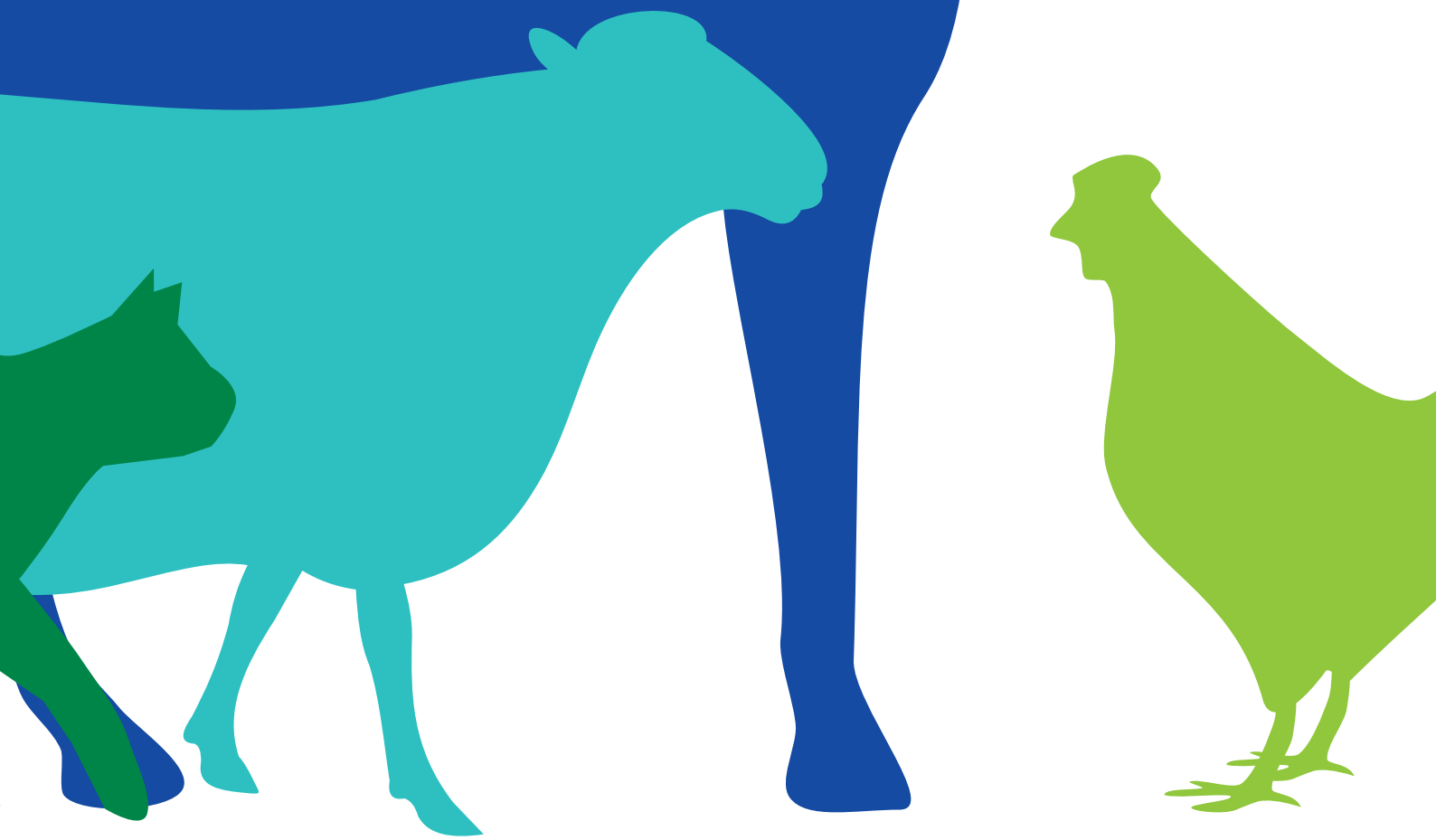
APPLICATIONS OF NOVEL ANALYTICAL METHODS IN EPIDEMIOLOGY

EDITED BY: Moh A. Alkhamis, Victoria J. Brookes and Kimberly VanderWaal
PUBLISHED IN: Frontiers in Veterinary Science



