A Data-driven Horizon Scan of Bacterial Pathogens at the Wildlife–livestock Interface

Michelle V. Evans1,2,3 and John M. Drake2,3

1MIVEGEC, Institut de Recherche pour le Développement, 34000 Montpellier, France
2Odum School of Ecology, University of Georgia, Athens 30606
3Center for Ecology of Infectious Diseases, University of Georgia, Athens 30606

Abstract: Many livestock diseases rely on wildlife for the transmission or maintenance of the pathogen, and the wildlife–livestock interface represents a potential site of disease emergence for novel pathogens in livestock. Predicting which pathogen species are most likely to emerge in the future is an important challenge for infectious disease surveillance and intelligence. We used a machine learning approach to conduct a data-driven horizon scan of bacterial associations at the wildlife–livestock interface for cows, sheep, and pigs. Our model identified and ranked from 76 to 189 potential novel bacterial species that might associate with each livestock species. Wildlife reservoirs of known and novel bacteria were shared among all three species, suggesting that targeting surveillance and/or control efforts towards these reservoirs could contribute disproportionately to reducing spillover risk to livestock. By predicting pathogen-host associations at the wildlife–livestock interface, we demonstrate one way to plan for and prevent disease emergence in livestock.

Keywords: spillover, livestock, bacteria, wildlife reservoirs

INTRODUCTION

Livestock diseases are a pressing economic and public health threat (Perry et al. 2013). While there is no official estimate of the total economic cost of livestock diseases, it includes both direct and indirect costs to public health, private health, household income, and agricultural production (Narrod et al. 2012). For example, the direct economic cost of production losses due to foot-and-mouth disease was estimated to be over 7.6 billion USD per year globally, without including indirect costs due to revenue losses and disease control (Knight-Jones and Rushton 2013). Further, domestic livestock often serve as reservoirs or bridge species for zoonotic diseases (Hassell et al. 2017), such as domestic swine for Nipah virus (Epstein et al. 2006) and domestic camels for Middle Eastern Respiratory Syndrome-related coronavirus (Azhar et al. 2014). The majority of livestock pathogens infect multiple hosts (Cleaveland et al. 2001), and 79% of diseases on the notifiable disease list of domestic livestock maintained by the World Organization for Animal Health (OIE) involve wildlife in the transmission or maintenance of the pathogen (Miller et al. 2013). In addition, wildlife can contribute to disease transmission and emergence in livestock via cross-
species transmission at the wildlife–livestock interface (i.e. transboundary disease (Siembieda et al. 2011)).

Transmission at the wildlife–livestock interface has been implicated in multiple livestock diseases. Often the directionality of the spillover is from domestic animals to wildlife, as for rinderpest virus in East Africa (Kock et al. 2006) and for rabies transmission between domestic dogs and African wild dogs (Prager et al. 2012). However, wildlife species also serve as reservoirs of livestock disease. Wild birds are reservoirs of avian influenza and have been implicated as the origin of multiple outbreaks of high-pathogenic avian influenza in domestic birds (Causey and Edwards 2008). Further, migration patterns of wild birds have allowed for rapid global spread of domestic avian influenza strains that would not have been possible without cross-species transmission from these wildlife reservoirs (Clark and Hall 2006). Other diseases are maintained primarily in wildlife reservoirs, thwarting eradication efforts that rely solely on livestock vaccination and disease control. For example, brucellosis prevalence in US cattle herds decreased from over 10% in the 1930s to near eradication by the early 2000s (Schumaker et al. 2012), and currently, outbreaks of brucellosis in cattle in the US are due almost exclusively to cross-species transmission from wild elk or bison (Rhyan et al. 2013).

Like all cross-species interfaces, the wildlife–livestock interface is also a site of novel pathogen emergence, whose frequency has been increasing over the past several decades (Jones et al. 2008; Wiethoelter et al. 2015). Examples of recently emerged diseases at this interface include Nipah and Hendra viruses, which spilled over from bats to pigs and horses, respectively (Daszak et al. 2006). One tool of disease emergence prevention is identifying candidate pathogens and hosts that may serve as sites of future disease spillover by predicting pathogens capable of cross-species transmission. In general, this method uses some form of statistical modeling, often based in machine-learning, to predict novel pathogen-host interactions by leveraging information on species traits and the network structure of known pathogen-host interactions (Becker et al. 2019b). This approach has been applied to predict rodent reservoirs of zoonotic diseases (Han et al. 2015), zoonotic spillover of viral pathogens from mammals (Olival et al. 2017) and parasite acquisition of introduced mammal species (Schatz and Park 2021). While the number of systematic, predictive studies of disease emergence, particularly zoonoses, has increased dramatically within the past ten years (Carlson et al. 2021), comprehensive, global studies of disease emergence at the wildlife–livestock interface remain relatively rare (Wiethoelter et al. 2015).

Determining which bacterial species can associate with which hosts is a first step toward predicting pathogen emergence. Here, we present a horizon scan of known and novel bacterial cross-species transmission at the wildlife–livestock mammal interface, applying established methods from previous systematic predictive studies (Han et al. 2016b; Evans et al. 2017; Majewski et al. 2021). We identify potential novel bacterial infections in three focal livestock species (Bos taurus, Ovis aries, Sus scrofa domesticus) by using a machine learning model to estimate the association propensity for each mammal-bacterium interaction, leveraging covariates created from the life history traits of each host and bacteria in the pair and phylogenetic relationships within the known host and bacterial communities. The objective of this study is to identify particular bacterial orders that are likely to contain novel pathogen associations, as well as wildlife reservoir species that could be targeted for veterinary disease surveillance, control, and prevention. We also aim to identify covariates associated with a higher propensity of a mammal-bacteria association. We hypothesize that hosts and bacteria that already have a high number of known associations would also have a high number of novel associations (Han et al. 2016a; Dallas et al. 2017). In addition, traits related to taxonomy and phylogeny have been shown to predict host-bacterium associations (Shaw et al. 2020; Albery et al. 2020), and we hypothesize they would be of similar importance in this study. The results of this study provide a foundation for future studies of emerging bacterial disease in livestock.

