese non-human sources.

Enteritidis

Of 208 human clinical isolates, 68.8% were attributed to poultry layers, followed by poultry broilers (19.7%), dogs (2.9%), cats (2.4%), other poultry (1.9%), reptiles (1.4%), turkeys (1.0%), horses/donkeys (1.0%), cattle (0.5%), and ducks (0.5%) (Fig 2). The majority of human clinical isolates were attributed to sources in Poland (46.6%) and Spain (25.0%), who submitted 49% and 23% of non-clinical isolates, followed by the Netherlands (10.1%), France, (8.2%), England and Wales (6.3%), and Portugal (3.8%). For Portuguese isolates, 59.3% were attributed to sources in Spain, followed by Portugal itself (18.5%). Attributions for England and Wales and the Netherlands and were similar to each other, with more isolates attributed to sources in Poland, 52.0% and 48.2% respectively. No Enteritidis isolates were submitted by Denmark, Ireland, and Sweden.

Infantis

The majority of human Infantis sequences were attributed to poultry broilers (56.1%). This was followed by pigs (14.4%), dogs (9.8%), layers (7.6%), cattle (6.8%), turkeys (4.5%) and wild mammals (0.8%). England and Wales had the largest (66.3%) attribution to broilers (Fig. 3). Of Dutch human infections, 42.3% were attributed to broilers, and 23.1% to dogs, whilst for Portugal most human isolates were attributed to turkeys (35.7%) followed by 21.4% to layers and 14.3% assigned to broilers. Most human clinical isolates were attributed to sources in Poland (37.9%) and the Netherlands (31.1%), 64.3% of Portuguese isolates to Poland, and none of them to Portugal itself. Half of Dutch isolates were attributed to the Netherlands, and no Dutch isolate was attributed to Spain. No Infantis sequences were submitted by Denmark, France, Ireland, and Sweden.

Newport

Most human Newport sequences were attributed to reptiles (28.1%), followed by cattle (25.9%), layers (17.8%), horses/donkeys (14.1%), turkeys (7.4%), dogs (3.7%) with cats, ducks, the environment, and other poultry at 0.7% (Fig. 4). Notably, 92.6% of Portuguese isolates were attributed to cattle, accounting for 25 of 35 cattle sequences, with the remainder attributed to reptiles (7.4%). For England and Wales, 34.3% of sequences were attributed to reptiles, and