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88945-658-1
DOI 10.3389/978-2-88945-658-1

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

APPLICATIONS OF NOVEL ANALYTICAL METHODS IN EPIDEMIOLOGY

Topic Editors:
Moh A. Alkhamis, Kuwait University, Kuwait
Victoria J. Brookes, University of Sydney, Australia
Kimberly VanderWaal, University of Minnesota, United States

The repertoire of quantitative analytical techniques in disciplines such as ecology, decision science, and evolutionary biology has grown, in part enabled by the development and increased availability of computational resources. Integration of cutting-edge, quantitative tools into veterinary epidemiology that have been borrowed from such disciplines has offered opportunities to advance the study of disease dynamics in animal populations, to improve and guide decision-making related to disease prevention, control, or eradication. Furthermore, the need to explore new analytical methods for veterinary epidemiology has been driven by the increasing availability and complexity of animal disease data. The objective of this eBook is to contribute to current methods in epidemiology by 1) presenting and discussing novel analytical tools that help advance our understanding of epidemiology; and 2) demonstrating how inferences emerging from the application of novel analytical tools can be incorporated into decision-making related to animal health. The eBook constitutes a collection of articles that explore the applications of a variety of analytical methods such as machine learning, Bayesian risk assessment and an advanced form of social network analysis in the modern epidemiologic study of animal diseases.

Citation: Alkhamis, M. A., Brookes, V. J., VanderWaal, K., eds. (2018). Applications of Novel Analytical Methods in Epidemiology. Lausanne: Frontiers Media.
doi: 10.3389/978-2-88945-658-1

Table of Contents

frontiers in Veterinary Science

EDITORIAL
published: 04 October 2018
doi: 10.3389/fvets.2018.00243

Editorial: Applications of Novel Analytical Methods in Epidemiology

Moh A. Alkhamis[1,2]*, *Victoria J. Brookes*[3] *and Kimberly VanderWaal*[2]

[1] *Department of Epidemiology and Biostatistics, Faculty of Public Health, Health Sciences Center, Kuwait University, Kuwait City, Kuwait,* [2] *Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota, St. Paul, MN, United States,* [3] *Sydney School of Veterinary Science, Faculty of Science, University of Sydney, Sydney, NSW, Australia*

Keywords: spatial epidemiology, social network analysis, risk assessment, disease surveillance, decision making

Editorial on the Research Topic

Applications of Novel Analytical Methods in Epidemiology

Over the past decades, the repertoire of quantitative analytical techniques in disciplines such as ecology, decision science and evolutionary biology has grown, in part enabled by the development and increased availability of computational resources. Integration of cutting-edge, quantitative tools into veterinary epidemiology that have been borrowed from such disciplines has offered opportunities to advance the study of disease dynamics in animal populations, to improve and guide decision-making related to disease prevention, control, or eradication. Furthermore, the need to explore new analytical methods for veterinary epidemiology has been driven by the increasing availability and complexity of animal disease data ("big data"). The term "novel" in this research topic indicates methodological approaches that are currently infrequently or previously not used in epidemiology. The objective of this research topic is to contribute to current methods in epidemiology by: (1) presenting and discussing novel analytical tools that help advance our understanding of epidemiology; and (2) demonstrating how inferences emerging from the application of novel analytical tools can be incorporated into decision-making related to animal health.

It is worth noting that traditional analytical methods will continue to be essential tools for guiding and improving disease surveillance because they are computationally less demanding and therefore, more widely accessible. Magro et al. demonstrated the utility of multivariable logistic regression to risk mapping of sinonasal aspergillosis in dogs and in California, whilst Gautam et al. used a Cox proportional hazards model to identify risk factors for culling, sales, and deaths in New Zealand dairy goats.