**METHODS**

**Data Collection**

**Mammal-Bacteria Association Matrix**

We constructed a species interaction network of known mammal-bacterium associations from three existing datasets: the Enhanced Infectious Diseases Organisms Interactions database (Wardeh et al. 2015), the Global Mammal Parasite Database (Stephens et al. 2017), and Shaw et al. (2020). For all three datasets, we only included mammal hosts (both wildlife and domestic, including humans) and bacterial species. Our set of mammal species comprised all mammals in any of the three datasets, including those that
were not known to host any bacterial species. Our set of bacteria species, on the other hand, included only those found in at least one mammal host. Instances where either the host or bacteria were not identified to species were not included. Mammal names were standardized to IUCN species names (IUCN 2021) and bacteria names were standardized to NCBI species names (Federhen 2012). This procedure resulted in a total of 4,753 known host-bacteria pairs across 1,223 unique mammal hosts and 1,672 unique bacterial species.

**Mammal Host Traits**

We compiled data on mammal host taxonomy, life history, spatial distribution, known bacterial interactions, and research effort. Mammal taxonomy, specifically host order, family, and genus, were collected from the NCBI database via the taxize package (Chamberlain and Szocs 2013). Host diet proportions, foraging strata, activity (e.g. diurnal, nocturnal, and/or crepuscular), and body mass were collected from the EltonTraits 1.0 database (Wilman et al. 2014). For those species missing from this database ($n = 11$), we used genus-level values ($n = 9$) or searched the primary literature ($n = 2$). We collected data on litter size from Cooke et al. (Cooke et al. 2019). For species missing from this database ($n = 150$), we used genus-level means ($n = 55$) or searched the primary literature ($n = 95$). Generation length was collected from Pacifici et al. (Pacifici et al. 2013). For species missing from this database ($n = 77$), we used traits from the primary literature.

We calculated covariates relevant to mammal host species spatial distributions from each species range map. The majority ($n = 1,206$) had existing range maps in the IUCN Red List of Threatened Species database (www.iucnredlist.org/). The remaining range maps were sourced from the primary literature or, for domestic mammals, created from the Gridded Livestock of the World database (Gilbert et al. 2018). From these range maps, we calculated the area of a species’ range in km$^2$, continental breadth (i.e. number of continents in which a species is present), and the latitude and longitude of the centroid of that species’ range. We also calculated habitat breadth, measured by the number of ecoregions within each species range. Ecoregions were defined via the Terrestrial Ecoregions of the World (Olson et al. 2001) and the Marine Ecoregions of the World (Spalding et al. 2007) datasets for terrestrial and marine mammals, respectively. All spatial data manipulation and analyses were conducted using the sf (Pebesma 2018), raster (Hijmans 2020), and rgdal (Bivand et al. 2021) packages in R v. 4.0.3 (R Core Team 2018).

From the mammal-bacteria association matrix, we calculated two covariates derived from the number of bacterial species known to associate with each host species. This included the number of known bacterial species and number of known bacterial families found in each host species.

As a measure of research intensity, we recorded the number of PubMed entries for each mammal host species and disease, using the search term “$host scientific name$” AND “disease” OR “bacteri*” OR “parasit*”. This was collected on Nov 19 2020 using the rentrez package (Winter 2017) in R v. 4.0.3 (R Core Team 2018).

**Bacterial Traits**

We compiled data on bacteria taxonomy, life history, patterns based on known host associations from the mammal-bacteria association matrix, and research effort. Bacteria taxonomy was collected from the NCBI Taxonomy Database. Life history traits (vector-borne, gram-stain, motility, and oxygen use) were primarily sourced from Shaw et al. (2020). For those species which were missing from this database ($n = 172$), we collected information on these traits from the NCBI Genome database or from the primary literature. For those species for which we could not find information (between 24 and 90, depending on the trait), we used genus-level trait values. Genomic data (genome size, GC content, and the number of genes) were primarily sourced from Shaw et al. (2020). For those species which were missing ($n = 639$), we either collected data from the NCBI Genome database ($n = 498$) or used the genus-level mean ($n = 141$). Finally, we recorded the number of PubMed entries for each bacteria species scientific name, accessed via the rentrez package (Winter 2017) in R v. 4.0.3 (R Core Team 2018) on Nov 19 2020 (Winter 2017).

We also created covariates derived from the known hosts of each bacteria species. This included the number of host species in which each bacterial species was found (i.e. host species breadth) and the mean phylogenetic distance between each known host species, referred to as phylogenetic host breadth, as described in Shaw et al. (2020). Node-based phylogenetic distances were calculated from a mean of a random subset of 100 mammal supertrees from Upham et al. (Upham et al. 2019). For each potential host-
bacterium pair, we also calculated the mean phylogenetic distance between that host and all known hosts of that bacterium. Phylogenetic analyses were conducted using the ape package (Paradis and Schliep 2019) in R v. 4.0.3 (R Core Team 2018).

We assessed the full collection of traits for collinearity among covariates and removed those covariates with a Spearman’s $\rho > 0.7$, resulting in 10 bacterial traits, 13 host traits, and one trait specific to each host-bacterium pair (phylogenetic distance to known hosts) included in the final model (Table S1).