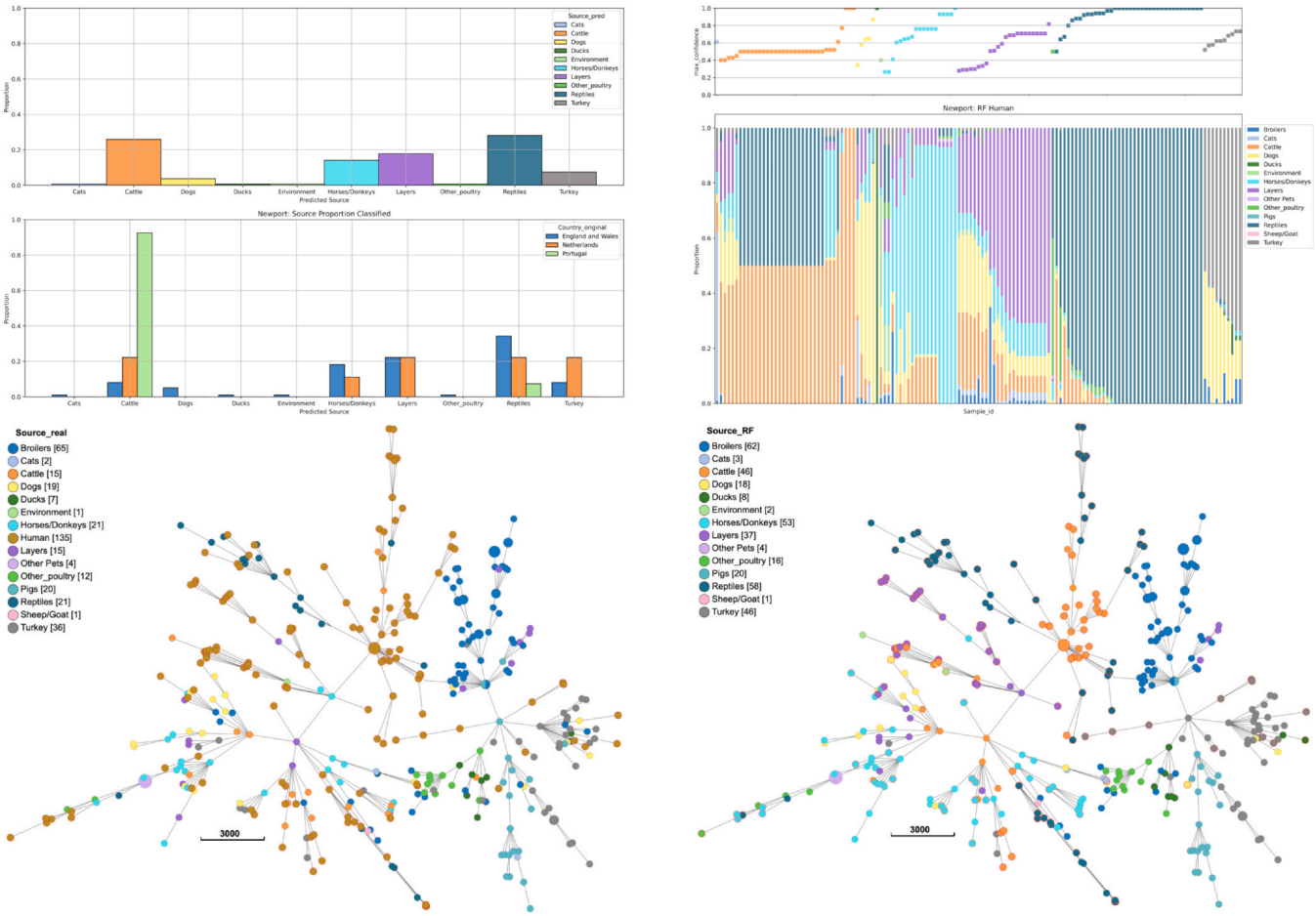

Fig. 4. Results of source attribution for *Salmonella* Newport. Top-left bar chart shows the proportion of sequences attributed to each source. Below it is the proportion of sequences originating from each country. Top-right shows the confidence level of each prediction and the proportion of RF trees voting for each source. Bottom-left shows a minimum-spanning tree of the Newport isolates and their associated source. Bottom-right then shows the same tree with the source predicted by the RF.

8.1% to cattle. The Dutch attributions were equally split between cattle, layers, reptiles and turkeys at 22.2%, with the final 11.1% attributed to horses/donkeys, though the small number of clinical Dutch sequences (n=9) had little impact on the overall distribution.

The non-clinical Newport sequences were from two countries: Poland, and England and Wales. The clinical attributions were split 71.9% to England and Wales and 28.1% to Poland. When breaking it down by country, all Portuguese sequences were attributed to Poland, as well as the majority of Dutch sequences (55.6%). The majority of English samples were attributed to England itself (66.7%). No Newport sequences were submitted by Denmark, France, Ireland, Spain and Sweden.

Derby

The majority (62.9%) of human Derby sequences were attributed to pigs. This was followed by turkeys (22.7%), cattle (6.1%), the environment (3.0%), layers (2.23%), and finally cats, dogs, other poultry, and wild birds (0.8%) (Fig. 5). All sequences attributed to cattle, the environment, layers, and turkeys were from England and Wales, all cats and wild birds were from the Netherlands, and finally, all sequences attributed to dogs and other poultry were Portuguese. Of the human sequences, 43.9% were attributed to sources in Poland, 37.9% to England and Wales, and 18.2% was attributed to Spain. For English sequences, 43.1% was attributed to Poland, 41.1% to England and Wales and the remaining 14.9% to Spain. Dutch isolates had 44.4% attributed to both Poland and Spain, and 11.1% to England and Wales. Finally, Portuguese sequences had 55.6% attributed to Poland, 33.3% to Spain and 11.1% to England and Wales. No Derby sequences were submitted by Denmark, France, and Ireland. In addition, no non-clinical isolates were submitted by the Netherlands and Portugal.

Discussion

Attributions to sources

Attributions of human clinical isolates to different sources were generally in line with previous findings for these serovars.[1] The majority of human infections of *Salmonella* Typhimurium were attributed to pigs,[1,11] the majority of human infections of *Salmonella* Enteritidis were attributed to layers,[1] *Salmonella* Infantis infections were mainly attributed to broilers,[1] and *Salmonella* Derby infections attributed to pigs and poultry.[1] One exception was *Salmonella* Newport, where human cases are frequently associated with poultry sources,[1] but which were equally attributed to reptile sources in this European dataset. While reported reptile-associated salmonellosis has been on the rise, they are mainly associated with other serovars and make up only a small fraction of cases.[24] Instead, this may be the result of the reptile isolates being a more homogenous population. Indeed, the RF appears to struggle distinguishing them for cattle, which is made up of more heterogeneous isolates.