Identification of suitable environmental and demographic factors for disease outbreaks is an essential component of risk-based surveillance. Escobar et al. offered two unique research articles in which he and his co-authors highlighted the potential of a novel algorithm in ecological niche modeling (ENM) for disease risk mapping of *Toxoplasma*, *Bartonella*, and *Heterosporis* spp.. They identified important environmental and demographic factors that shaped their predicted spatial distribution, as well as the relative contribution of these factors to this distribution. The two articles set the scene for further development of a powerful analytical approach for predicting disease distribution, thus contributing to the expanding field of spatial epidemiology.

Animal movement between premises has been identified as a critical factor for infectious disease introduction and spread. Valdes-Donoso et al. integrated methods in this research topic by combining data science techniques with social network analysis (SNA) to infer unobserved pig movements between farms. They selected relevant spatial and demographic factors and replicated the structure of a pig movement network from incomplete data by predicting the probability

OPEN ACCESS

Edited and reviewed by:
Ioannis Magouras,
City University of Hong Kong, Hong Kong

***Correspondence:**
Moh A. Alkhamis
m.alkhamis@hsc.edu.kw

Specialty section:
This article was submitted to Veterinary Epidemiology and Economics, a section of the journal Frontiers in Veterinary Science

Received: *09 August 2018*
Accepted: *14 September 2018*
Published: *04 October 2018*

Citation:
Alkhamis MA, Brookes VJ and VanderWaal K (2018) Editorial: Applications of Novel Analytical Methods in Epidemiology. Front. Vet. Sci. 5:243. doi: 10.3389/fvets.2018.00243

of unobserved movement between sites using supervised machine learning. Their inferences approximated actual unobserved movements between sites; hence this predictive method could guide control strategies between and within geographical regions.

McCabe and Nunn also utilized SNA to offer a unique approach to modeling transmission networks of infectious diseases. They proposed a novel measure for describing a network—"effective network size"—which accounts for the heterogeneity of contacts between the nodes of a given network. They applied this metric to disease outbreak networks simulated from different disease spread models as well as empirical disease networks described in the literature. They found that their metric is highly associated with many traditional network metrics, and hence might provide additional epidemiological insights when using SNA to guide disease intervention.

Assessment of the temporal dynamics of disease outbreaks is another analytical pillar needed to formulate risked-based surveillance systems. Arruda et al. proposed the effective time-dependent reproductive number as a complementary measure of disease spread to use alongside incidence and frequently used spatial analyses such as cluster detection. They demonstrated the utility of computing sequential effective reproductive numbers from case-series data to assess the endemicity of porcine reproductive and respiratory syndrome across regions in the United States. Furthermore, they showed how such a measure could be used to guide and improve intervention strategies to control endemic swine diseases.

Transforming common analytical methods in epidemiology into a Bayesian statistical framework has recently become more common due to the substantial growth in computational resources. Bayesian analytical methods require fewer assumptions about the data than frequentist methods, provide methods to account for uncertainties in data, and can accommodate more complex biological parameters for estimating posterior risk measures of infectious diseases. Gamado et al. developed a toolkit that can be used for risk assessment of epidemic models that are based on discrete-state, continuous-time Markov and semi-Markov processes, using data-augmentation Markov Chain Monte Carlo techniques within a Bayesian framework. They demonstrated how their toolkit could be reliably used to assess risks from potential disease introductions, which subsequently can be used to support and guide prompt disease intervention efforts.

Perception of risk and quantifying preferences among animal producers, health workers, and other stakeholders is another important factor in guiding the implementation of disease intervention efforts. Opinions and perceptions impact policymaking, and thus, advanced analytical methods that can decipher the complexity and diversity of stakeholder preferences are required. Van den Borne et al. borrowed and adapted conjoint analysis (CA) from decision-science. This method has had limited use in veterinary science, with applications previously focused on disease prioritization. Van den Borne et al. used a computer-based adaptive choice-based conjoint analysis to elicit respondents' preferences for design characteristics of a new udder health and antimicrobial usage improvement program in Switzerland. They demonstrated the novelty and advantages of their approach in guiding decision-makers to both administer current animal health programs and develop new programs based on the independent opinions of veterinarians and animal producers.