Model Fitting & Validation

We used boosted regression trees (BRT) to estimate the association propensity of each host-bacterium association. BRT trains regression trees to predict an outcome variable based on recursive binary splits of predictor covariates (Elith et al. 2008). In our approach, the outcome variable is binary, representing known and unknown, or absence of known associations, similar to cases and contaminated-controls in other analyses of binary classification, such as logistic regression (Lancaster and Imbens 1996). The cases represent sampled “uses” of a host by a bacterium (e.g. a known host-bacterium association), and the controls represent available hosts for each bacterium (e.g. absence of a known host-bacterium association). This is similar to use-availability data in habitat selection studies (Keating and Cherry 2004) or pseudo-absence data in species distribution modeling (Elith and Leathwick 2009). We use the model to identify the potential bacterium-host associations that have not been sampled yet, but are biologically possible (e.g. the controls that are contaminated). We treat the unknown host-bacterium associations as ‘true absences’ in the model (Ward et al. 2009) and so are unable to estimate the absolute probability of an association because the number of host-bacterium associations sampled is not necessarily proportional to the occurrence of associations in the population (Rota et al. 2013, but see Royle et al. 2012 for instances where this is not the case). Instead, the estimated value increases monotonically with the absolute probability, and therefore has an identical rank. Following Evans et al. (2017), we refer to this value as the propensity of a host-bacterium association to avoid confusing it with the absolute probability of an association. Individual trees are iteratively trained via adaptive resampling (“boosting”) at each split. Finally, all trees are combined to create a strongly predictive ensemble consisting of a collection of relatively poor learners. We used the xgboost algorithm to train our model (Chen and Guestrin 2016). To reduce the effects of bias in the training data, we used a subset of the data at each iteration (e.g. start of a new tree) by only fitting each tree to a random subset of half of the observations and half of the covariates. Our data had a very low number of known associations relative to unknown associations (0.002 prevalence), which is known to bias prediction in the xgboost algorithm where the default learning parameters weight misclassification of cases and controls equally (Chen and Guestrin 2016). We sought to manage this imbalance in our data by scaling the gradient of the loss function of known associations by the ratio of unknown associations (controls) to known associations (cases) (435) within the xgboost algorithm (Chen and Guestrin 2016), so that misclassification of known associations resulted in larger corrections to the model.

The final dataset was randomly split into training and testing sets consisting of 70% and 30% of the data, respectively, preserving the proportion of positive and negative outcomes between the two. We used the training data to tune six parameters of the model (Table S2) via ten-fold cross-validation that incorporated three performance metrics: AUC (area under the receiver operating curve), true skill statistic (TSS), and the Boyce Index (BI, a measure of model performance that has been developed for classification using contaminated-control data) (Boyce et al. 2002). Because we accounted for the imbalance of positive and negative samples by scaling the positive observations directly in the xgboost algorithm, we used a balanced threshold of 0.5 to turn continuous predictions into binary responses to evaluate the TSS. After the initial tuning step, we fit a model with the set of six tuned parameters and used ten-fold cross-validation to determine the optimal number of iterations (i.e. trees) for this specific parameter set, up to 300. We chose the number of iterations that maximized performance on the proportion of folds not used for fitting (i.e., out-of-fold data) while minimizing the difference in performance between in-fold and out-of-fold data, a form of early stopping used to regularize boosting models (Zhang and Yu 2005). Cross-validation was evaluated via AUC. The optimal number of iterations was 120 (Fig. S1). The tuned model was then used to predict over the testing dataset and performance was evaluated via AUC, TSS, and BI.

The outputs of our BRT model are continuous predicted probabilities of each host-bacterium association, which increase monotonically with the absolute probability
(Lancaster and Imbens 1996). To avoid confounding this value with absolute probability, we refer to these predictions as propensities. To transform these continuous values into binary predicted novel associations, we defined novel bacteria as those ranked above the lowest known bacteria species for each livestock species following Evans et al. (2017), a threshold we refer to as the “least known case”. Predictive models are necessarily biased by the threshold choice when converting continuous values to binary classifications (Nenzen and Araujo 2011), particularly when controls are contaminated (Liu et al. 2015). The choice of threshold is determined by both the statistical approach and the scientific question. We chose the “least known case” threshold because it is the most conservative threshold that retains all known associations in the training data set (Pearson et al. 2006), and therefore translates well to a list of bacteria or hosts to prioritize for surveillance and diagnostics. Less conservative thresholds would result in a higher number of novel associations, but would not impact the ranking of top bacteria or reservoir species. We then examined the known and predicted bacterial and wildlife reservoir community of three focal livestock species chosen for their economic importance: cattle (Bos taurus), sheep (Ovis aries), and domestic swine (Sus scrofa domesticus). Sheep and cattle are the two largest domestic mammal stocks globally, and cattle and swine provide the majority of the world’s meat consumption of domestic mammal species (Food and Agriculture Organization of the United Nations 2022). We also examined the traits of livestock diseases identified as notifiable diseases by the OIE (World Organization for Animal Health 2021) in the context of our model results to explore our model’s ability to provide information about specific diseases of concern. Specifically, we summarized the propensities and predictive traits of OIE notifiable bacterial diseases in our dataset.

Recognizing that our data suffers from bias due to variable intensity in scientific scrutiny, we performed two sensitivity analyses to assess the robustness of our findings. Specially, the data exhibit bias towards bacteria known to be found in humans and North American mammals. We therefore repeated the analysis with subsets of the data where 1) human-bacteria interactions were excluded from the data, and 2) the data only included mammals from North America. The methods and full results of these supplemental models are reported in the supplement.

All data and code to reproduce the analyses are available on figshare (https://doi.org/10.6084/m9.figshare.15073299).

RESULTS

Our final data set consisted of 1,223 unique mammal hosts and 1,672 unique bacterial species. Of the possible 2,044,856 bacteria-host associations, 4,753 (0.23%) have been reported. Our coverage of bacterial orders was relatively even, while our coverage of mammal host orders was over-represented by those species found within the Global Mammal Parasite Database (i.e., Primates, Carnivora, and Artiodactyla), which together represented 80.0% of known host-bacteria interactions (Fig. 1). The high coverage in primates was primarily due to host-bacterium associations involving humans, which made up 27.4% of all host-bacterium associations. In addition, 77.8% of all bacteria have been found in humans, and 64.6% of all bacterial species in our data have been found only in humans. Coverage was also high among domestic mammals: after humans, five of the top seven mammal species with the most known bacteria are domestic (Bos taurus, Felis catus, Ovis aries, Sus scrofa domesticus, Capra hircus). Most bacteria and hosts were included in only a few known associations. The majority of mammal hosts in our data were not known to host any bacteria (n = 734) or hosted only one bacteria species (n = 223). Similarly, 1,066 bacterial species were known to be found in only one mammal host.

Our BRT model performed well on both the training and testing datasets (Table 1). Similar performance on both data sets indicates that the model was not overfit to the training data. In addition, our model predictions range from 0.001 to 0.998, with a median of 0.003. This median value is close to the mean prevalence in our full dataset (0.002), suggesting that our model appropriately accounted for the high imbalance between known and unknown associations. Taken together, these results suggest our model adjusted well to the biases in our data, while retaining its generalizability to the testing data set.