The random forest (RF) classifiers were trained on variations encoded in the cgMLST scheme. These genomic variations in the core genome can be considered reflective of the population structure, since genomes closely related in terms of evolutionary distance are likely to be from the same source. Therefore, the visualisation of the

Fig. 5. Results of source attribution for *Salmonella* Derby. Top-left bar chart shows the proportion of sequences attributed to each. Below it is the proportion of sequences originating from each country. Top-right shows the confidence level of each prediction and the proportion of RF trees voting for each source. Bottom-left shows a minimum-spanning tree of the Derby isolates and their associated source. Bottom-right then shows the same tree with the source predicted by the RF.

population structure of the isolates can provide insights into the variation observed in self-attribution accuracy. For example, for *Salmonella* Enteritidis, over 60% of human infections were attributed to layers compared to approximately 20% to broilers. However, when considering the minimum-spanning tree, sequences from broilers and layers showed considerable mixing. This is reflected in the confidence chart for their attributions, which showed that for isolates attributed to broilers, a large proportion of the trees in the RF voted for layers and *vice versa*. While broilers and layers are functionally separated sectors raising the same species (*Gallus gallus*), our results suggest that their *Salmonella* strains do overlap. In contrast, for *Salmonella* Infantis the RF showed greater confidence differentiating these groups. However, only 9 layers isolates were included, and the RF had lower accuracy in attributing layers.

Country attributions

Understanding the proportion of human cases that are exposed to domestic sources of *Salmonella* as opposed to imported sources is important for control and mitigation prioritisation and strategies. This dataset included *Salmonella* isolates from human and non-human sources within the same country and thus allowed us to examine how this changed per serovar and per country.

Overall, country attributions showed that sequences were frequently attributed to their own countries (33.1%). For example, most Typhimurium sequences were attributed to England and Wales, where most of the isolates were also sourced from. Additionally, Poland, the source of a prolonged *S.* Enteritidis outbreak in the EU/EEA in recent years,[25] had a relatively high proportion of sequences attributed to from other countries (32.2%), specifically of Typhimurium in England and Wales, Portugal, and Denmark (all around 20% of sequences attributed to Poland). A significant percentage of non-clinical sequences were provided by Poland (39.5%). In addition, Poland and England and Wales are the only countries to submit non-clinical sequences for each serovar. Thus, an overrepresentation in the attributions is not unexpected. On the other hand, attributions are not proportional to the number of sequences for all other countries and/or sources for each serovar. Additionally, the model performs well for country attributions (Table 2).

Model accuracy

The source attribution models achieved moderate accuracy (F1=0.6–0.7), but lower than in previous studies using ML on WGS data[10] including a *Salmonella* Typhimurium data set.[11] One explanation is that these studies used isolates from fewer sources (n=5) and fewer countries (domestic and imported). In general, the most numerous livestock sources were predicted with high accuracy and low abundant sources often representing wild animals and pets were predicted with low accuracy. Even with over-sampling strategies low classes with low abundance can be problematic to classify correctly if they are not a single homogenous population. This

highlights the importance of balanced datasets in order to avoid potential bias, as other work on *Campylobacter* has also shown.[7]

An example of this can be found within *Salmonella* Typhimurium where 26/28 'other pet' sequences clustered monophyletically and 2 isolates clustered elsewhere in the population. The RF successfully classified those 26 sequences and miss-assigned the 2 other isolates. Furthermore, no additional sequences outside of that branch were attributed to this class. In comparison, a source of similar size, cats, had an F1 of 0.107. For this source, most of its sequences were distributed across the phylogeny and miss-classified. These sources with no clonal signal are probably unreliable for attribution. The phylogenetic analysis suggests that there is unlikely on-going transmission within these populations and therefore they represent transient infections, presumably originating from the main source they are embedded within.

Despite these limitations, even moderate improvements in predicting minor sources could provide meaningful added value over traditional typing methods, which already perform well for common sources, and could support real-time surveillance at national and EU levels. ML could therefore complement existing genomic surveillance with limited efforts, though reliance on sequencing data may limit implementation in low- and middle-income countries.