In this research topic, a selected array of novel analytical methods from a variety of scientific disciplines was demonstrated, with the goal of complementing and advancing the field of epidemiology. Our intention is that this research topic further motivates epidemiologists to seek and develop analytical methods to overcome current analytical limitations; application of novel and interdisciplinary methods new to the field of veterinary epidemiology has the potential to expand the horizons of veterinary epidemiological research.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Veterinary Science

ORIGINAL RESEARCH
published: 03 July 2017
doi: 10.3389/fvets.2017.00104

Spatial Patterns and Impacts of Environmental and Climatic Factors on Canine Sinonasal Aspergillosis in Northern California

*Monise Magro[1], Jane Sykes[2], Polina Vishkautsan[3] and Beatriz Martínez-López[1]**

[1] Center for Animal Disease Modeling and Surveillance (CADMS), Department of Medicine and Epidemiology, School of Veterinary Medicine, University of California Davis, Davis, CA, United States, [2] William R. Pritchard Veterinary Medical Teaching Hospital (VMTH), Department of Medicine and Epidemiology, School of Veterinary Medicine, University of California Davis, Davis, CA, United States, [3] Internal Medicine, Veterinary Specialty Center of Tucson, Tucson, AZ, United States

Sinonasal aspergillosis (SNA) causes chronic nasal discharge in dogs and has a worldwide distribution, although most reports of SNA in North America originate from the western USA. SNA is mainly caused by *Aspergillus fumigatus*, a ubiquitous saprophytic filamentous fungus. Infection is thought to follow inhalation of spores. SNA is a disease of the nasal cavity and/or sinuses with variable degrees of local invasion and destruction. While some host factors appear to predispose to SNA (such as belonging to a dolichocephalic breed), environmental risk factors have been scarcely studied. Because *A. fumigatus* is also the main cause of invasive aspergillosis in humans, unraveling the distribution and the environmental and climatic risk factors for this agent in dogs would be of great benefit for public health studies, advancing understanding of both distribution and risk factors in humans. In this study, we reviewed electronic medical records of 250 dogs diagnosed with SNA between 1990 and 2014 at the University of California Davis Veterinary Medical Teaching Hospital (VMTH). A 145-mile radius catchment area around the VMTH was selected. Data were aggregated by zip code and incorporated into a multivariate logistic regression model. The logistic regression model was compared to an autologistic regression model to evaluate the effect of spatial autocorrelation. Traffic density, active composting sites, and environmental and climatic factors related with wind and temperature were significantly associated with increase in disease occurrence in dogs. Results provide valuable information about the risk factors and spatial distribution of SNA in dogs in Northern California. Our ultimate goal is to utilize the results to investigate risk-based interventions, promote awareness, and serve as a model for further studies of aspergillosis in humans.

Keywords: *Aspergillus*, risk map, fungal infection, dogs, epidemiology, public health, spores, spatial analysis

OPEN ACCESS

Edited by:
Timothée Vergne, Institut de Recherche pour le Développement (IRD), France

Reviewed by:
Kim Stevens, Royal Veterinary College, United Kingdom
Dan Gerard O'Neill, Royal Veterinary College, United Kingdom

****Correspondence:***
Beatriz Martínez-López
beamartinezlopez@ucdavis.edu

Specialty section:
This article was submitted to Veterinary Epidemiology and Economics, a section of the journal Frontiers in Veterinary Science

Received: *16 February 2017*
Accepted: *15 June 2017*
Published: *03 July 2017*

Citation:
Magro M, Sykes J, Vishkautsan P and Martínez-López B (2017) Spatial Patterns and Impacts of Environmental and Climatic Factors on Canine Sinonasal Aspergillosis in Northern California. Front. Vet. Sci. 4:104. doi: 10.3389/fvets.2017.00104