All 24 covariates were included in the model at least once, but the majority contributed very little to model performance with only ten covariates resulting in a gain of more than 0.01 AUC over all iterations (Fig. 2). The four covariates that contributed the most to model performance were associated with the existing diversity of the known community of hosts or bacteria for each species, including measures of both phylogenetic and taxonomic diversity. All three diversity covariates (bacterial breadth, host breadth, and phylogenetic host breadth) had an asymptotically increasing relationship with the propensity of a specific
host-parasite association (Fig. 3A,C,D). That is, increasing diversity led to large increases in the propensity of an association at low diversity levels, but had little effect at higher levels of diversity. Similarly, the relationship between the phylogenetic distance to known hosts and propensity of association was exponentially decreasing (Fig. 3B). If a potential mammal host was closely related to the community of known hosts of a bacterium, the model assigned a high propensity to that host-bacterium association. Hosts with a phylogenetic distance value from known hosts of over 150 Ma had low propensity to associate with that bacterium. For context, this distance is similar to that between humans and our three livestock hosts. Research effort of both hosts and bacteria also arose as important in our model (Fig. 2), with increasing propensity of association for more highly studied species (Fig. 3E). Our three livestock species had research effort values (i.e. number of PubMed entries) of 73,852, 3,091, and 51 for B. taurus, O. aries, and S. scrofa domesticus, respectively. It should be noted that Sus scrofa, of which S. scrofa domesticus is the domesticated subspecies, had a research effort of 6,299.

We defined novel bacteria as those ranked above the lowest known bacteria species for each livestock host species. This resulted in 285, 168, 154 known and 76, 88, and 189 novel bacteria for B. taurus, O. aries, and S. scrofa domesticus, respectively (Table S3, Fig. S2). While B. taurus and O. aries had similar ratios of known to novel bacteria species, our model predicted more novel than known bacteria for S. scrofa domesticus (Fig. 4). S. scrofa domesticus also had the most high-ranking novel bacteria of the three livestock species (Fig. S2). All three livestock species are predicted to host many bacteria (both known and novel) in three bacteria orders: Lactobacillales, Corynebacteriales, and Bacillales (Fig. 4).

Table 1. Performance of Boosted Regression Tree Model on Training and Testing Datasets as Measured by Area Under the Receiving Operator Curve (AUC), True Skill Statistic (TSS), and Boyce Index (BI).

|               | Training | Testing |
|---------------|----------|---------|
| AUC           | 0.9985   | 0.9980  |
| True Skill Statistic | 0.9768   | 0.9598  |
| Boyce Index   | 0.9980   | 0.9542  |

TSS and BI were calculated using a threshold of 0.5 to transform continuous predictions to binary values.

Figure 1. Number of known host-bacterium associations at the species level within each paired group of host and bacteria orders. Darker red colors represent more known associations of bacteria of that order in that mammal host order, on a natural log scale, and white represents no known associations.

Table 1. Performance of Boosted Regression Tree Model on Training and Testing Datasets as Measured by Area Under the Receiving Operator Curve (AUC), True Skill Statistic (TSS), and Boyce Index (BI).

|               | Training | Testing |
|---------------|----------|---------|
| AUC           | 0.9985   | 0.9980  |
| True Skill Statistic | 0.9768   | 0.9598  |
| Boyce Index   | 0.9980   | 0.9542  |

TSS and BI were calculated using a threshold of 0.5 to transform continuous predictions to binary values.
Figure 2. Importance of top ten covariates, as measured by total gain in AUC due to tree splits including the covariate. Covariates are colored based on whether they were a trait of the bacteria species, host species, or unique to that host-bacterium pair.

Figure 3. Partial-dependence plots of top six most important covariates. Rugs along the x-axis represent the distribution of each covariate in our training dataset.
mean association with our three focal livestock species of 0.982, signifying that they are very likely to infect livestock species. However, their mean predicted propensity of infecting all mammal species was 0.23. Interestingly, these diseases did not necessarily have traits that our model associated with higher predicted propensities of mammal host infection. For example, the phylogenetic host breadth ranged between 40.45—146.15 for these bacteria species, which includes values associated with low propensity of association (Fig. 3 D). In addition, the mean range of host breadth of these bacteria was 26.26, but this varied widely depending on the bacterial species, with a range of 3—88 mammal hosts across the nineteen bacteria.

From our dataset of predictions, we identified the top ten reservoir wildlife species for each livestock species, including both known and novel bacteria associations (Table 2). This does not include predicting novel species of wildlife, but rather identifying wildlife species that are known to associate with predicted novel species of livestock. In general, wildlife species known to host novel bacteria of livestock species are also hosts of known bacteria species. Many of the wildlife reservoirs were shared across livestock species, with seven wildlife species (white-tailed deer (*Odocoileus virginianus*), bighorn sheep (*Ovis canadensis*), harbor seal (*Phoca vitulina*), racoon (*Procyon lotor*), American black bear (*Ursus americanus*), red fox (*Vulpes vulpes*), California sea lion (*Zalophus californianus*)) ranking highly for all three livestock species. Wildlife reservoirs of bacteria able to infect *S. scrofa domesticus* were predicted to have high numbers of both novel and known bacterial species, while reservoirs of bacteria able to infect *B. taurus* primarily were associated with bacterial species already known to infect *B. taurus*. In addition to the top-ranked reservoir species, 266 wild mammals are known to host at least one predicted novel bacteria of the three livestock species and 378 wild mammal species are known to host at least one known livestock bacterial species. This high number of potential wildlife hosts of livestock bacteria suggests there is high potential for transboundary disease transmission at the wildlife–livestock interface.

Results from our two supplemental models had high agreement with our original model, suggesting our results are not an artifact of sampling bias in our data. Relative variable importance was similar across all three models.
Fig. S4). In addition, the predicted association propensities from the two models were highly correlated with predictions from our original model (model without humans $\rho = 0.964$, N America model $\rho = 0.859$). Additional information on supplemental model fit and performance is reported in the Supplemental Materials.