Several animal sources included in this work are not known to be natural reservoirs of *Salmonella*, and they most likely acquire the infection from the same sources as the clinical isolates. Thus, *Salmonella* strains isolated from dogs may originate from the same sources as for humans. In recent years, raw pet food has become recognised as a higher risk factor for exposure to pathogens such as *Salmonella*.[26] Dogs are usually asymptomatic carriers, shedding the bacteria for more than 6 weeks.[27] Cats are not considered to be a major source of human salmonellosis, but may play a role in spreading infection in rural environments e.g. recolonisation of livestock farms.[27] Cats also suffer recurring outbreaks of passerine-adapted lineages of *S.* Typhimurium, which can to lead to clusters of human cases.[28] Considering these sources as either 'intermediate' infections or clinical cases may improve the accuracy of the model and therefore the attribution estimates, but further data collection from currently under sampled categories would be required (e.g. wildlife).

Conclusions

This study provides comprehensive source attribution estimates for the five of the most abundant *Salmonella* serovars in Europe. The findings are in line with past attribution estimates utilising smaller datasets or less discriminatory methods. We demonstrated the challenges of using genomic population-based features in ML source attribution, whereby the lack of clonality of some sources can hamper their prediction. However, it is also likely that these sources (e.g., pets) do not significantly contribute to the transmission of human infection. Therefore, attributing them to the livestock compartments in which they are often found is both logical and appropriate. Moreover, there are clear indications that livestock remains the main source for human infection. National action plans in the poultry sector have been shown to successfully reduce the prevalence as well as human infections.[1] In contrast, implementation of effective *Salmonella* control efforts in pigs and pork is still missing in most EU countries.

Country attributions revealed that many human clinical isolates are attributed to countries other than their own, highlighting the geographical interconnection of sources, and the value of internationally harmonised *Salmonella*-control policies in the food production chain.

In the future, improving sampling and data availability from both the human and veterinary sectors, as well as refocussing on the primary reservoirs, is likely to further refine the results. Additionally, different genomic features, such as those genetic markers that may be representative of host adaption (e.g., acquired/lost genes) may improve performance.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.

Research at the National Veterinary Research Institute (PIWet), Poland, was also partially supported by the Polish Ministry of Education and Science from the funds for science in the years 2018–2022 allocated for the implementation of a co-financed international project.

Marie Anne Chattaway is affiliated to the National Institute for Health Research, Health Protection Research Unit (NIHR HPRU) in Genomics and Enabling Data at University of Warwick in partnership with the UK Health Security Agency (UKHSA). Marie Anne Chattaway is based at UKHSA. The views expressed are those of the author(s) and not necessarily those of the NIHR, the Department of Health and Social Care or the UK Health Security Agency.

Declaration of Competing Interest

The authors have no interests to declare.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jinf.2025.106632.

References

1. Food Safety Authority E, Centre for Disease Prevention E. *The European Union One Health 2023 zoonoses report. EFSA J* 2024;**22**(12):e9106. [cited 2025 Apr 12]. ⟨https://onlinelibrary.wiley.com/doi/full/10.2903/j.efsa.2024.9106⟩.
2. Mulvey MR, Boyd DA, Olson AB, Doublet B, Cloeckaert A. *The genetics of Salmonella genomic island 1. Microbes Infect* 2006;**8**(7):1915–22.
3. Knodler LA, Elfenbein JR. *Salmonella enterica. Trends Microbiol* 2019;**27**(11):964–5.
4. Pires SM, Evers EG, Van Pelt W, Ayers T, Scallan E, Angulo FJ, et al. *Attributing the human disease burden of foodborne infections to specific sources. Foodborne Pathog Dis* 2009;**6**(4):417–24.
5. Koutsoumanis K, Allende A, Alvarez-Ordóñez A, Bolton D, Bover-Cid S, Chemaly M, et al. *Whole genome sequencing and metagenomics for outbreak investigation, source attribution and risk assessment of food-borne microorganisms. EFSA J* 2019;**17**(12):e05898. cited 2023 Jul 19];. ⟨https://onlinelibrary.wiley.com/doi/full/10.2903/j.efsa.2019.5898⟩.
6. Helwigh B, Munck NSM, Litrup E. *Trends and sources in human salmonellosis.* Annual report on zoonoses in Denmark 2019. Technical University of Denmark; 2020. p. 20–3.
7. Thystrup C, Brinch ML, Henri C, Mughini-Gras L, Franz E, Wieczorek K, et al. *Source attribution of human Campylobacter infection: a multi-country model in the European Union. Front Microbiol* 2025;**16**:1519189. [cited 2025 Apr 12]. ⟨https://onehealthejp.eu/jrp-discover⟩.
8. Murigu Kamau Njage P, Henri C, Leekitcharoenphon P, Mistou MY, Hendriksen RS, Hald T. *Machine learning methods as a tool for predicting risk of illness applying next-generation sequencing data. Risk Anal* 2019;**39**(6):1397–413.. [cited 2023 Jul 19]. ⟨https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.13239⟩.
9. Lupolova N, Dallman TJ, Holden NJ, Gally DL. *Patchy promiscuity: machine learning applied to predict the host specificity of Salmonella enterica and Escherichia coli. Microb Genom* 2017;**3**(10):e000135.
10. Arning N, Sheppard SK, Bayliss S, Clifton DA, Wilson DJ. *Machine learning to predict the source of campylobacteriosis using whole genome data. PLoS Genet* 2021;**17**(10):e1009436.
11. Munck N, Murigu P, Njage K, Leekitcharoenphon P, Litrup E, Hald T. *Application of whole-genome sequences and machine learning in source attribution of Salmonella typhimurium. Risk Anal* 2020;**40**(9):1693–705.. [cited 2023 Jul 19]. ⟨https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.13510⟩.
12. Mughini-Gras L, Paganini JA, Guo R, Coipan CE, Friesema IHM, van Hoek AHAM, et al. *Source attribution of Listeria monocytogenes in the Netherlands. Int J Food Microbiol* 2025;**427**(16):110953.
13. Hald T. ⟨https://onehealthejp.eu/projects/foodborne-zoonoses/jrp-discover⟩. DISCoVeR.
14. Deneke C, Uelze L, Brendebach H, Tausch SH, Malorny B. *Decentralized investigation of bacterial outbreaks based on hashed cgMLST. Front Microbiol* 2021;**12**:649517.