INTRODUCTION

Aspergillosis has gained clinical significance in both the veterinary medicine and public health fields. It affects a wide variety of hosts such as dogs, cats, birds, and humans causing morbidity and mortality worldwide (1–11). Canine sinonasal aspergillosis (SNA) is a common presentation of aspergillosis in dogs (1, 3, 12, 13) and also one of the most frequent causes of chronic sinonasal disease in

dogs (6, 14–16). SNA is mainly caused by *Aspergillus fumigatus*, and less frequently by other species, such as *Aspergillus niger*, *Aspergillus flavus* (16, 17), and *Penicillium* spp. (16). *A. fumigatus* is a ubiquitous filamentous saprophytic fungus (18), and its conidia (spores) are found in soil, air, water, decaying vegetation, and dust. It is an airborne pathogen (19–24), thus transmission occurs through inhalation of conidia from the environment. SNA generally involves the nasal cavity and/or frontal sinuses (17, 25) of otherwise apparently systemically healthy dogs (17), causing destruction of nasal turbinates (9). Extensive damage to the nasal bones, cribriform plate, and orbit can occur in severe cases (26, 27). Clinical signs include mucoid to mucopurulent nasal discharge, sneezing, depigmentation and/or ulceration of the nares, nasal pain, and epistaxis (1, 26). Ocular discharge occurs as a result of nasolacrimal duct destruction and orbit invasion in advanced cases (1). Neurologic signs may occur if the cribriform plate is affected (26, 27). An important differential diagnosis for SNA in dogs is nasal neoplasia (14). Diagnosis is costly and requires a mixture of invasive and non-invasive diagnostic tests such as advanced imaging, serology, rhinoscopy for identification of fungal plaques, and cytology and/or histopathology (12, 27–31). Treatment is often challenging and involves a combination of tissue debridement, topical, and systemic antifungal therapy (9, 26).

We focused our study on SNA because it is a frequent form of aspergillosis seen in dogs (1, 3, 12, 13), and because the etiology and epidemiology of SNA are distinct from the other types of aspergillosis in dogs and may have implications for human disease. Nasal aspergillosis accounts for 7–34% of dogs with nasal disorders and is the second most common cause of chronic nasal discharge (6, 14, 32, 33). A study in humans shows that worldwide approximately 2.5% (4.8 million people) of adults who have asthma also have allergic bronchopulmonary aspergillosis (34). Invasive aspergillosis in humans generally affects immunocompromised people and is one of the most common fungal infections on organ transplant recipients (35). Studies have shown that *A. fumigatus* are present and propagate under a variety of environmental and climatic conditions such as water, soil, decaying vegetation, and compost. *A. fumigatus* present in the air and water for example have been associated with aspergillosis in humans (11, 19–23, 36, 37). *A. fumigatus* spores can also disperse easily in the air when compared to other types of fungi. Moisture and inappropriate temperature have been associated with the disease in birds (10).

We hypothesized that environmental and climatic factors significantly favor pathogens' growth, spread, and disease transmission. Studies of geographical distribution with identification of potential high-risk areas and quantification of the influence of environmental and climatic factors on canine SNA are lacking.

Our specific study aims were as follows: (1) to analyze the spatial patterns of canine SNA in Northern California from electronic medical records of the University of California Davis William R. Pritchard Veterinary Medical Teaching Hospital (VMTH) to aid identification of high-risk areas for disease occurrence; (2) to assess the association between environmental/climatic risk factors and disease occurrence in Northern California using a multivariate logistic model. Although SNA does not pose zoonotic potential, dogs may be important sentinels for human exposure as dogs and humans cohabit the same environment. Therefore, this study has potential implications for both animal and public health with the ultimate goal of early detection, prevention, and mitigation of disease.