**DISCUSSION**

Although relatively understudied compared to zoonotic spillover, transboundary transmission at the wildlife–livestock interface poses large economic and human health risks (Wiethoelter et al. 2015). To increase our understanding of disease spillover at this interface, we conducted a data-driven approach to create a horizon scan of bacterial pathogens of mammal livestock to identify candidate bacterial groups that could pose a threat to the livestock industry. Our model identifies several bacterial species that have the potential to associate with livestock species, which could increase the size of the bacterial community in each species from 23 to 123%. In addition, our results highlight wildlife species that are likely to serve as reservoirs of infectious diseases, providing a guide for future One Health-based interventions.

Several bacterial orders had high numbers of known and predicted bacteria capable of infecting livestock, including some with pathogenic properties in livestock and humans. The order *Bacillales* includes bacteria that pose severe health risks to humans, such as *B. anthracis*, as well as bacteria that serve as contaminating organisms in agricultural and dairy production, such as listeria (Logan 1988). Similarly, many species within the order *Corynebacteriales* are known pathogens for mammal hosts, particularly those in the *Mycobacterium* and *Corynebacterium* genera (Coimbra et al. 2020), although bacteria in other genera are known to cause infection in immuno-compromised hosts (Sowani et al. 2018). *M. bovis*, the causative agent of tuberculosis in cattle, is a bacterium in this order that poses a known threat at the wildlife–livestock interface (Woodroffe et al. 2009). Bacteria in the order *Lactobacillales* are primarily associated with natural fermentation (Vinderola et al. 2019). However, some *Lactobacillales* bacteria of the genera *Streptococcus* and *Ente-

**Table 2.** Top Ten Wildlife Reservoirs of Known and Novel Bacteria Predicted to be Associated with Livestock Species, Ordered in Decreasing Order of Total Shared Bacteria Known to be Found in Each Wildlife Reservoir.

| Host                  | Novel | Known | Total |
|-----------------------|-------|-------|-------|
| *Ovis canadensis*     | 8     | 25    | 33    |
| *Procyon lotor*       | 7     | 24    | 31    |
| *Odocoileus virginianus* | 5   | 25    | 30    |
| *Phoca vitulina*      | 1     | 27    | 28    |
| *Zalophus californianus* | 5  | 23    | 28    |
| *Vulpes vulpes*       | 3     | 17    | 20    |
| *Ursus americanus*    | 1     | 18    | 19    |
| *Mirounga angustirostris* | 0  | 18    | 18    |
| *Odocoileus hemionus* | 3     | 14    | 17    |
| *Alces alces*         | 0     | 16    | 16    |
| *Ovis aries*          | 11    | 28    | 39    |
| *Odocoileus virginianus* | 4  | 30    | 34    |
| *Phoca vitulina*      | 5     | 21    | 26    |
| *Vulpes vulpes*       | 3     | 21    | 24    |
| *Zalophus californianus* | 5  | 17    | 22    |
| *Ursus americanus*    | 1     | 20    | 21    |
| *Alces alces*         | 2     | 18    | 20    |
| *Rupicapra rupicapra* | 2     | 17    | 19    |
| *Odocoileus hemionus* | 4     | 14    | 18    |
| *Canis latrans*       | 10    | 6     | 16    |

Known bacteria species are those currently found in the livestock host and novel bacteria species are those bacteria species ranked above the lowest known positive for that livestock host by our model. All bacterial species are known to associate with that wildlife reservoir (e.g. are not based on model predictions).
**Bacteria at the Wildlife–Livestock Interface**

...rococcus are pathogenic in livestock, causing meningitis and mastitis (Chanter 1997). In contrast, the order Rhizobiales (Hyphomicrobiales) contains the causative agents of bartonella and brucellosis, two diseases of concern at the wildlife–livestock interface (Chang et al. 2000; Godfroid 2017), but this order contained relatively few known and novel bacteria species for all three livestock.

OIE notifiable diseases were correctly identified as diseases of livestock, but their traits were not necessarily associated with high propensity of infection in mammals. This may be because our model’s objective is only to predict the association propensity of each host-bacterium association, and not the pathogenicity of this association. A bacteria species’ tendency to cause illness or mortality in a host is dependent on both its ability to infect the host and its effect on the host following establishment (Morse 1995). Importantly, our study focused only on the first step, providing the propensity of each host-bacterium association (i.e. infection), many of which likely have a negligible effect on host fitness. Further, the relative importance of both taxonomic and phylogenetic host breadth covariates in our model suggests that those bacteria species with average high propensity of association with mammals in our model are also likely to be generalists. Generalist pathogens are predicted to be less virulent than specialist pathogens, due to the increased fitness costs of adapting to novel species (Antonovics et al. 2013). It is therefore not surprising that the OIE notifiable diseases, many of which have high mortality rates in livestock, have predicted infection distributions more closely resembling a specialist bacterial pathogen, and do not have traits associated with high propensity of association with mammal hosts.

The wildlife reservoir species identified by our model are potential target species for interventions or increased surveillance at the wildlife–livestock interface. These interventions, which aim to control disease in both wildlife and livestock populations, have successfully led to the extirpation or extinction of livestock diseases in the past, notably rinderpest (Morens et al. 2011), and are an integral part of veterinary disease management programs (Gortazar et al. 2015). For example, as B. abortus nears elimination in cattle populations in North America, wild herds of elk and bison serve as reservoirs of the pathogen (Rhyian et al. 2013), and control of disease in wildlife populations is necessary to reduce the risk to livestock (Davis and Elzer 2002; Olsen 2010). In addition, the majority of wildlife reservoir species identified by our model were predicted to serve as reservoirs for multiple bacteria of all three livestock species. Interventions that target these species have the additional benefit of reducing the risk of cross-species transmission for multiple livestock species of economic importance and may represent an efficient use of limited resources.