15. Dyer NP, Päuker B, Baxter L, Gupta A, Bunk B, Overmann J, et al. *EnteroBase in 2025: exploring the genomic epidemiology of bacterial pathogens. Nucleic Acids Res* 2025;**53**(D1):D757–62. https://doi.org/10.1093/nar/gkae902. [cited 2025 Apr 16].
16. Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, et al. *GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res* 2018;**28**(9):1395–404.
17. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. *eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol* 2004;**186**(5):1518–30.
18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. *Scikit-learn: machine learning in Python. J Mach Learn Res* 2011;**12**:2825–30.
19. Ul Haq I, Gondal I, Vamplew P, Brown S. *Categorical features transformation with compact one-hot encoder for fraud detection in distributed environment.* Data mining: 16th Australasian conference, AusDM 2018, Bahrurst, NSW, Australia, November 28–30, 2018, revised selected papers 16. Springer; 2019. p. 69–80.
20. Kohavi R, John GH. *Wrappers for feature subset selection. Artif Intell* 1997;**97**(1–2):273–324.
21. Shi L, Westerhuis JA, Rosén J, Landberg R, Brunius C. *Variable selection and validation in multivariate modelling. Bioinformatics* 2019;**35**(6):972–80.
22. Lemaître G, Nogueira F, Aridas CK. *Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res* 2017;**18**(1):559–63.
23. Kuhn M, Johnson K. *Applied predictive modeling.* New York: Springer; 2013.
24. Mughini-Gras L, Heck M, van Pelt W. *Increase in reptile-associated human salmonellosis and shift toward adulthood in the age groups at risk, the Netherlands, 1985 to 2014. Eurosurveillance* 2016;**21**(34):30324.
25. Pijnacker R, Dallman TJ, Tijsma ASL, Hawkins G, Larkin L, Kotila SM, et al. *An international outbreak of Salmonella enterica serotype Enteritidis linked to eggs from Poland: a microbiological and epidemiological study. Lancet Infect Dis* 2019;**19**(7):778–86.
26. Morgan G, Pinchbeck G, Taymaz E, Chattaway MA, Schmidt V, Williams N. *An investigation of the presence and antimicrobial susceptibility of Enterobacteriaceae in raw and cooked kibble diets for dogs in the United Kingdom. Front Microbiol* 2023;**14**:1301841.
27. Dróżdż M, Małaszczuk M, Paluch E, Pawlak A. *Zoonotic potential and prevalence of Salmonella serovars isolated from pets. Infect Ecol Epidemiol* 2021;**11**(1):1975530.
28. Söderlund R, Jernberg C, Trönnberg L, Pääjärvi A, Ågren E, Lahti E. *Linked seasonal outbreaks of Salmonella Typhimurium among passerine birds, domestic cats and humans, Sweden, 2009 to 2016. Eur Surveill* 2019;**24**(34):1900074.