MATERIALS AND METHODS

Data Description

The VMTH database was searched from January 1st of 1990 to December 31st of 2014 to identify cases of canine SNA. The keyword *aspergi** was used to capture all canine aspergillosis cases. Dogs were then identified that had a clinical diagnosis of SNA as determined by the attending clinician. Cases were included if clinical findings were consistent with SNA based on thorough review of medical records by a board-certified internal medicine specialist (Polina Vishkautsan) including consistent medical history, physical examination findings, imaging findings (computed tomography or magnetic resonance imaging), presence of fungal plaques on rhinoscopy, positive *Aspergillus* gel immunodiffusion serology, culture of *Aspergillus* from nasal biopsy specimens and identification of fungal hyphae and conidiophores on histopathology. Not all dogs had all diagnostic tests done, but all dogs diagnosed with SNA had advanced imaging (usually a computed tomography scan) and rhinoscopy, and one or more of either visualization of fungal plaques on rhinoscopy, fungal structures on biopsy, or growth of *A. fumigatus* from a biopsy of nasal tissue. Both primary care and referral cases were included, although the majority was referred because of the need for special expertise for diagnosis and treatment of the condition.

In addition, a reference dog population was determined, which included all dogs seen at the VMTH during the 25 years of the study period residing in the zip codes within a catchment area of 145 miles around the UC Davis VMTH. This area was empirically established and assumed to represent a reasonable distance that an owner could drive to seek care. Data of cases and the dog reference hospital population were aggregated at the zip code level (i.e., unit of analysis) (Supplementary Figures S1–S3 are provided for reference).

Data collected from dogs with SNA were residential address (zip code), date of SNA diagnosis, whether diagnosis was obtained by the referring veterinarian or at the VMTH, and fungal culture results when available.

The environmental and climatic factors evaluated in this study for potential association with canine SNA occurrence are described in **Table 1**. We included as predictors 38 environmental and climatic factors that have a likely biological plausibility and/or have been previously described either to favor fungus growth or cause alteration or damage of respiratory track (11, 19, 22, 36–41).

Spatial Analysis

Spatial analysis was performed using *ArcGIS version 10.2.2 (ESRI®, 2015)* and *R Studio* (version 0.98.1091). ZIP Code Tabulation Areas (ZCTAs) shapefile for the state of California was obtained from U.S. Census Bureau (42) and joined using the *table join* function in *ArcMap* with all the information collected

TABLE 1 | Environmental and climatic factors assessed for association with canine sinonasal aspergillosis occurrence over a 25-year study period.