The taxonomic scale of ecological interactions is also an important component of spillover (Becker et al. 2019a). Our study provided a horizon-scan of bacterial pathogens at the wildlife–livestock interface applied at a very coarse taxonomic scale, and consequently lacks the resolution to predict specific associations and trends within smaller taxonomic groupings. Moreover, host–pathogen association patterns are known to differ according to the phylogenetic scale of analysis (Dallas and Becker 2021), and consideration of these associations across multiple, finer phylogenetic scales would reveal further cladistic patterns in the network (Becker et al. 2019a). Our dataset also suffered from a bias towards human bacteria and North American mammals due to bias in scientific effort and sampling. While our supplementary analysis found that our results were robust to these biases, North American mammals remained overrepresented in our list of top wildlife reservoirs (Table 2). This is because this ranking is based on the total known and potential livestock bacterial species that are known to associate with that wildlife reservoir, and North American mammals share more known bacteria with livestock than in other regions (Fig. S3). When considering all predicted wildlife reservoirs, and not just the top ten, this geographic bias becomes less pronounced, but continues to resemble geographic patterns in wildlife reservoirs of known livestock bacteria (Fig. S3). The goal of this study is to identify novel bacterial species in livestock, and does not extend these predictions to wildlife. The identification of wildlife reservoirs is therefore limited by the data on known bacterial associations with those hosts. In addition, we were limited in our choice of covariates to those that applied to the full range of mammals and bacteria and chose to ignore strain-level differences. Performing similar analyses at the level of bacterial orders, and including more specific biology and traits, such as specific genome mutations, would provide more confident discrimination of species of concern within a family or order. Our model also did not consider mortality and morbidity of diseases, focusing simply on the propensity of a mammal-bacterium association. Most bacteria are non-pathogenic and may even be in commensal relationships with their hosts. While we qualitatively consider our predictions in the context of their OIE
notification status, a logical next step would be to focus a similar predictive analysis on specific traits of concern for that disease (e.g., mortality, zoonotic potential, transmissibility). Indeed, as creation and collation of host–pathogen networks and associated life history traits continues (e.g. CLOVER dataset (Gibb et al. 2021)), analyses such as this will become more feasible.

Finally, we believe our predictive analysis provides further insight into transboundary disease transmission at the wildlife–livestock interface by highlighting macroecological patterns and identifying potential novel bacteria with a high propensity to associate with three focal livestock species. We found that *S. scrofa domesticus* had much higher numbers of predicted novel bacteria than the other two livestock species. Disease emergence at the wildlife–swine interface is less researched than at the wildlife–cattle and wildlife–poultry interfaces (Wiethoelter et al. 2015), and our analysis suggests this interface may be a site of undiscovered livestock bacteria or future spillover events. Importantly, this study contributes to a growing body of literature applying data-driven methods based in macroecological theory to the prediction of disease spillover at non-human interfaces (Luis et al. 2015; Han et al. 2019; Schatz and Park 2021). Although zoonotic disease emergence represents a clear public health threat, disease emergence in livestock also has important economic and public health implications (e.g., when livestock act as bridge species). Our study shows that disease emergence events in livestock need not be viewed as idiosyncratic events, but rather may be anticipated and planned for.

**FUNDING**

Funding was provided by National Science Foundation (DEB 1717282).

**REFERENCES**

Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. 2002. Introduction to Pathogens. Molecular Biology of the Cell. 4th edition.

Albery GF, Eskew EA, Ross N, Olival KJ (2020) Predicting the global mammalian viral sharing network using phylogeography. *Nature Communications* 11:1–9

Antonovics J, Boots M, Ebert D, Koskella B, Poss M, Sadd BM (2013) The origin of specificity by means of natural selection: evolved and nonhost resistance in host-pathogen interactions. *Evolution* 67:1–9

Azhar EI, El-Kafrawy SA, Farraj SA, Hassan AM, Al-Saeed MS, Hashem AM, Madani TA (2014) Evidence for Camel-to-Human Transmission of MERS Coronavirus. *New England Journal of Medicine* 370:2499–2505

Becker DJ, Washburne AD, Faust CL, Mordecai EA, Plowright RK (2019) The problem of scale in the prediction and management of pathogen spillover. *Philosophical Transactions of the Royal Society b: Biological Sciences* 374:20190022

Becker DJ, Washburne AD, Faust CL, Pulliam JRC, Mordecai EA, Lloyd-Smith JO, Plowright RK (2019) Dynamic and integrative approaches to understanding pathogen spillover. *Philosophical Transactions of the Royal Society b: Biological Sciences* 374:20190014

Bivand, R., T. Keitt, and B. Rowlingson. 2021. rgdal: Bindings for the “geospatial” data abstraction library.

Boyce MS, Vernier PR, Nielsen SE, Schmiegelow FKA (2002) Evaluating resource selection functions. *Ecological Modelling* 157:281–300

Carlson, C., M. Farrell, Z. Grange, B. Han, N. Mollenkamp, E. Phelan, A. Rasmussen, G. Albery, B. Bett, D. Brett-Major, L. Cohen, T. Dallas, E. Eskew, A. Fagre, K. Forbes, R. Gibb, S. Halabi, C. Hammer, R. Katz, J. Kindrachuk, R. Muyalera, F. Nutter, J. Ogola, K. Olival, M. Rourke, S. Ryan, N. Ross, S. Seifert, T. Sironen, C. Standley, K. Taylor, M. Venter, and P. Webala. 2021. Zoonotic Risk Technology Enters the Viral Emergence Toolkit. Preprints:20210402002

Causey D, Edwards SV (2008) Ecology of Avian Influenza Virus in Birds. *The Journal of Infectious Diseases* 197:S29–S33

Chamberlain S, Szocs E (2013) taxize - taxonomic search and retrieval in R. *F1000Research* 2:191

Chang CC, Chomel BB, Kasten RW, Heller RM, Ueno H, Yamamoto K, Bleich VC, Pierce BM, Gonzales BJ, Swift PK, Boyce WM, Jang SS, Boulouis HJ, Piémont Y, Rossolini GM, Riccio ML, Cornaglia G, Pagani L, Lagatolla C, Selan L, Fontana R (2000) Bartonella spp. isolated from wild and domestic ruminants in North America. *Emerging Infectious Diseases* 6:306–311

Chanter N (1997) Streptococci and enterococci as animal pathogens. *Journal of Applied Microbiology* 83:100S–1095

Chen, T., and C. Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. Pages 785–794 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA.

Clark L, Hall J (2006) Avian Influenza in Wild Birds: Status as Reservoirs, and Risks to Humans and Agriculture. *Ornithological Monographs* 60:3–29

Cleaveland S, Laurenson MK, Taylor LH (2001) Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philosophical Transactions of the Royal Society of London. Series b, Biological Sciences* 356:991–999

Coimbra NDR, Goes-Neto A, Azevedo V, Ouangraoua A (2020) Reconstructing the phylogeny of Corynebacteriales while accounting for horizontal gene transfer. *Genome Biology and Evolution* 12:381–395

Cooke RSC, Bates AE, Eigenbrod F (2019) Global trade-offs of functional redundancy and functional dispersion for birds and mammals. *Global Ecology and Biogeography* 28:484–495
Dallas TA, Becker DJ (2021) Taxonomic resolution affects host—parasite association model performance. Parasitology 148:584–590.

Dallas T, Park AW, Drake JM (2017) Predictability of helminth parasite host range using information on geography, host traits and parasite community structure. Parasitology 144:200–205.

Daszak P., R. Plotz, J. Epstein, J. Pulliam, S. Abdul Rahman, H. Field, A. Jamaluddin, S. Sharifah, C. Smith, K. Olival, S. Luby, K. Halpin, A. Hyatt, A. Cunningham, and Henipavirus Ecology Research Group. 2006. The emergence of Nipah and Hendra virus: pathogen dynamics across a wildlife–livestock–human continuum. Page in S. K. Collinge and C. Ray, editors. Disease Ecology. Oxford University Press.

Davis DS, Elzer PH (2002) Brucella vaccines in wildlife. Veterinary Microbiology 90:533–544.

Elith J, Leathwick JR (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics 40:677–697.

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. Journal of Animal Ecology 77:802–813.

Epstein JH, Field HE, Luby S, Pulliam JRC, Daszak P (2006) Nipah virus: Impact, origins, and causes of emergence. Current Infectious Disease Reports 8:59–65.

Evans MV, Dallas TA, Han BA, Murdock GC, Drake JM (2017) Data-driven identification of potential Zika virus vectors. eLife 6:e22053.

Federhen S (2012) The NCBI Taxonomy database. Nucleic Acids Research 40:D136–D143.

Food and Agriculture Organization of the United Nations. 2022. FAOSTAT: Statistics Database.

Gibb, R., G. F. Albery, D. J. Becker, L. Brierley, R. Connor, T. A. Dallas, E. A. Eskew, M. J. Farrell, A. L. Rasmussen, S. J. Ryan, A. Sweeney, C. J. Carlsson, and T. Poisot. 2021. Data proliferation, reconciliation, and synthesis in viral ecology. bioRxiv:2021.01.14.426572.

Gilbert M, Nicolas G, Cinardi G, Van Boeckel TP, Vanwambeke SO, Wint GRW, Robinson TP (2018) Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. Scientific Data 5:180227.

Godfroid J (2017) Brucellosis in livestock and wildlife: zoonotic diseases without pandemic potential in need of innovative one health approaches. Archives of Public Health 75:1–6.

Gortazar C, Diez-Delgado I, Barasona JA, Vicente J, De La Fuente A, Boadella M (2015) The wild side of disease control at the wildlife–livestock–human interface: A review. Frontiers in Veterinary Science 1:1–12.

Han, B. A., A. M. Kramer, and J. M. Drake. 2016a. Global Patterns of Zoonotic Disease in Mammals. Trends in Parasitology 1–13.

Han BA, Majumdar S, Calmon FP, Glicksberg BS, Horsesh R, Kumar A, Perer A, von Marschall EB, Wei D, Mojsilovic A, Varshney KR (2019) Confronting data sparsity to identify potential sources of Zika virus spillover infection among primates. Epidemics 27:59–65.

Han BA, Schmidt JP, Alexander LW, Bowden SE, Hayman DTS, Drake JM (2016) Undiscovered bat hosts of Filoviruses. PLoS Neglected Tropical Diseases 10:e0004815–e4910.

Han BA, Schmidt JP, Bowden SE, Drake JM (2015) Rodent reservoirs of future zoonotic diseases. Proceedings of the National Academy of Sciences 112:7039–7044.

Hassell JM, Begon M, Ward MJ, Fève M (2017) Urbanization and Disease Emergence: Dynamics at the Wildlife–Livestock–Human Interface. Trends in Ecology & Evolution 32:55–67.

Hijmans, R. 2020. raster: Geographic data analysis and modeling. IUCN. 2021. The IUCN Red List of Threatened Species.

Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P (2008) Global trends in emerging infectious diseases. Nature 451:990–993.

Keating KA, Cherry S (2004) Use and Interpretation of Logistic Regression in Habitat-Selection Studies. The Journal of Wildlife Management 68:774–789.

Knight-Jones TJ, Rushston J (2013) The economic impacts of foot and mouth disease – What are they, how big are they and where do they occur? Preventive Veterinary Medicine 112:161–173.

Kock RA, Warnway HM, Rossiter PB, Lieberman G, Wambwa E, Okori J, Shiferaw FS, Mengeya TD (2006) Re-infection of wildlife populations with rinderpest virus on the periphery of the Somali ecosystem in East Africa. Preventive Veterinary Medicine 75:63–80.

Lancaster T, Imbens G (1996) Case-control studies with contaminated controls. Journal of Econometrics 71:145–160.

Liu C, Newell G, White M (2015) On the selection of thresholds for predicting species occurrence with presence-only data. Ecology and Evolution 6:337–348.

Logan NAY (1988) 1988. Bacillus species of medical and veterinary importance. Journal of Medical Microbiology 25:157–165.

Luis AD, O’Shea TJ, Hayman DTS, Wood JLN, Cunningham AA, Gilbert AT, Mills IN, Webb CT (2015) Network analysis of host–virus communities in bats and rodents reveals determinants of cross-species transmission. Ecology Letters 18:1153–1162.

Majewska AA, Huang T, Han B, Drake JM (2021) Predictors of zoonotic potential in helminths. Philosophical Transactions of the Royal Society B: Biological Sciences 376:20200356.

Miller RS, Farnsworth ML, Malmeberg JL (2013) Diseases at the livestock–wildlife interface: Status, challenges, and opportunities in the United States. Preventive Veterinary Medicine 110:119–132.

Morens DM, Holmes EC, Davis AS, Taubenberger JK (2011) Global rinderpest eradication: Lessons learned and why humans should celebrate too. The Journal of Infectious Diseases 204:502–505.

Morse SS (1995) Factors in the emergence of infectious diseases. Emerging Infectious Diseases 1:7–15.