Variable	Description (unit)	Source
Water	Open water and perennial ice/snow areas (%)	Geospatial Data Getaway USDA website National Land Cover Database 2011 by state (California) https://gdg.sc.egov.usda.gov/
Developed areas	Open space, low, medium, and high intensity areas of development (%)	
Barren	Barren land areas (%)	
Forest	Areas of deciduous, evergreen, and mixed forests (%)	
Shrub land	Areas dominated by shrubs (%)	
Grassland/herbaceous	Areas dominated by graminoid or herbaceous vegetation (%)	
Agriculture	Pasture/hay (areas of grasses, legumes, or grass–legume mixtures planted for livestock grazing or the production of seed or hay crops) and cultivated crops (areas used for the production of annual crops) (%)	
Wetlands	Woody and emergent herbaceous wetlands (%)	
Soil moisture	Mean soil moisture (average values from 1990 to 2014) (%)	Cal-Adapt website http://cal-adapt.org/data/download/
Soil moisture difference	Difference in soil moisture (average values from 2014 minus average values from 1990) (%)	
Relative humidity	Relative humidity (average values from 1990 to 2014) (%)	
Clay	Mean soil percent clay (%)	Data Basin website SSURGO percent soil clay, sand, silt, and pH for California, USA https://databasin.org/datasets/
Sand	Mean soil percent sand (%)	
Silt	Mean soil percent silt (%)	
Soil percent organic matter	Mean percentage soil organic matter (%)	
Soil pH	Mean soil pH (pH)	
Maximum soil pH	Maximum soil pH (pH)	
Minimum soil pH	Minimum soil pH (pH)	
Total land fire acre	Total acres of land at time of fire control from 1990 to 2014 (acre)	Federal Fire Occurrence website http://wildfire.cr.usgs.gov/firehistory/data.html
Fire history	Counts of fire episodes from 1990 to 2014 (count)	
Active composting sites	Active composting facilities (food, green, wood, biosolid, agricultural waste) (count)	CalRecycle website at http://www.calrecycle.ca.gov/
Wind	Mean wind speed (average values from 1990 to 2014) (m/s)	Cal-Adapt website http://cal-adapt.org/data/download/
Wind difference	Difference in wind speed (average wind speed values from 2014 minus average values from 1990) (m/s)	
Temperature	Mean temperature (average temperature values from 1990 to 2014) (°C)	
Temperature difference	Difference in temperature (average values from 2014 minus average values from 1990) (°C)	
Maximum temperature	Maximum temperature (maximum values from average maximum values from 1981 to 2010) (°F)	Prism Climate Group Oregon State University http://prism.oregonstate.edu/normals/
Minimum temperature	Minimum temperature (minimum values from average minimum values from 1981 to 2010) (°F)	
Precipitation	Mean precipitation (average values from 1981 to 2010) (″)	
Precipitation difference	Difference in precipitation (average values of 2014 minus average values of 1990) (mm)	Cal-Adapt website at http://cal-adapt.org/data/download/
Ozone	Ozone [portion of the daily maximum 8-h ozone concentration over the federal 8-h standard (0.075 ppm), averaged over 3 years (2007–2009)] (ppm)	Office of Environmental Health Hazard Assessment website http://oehha.ca.gov/ej/ces11.html
PM 2.5	Fine particulate matter annual mean concentrations (average of quarterly means), over 3 years (2007–2009) ($\mu g/m^3$)	
Diesel PM	Diesel particulate matter emissions from on-road and non-road sources for a 2010 July day (kg/day)	
Pesticide use	Total pounds of selected active pesticide ingredients used in production-agriculture per square mile ($lb/mile^2$)	
Toxic release	Total toxicity-weighted pounds of chemicals released to air or water from all facilities within the ZIP code or within 1 km of ZIP code (toxicity-weighted pounds)	
Traffic density	Sum of traffic volume (vehicle/kilometers per hour) by total road length (km) within 150 m of the ZIP code boundary [vehicle kilometers per hour/total road length (km)]	
Cleanup sites	Sum of weighted sites per ZIP codes (weighted sites)	

at the ZCTA level. The *select by location function* was used to select the ZCTA within a distance of 145 miles. We calculated the regionwide disease hospital incidence rate (i.e., sum of dog cases/sum of reference dog population) and used it to calculate the expected cases per ZCTA (sum reference dog population per ZCTA × regionwide disease incidence rate). Then, the standardized incidence ratio (SIR) was calculated by dividing the observed number of cases per ZCTA by the expected number of cases per ZCTA, and its mean 1.34 used as the cut point to create the binomial response variable for the multivariate logistic regression analysis (if the SIR value was larger than the SIR mean, it was assigned one, and if it was smaller

than the mean, it was assigned 0). For the predictors, the mean value of each variable for each ZCTA was obtained using the *extract function* (*raster* package) in *R*. The *tabulate area* tool in *ArcMap* was used to obtain the percentage of land use type for each ZCTA. The geographic coordinate system used for all the shapefiles and rasters were North American 1983, the Datum North American 1983, and the projection NAD 1983 UTM Zone 11N (linear unit in meters).

Statistical Analysis

All continuous variables were transformed and evaluated both in the standardized ($Z = X - \mu/\sigma$; i.e., the observed value minus the overall mean divided by the SD) and binomial (using the median as cut point: observed value > median = 1; 0 otherwise) forms. Univariate logistic regression analysis was first applied to each variable and only those that had a *p*-value smaller than 0.25 were considered for inclusion in the full model. Then, a multivariate logistic regression analysis was used to identify the factors significantly associated with the disease occurrence. The model was specified as follows:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1\chi_1 + \beta_2\chi_2 \ldots + \beta_k\chi_k$$

where p_i is the probability of SNA occurrence at ZCTA level, α is the intercept and, β are the regression coefficients for each predictor χ.