Narrod C, Zinsstag J, Tiongco M (2012) A One Health Framework for Estimating the Economic Costs of Zoonotic Diseases on Society. EcoHealth 9:150–162.

Nenžin HK, Araújo MB (2011) Choice of threshold alters predictions of cross-species transmission. Proceedings of the Royal Society B: Biological Sciences 278:F73–F76.

Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL, Daszak P (2017) Host and viral traits predict zoonotic spillover from mammals. Nature 546:646–650.

Olsen SC (2010) Brucellosis in the United States: Role and significance of wildlife reservoirs. Vaccine 28:F73–F76.

Olson DM, Dinerstein E, Wikramanayake ED, Burgess ND, Powell GVN, Underwood EC, D’amico JA, Itoua I, Strand HE, Morrison JC, Loucks CJ, Allnutt TF, Ricketts TH, Kura Y, Lameurex JF, Wettengel WW, Hedao P, Kassem KR (2001) Ter-
restrial ecoregions of the world: A new map of life on earth. *BioScience* 51:993–938

Pacifi M, Santini L, Marco MD, Basiero D, Francuelli L, Marasini GG, Visconti P, Rondinini C (2013) Generation Length for Mammals. *Nature Conservation* 5:89–94

Paradies E, Schliew K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35:526–528

Pearson RG, Raxworthy CJ, Nakamura M, Townsend Peterson A (2006) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography* 34:102–117

Pebesma E (2018) Simple features for R: Standardized support for spatial vector data. *The R Journal* 10:439–446

Perry BD, Grace D, Sones K (2013) Current drivers and future directions of global livestock disease dynamics. *Proceedings of the National Academy of Sciences* 110:20871–20877

Prager KC, Mazet JAK, Dubovi EJ, Frank LG, Munson L, Wagner AP, Woodroffe R (2012) Rabies virus and canine distemper virus in wild and domestic carnivores in northern Kenya: Are domestic dogs the reservoir? *EcoHealth* 9:483–498

R Core Team (2018) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing

Rhyan JC, Nol P, Quance C, Gertonson A, Belfrage J, Harris L, Straka K, Robbe-Austerman S (2013) Transmission of Brucellosis from Elk to Cattle and Bison, Greater Yellowstone Area, USA, 2002–2012. *Emerging Infectious Diseases* 19:1992–1995

Rota CT, Millsbaugh JJ, Kesler DC, Lehman CP, Rumble MA, Jachowski CMB (2013) A re-evaluation of a case–control model with contaminated controls for resource selection studies. *Journal of Animal Ecology* 82:1165–1173

Royce JA, Chandler RB, Yackulic C, Nichols JD (2012) Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution* 3:343–354

Schatz AM, Park AW (2021) Host and parasite traits predict cross-species parasite acquisition by introduced mammals. *Proceedings of the Royal Society b: Biological Sciences* 288:20210341

Schumaker BA, Peck DE, Kauffman ME (2012) Brucellosis in the Greater Yellowstone area: disease management at the wildlife–livestock interface. *Human-Wildlife Interactions* 6:48–63

Shaw LP, Wang AD, Dylus D, Meier M, Pogacnik G, Dessimoz C, Balloux F (2020) The phylogenetic range of bacterial and viral pathogens of vertebrates. *Molecular Ecology* 29:3361–3379

Siembieda JL, Kock RA, McCracken TA, Newman SH (2011) The role of wildlife in transboundary animal diseases. *Animal Health Research Reviews* 12:95–111

Sowani H, Kulkarni M, Zinjarde S (2018) An insight into the ecology, diversity and adaptations of Gordonia species. *Critical Reviews in Microbiology* 44:393–413

Spalding MD, Fox HE, Allen GR, Davidson N, Feredia ZA, Finlayson M, Halpern BS, Jorge MA, Lombana A, Lourie SA, Martin KD, McManus E, Molnar J, Recchia CA, Robertson J (2007) Marine ecoregions of the world: A bioregionalization of coastal and shelf areas. *Science* 375:573–583

Stephens PR, Pappalardo P, Huang S, Byers JE, Farrell MJ, Gehman A, Ghai RR, Haas SE, Han B, Park AW, Schmidt JP, Altizer S, Ezenwa VO, Nunn CL (2017) Global Mammal Parasite Database version 2.0. *Ecology* 98:1476–1476

Upham NS, Esselstyn JA, Jetz W (2019) Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLOS Biology* 17:e3000494

Vinderola GN: Ouwehand ACSalminen SWright Avon (editors) (2019) *Lactic Acid Bacteria: Microbiological and Functional Aspects*, 5th ed., Boca Raton: CRC Press

Ward G, Hastie T, Barry S, Eliot J, Leathwick JR (2009) Presence-Only Data and the EM Algorithm. *Biometrics* 65:554–563

Wardeh M, Risley C, McIntyre MK, Setzkorn C, Baylis M (2015) Database of host-pathogen and related species interactions, and their global distribution. *Scientific Data* 2:150049

Wiothoelter AK, Beltrán-Alcudio D, Kock R, Mor SM (2015) Global trends in infectious diseases at the wildlife–livestock interface. *Proceedings of the National Academy of Sciences* 112:9662–9667

Wilman H, Belmaker J, Simpson J, de la Rosa C, Rivadeneira MM, Jetz W (2014) EltonTraits 1.0: Species-level foraging attributes of the world’s birds and mammals. *Ecology* 95:2027–2027

Winter DJ (2017) rentrez: an R package for the NCBI eUtils API. *The R Journal* 9:520–526

Woodroffe R, Donnelly CA, Cox DR, Gilks P, Jenkins HE, Johnston WT, Le Fevre AM, Bourne FJ, Cheeseman CL, Clifton-Hadley RS, Gettinby G, Hewinson RG, McNerney JP, Mitchell AP, Morrison WI, Watkins GH (2009) Bovine Tuberculosis in cattle and badgers in localized culling areas. *Journal of Wildlife Diseases* 45:128–143

World Organization for Animal Health. 2021. Terrestrial Animal Health Code.

Zhang T, Yu B (2005) Boosting with early stopping: Convergence and consistency. *The Annals of Statistics* 33:1538–1579