Model selection was conducted using forward selection (*step function* from *stats package* from *R*) and using the *Akaike information criterion* to determine the best fitting model. Model diagnostics were verified by checking deviance residuals and the variance inflation factor to evaluate multicollinearity problems. The predictive ability of the model was conducted analyzing the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. All possible two-way interaction terms and potential confounders were also evaluated. Fitted values obtained were then used to create a canine SNA risk map using the *spplot function* (*sp* package) in *R*. The *spplot function* was also used to create a deviance residuals map obtained from the logistic regression model to analyze the distribution and presence of extreme residual values. Because the neglect of spatial autocorrelation can result in biased regression coefficient estimates for the SNA occurrence, an autologistic regression analysis was run and compared with the ordinary multivariate logistic regression analysis results and check for any effect on spatial autocorrelation. The autologistic regression model introduced by Besag (43) is an extension of an ordinary logistic regression model and corrects for spatial dependence in the observations by incorporating an autocovariate variable. In this case, the autocovariate variable was generated from a neighborhood-weighting matrix based on the probabilities of SNA occurrence obtained from the final multivariate logistic regression model. Among the several spatial weights tested (i.e., rook and queen contiguity, inverse-distance) to create the spatial weights matrix, queen contiguity was considered the most appropriate for this neighborhood structure and was chosen. Queen contiguity defines ZCTAs neighbors when they either share a border or vertex. The autocovariate formula was specified as follows:

$$\text{Autocov}_i = \frac{\sum_{j=1}^{k_i} w_{ij}\hat{p}_j}{\sum_{j=1}^{k_i} w_{ij}}$$

The autocovariate variable (*Auto cov*$_i$) is a weighted average of the probabilities of the geographic units (ZCTAs) amongst a set of k_i neighbors of the geographic unit i, w_{ij} is the spatial weight between the geographic unit i and j, $\hat{p}_j$ is the probability estimated by the logistic regression model. All analyses were conducted in *R* Language using *R Studio* (44, 45).

RESULTS

A total of 250 cases (out of 286 cases retrieved from ZCTA in the entire state of California) and 190,894 dogs (reference dog population) were part of the study catchment area ZCTAs and were considered for the analysis. The mean number of cases per ZCTA in the study area was 0.33 (SD = 0.71; min = 0; max = 5). The overall cumulative incidence (incidence proportion) for the study area was 1.31 SNA cases per 1,000 dogs for the entire studied period. The mean SIR was 1.34 (SD = 4.96; min = 0; max = 54.54). The study area contained 768 ZCTAs (43% of the total number of ZCTAs in California), from which 140 ZCTAs had an SIR > mean.

The final model contained six variables (five main effects and one interaction term), three of which were environmental [i.e., traffic density (OR = 1.7; $p = 0.0311$), active composting sites (OR = 1.2; $p = 0.0299$), agriculture (OR = 0.67; $p = 0.00345$)] and three were climatic [i.e., wind difference (OR = 1.3; $p = 0.0621$), temperature difference (OR = 0.69; $p = 0.0104$), interaction between wind and temperature differences (OR = 1.6; $p = 0.0134$)] (**Table 2**). The predictive ability of the model based on the AUC value of the ROC curve was 73.1%. The autologistic regression model showed similar results than the final logistic regression model, with similar regression coefficients, deviance residuals and AUC and, for that reason, the simplest logistic regression model was selected.

TABLE 2 | Association between environmental and climatic variables and canine sinonasal aspergillosis occurrence in California obtained with a multivariate logistic regression model.

Variable	Odds ratio (95% confidence interval)	*p*-Value (Wald test)
Traffic density[a] low (≤503.2)	Reference = 1.0	
High (>503.2)	1.7 (1.1, 2.8)	0.0311
Wind difference (2014–1990)[b]	1.3 (0.99, 1.7)	0.0621
Active composting sites[b]	1.2 (1.0, 1.4)	0.0299
Temperature difference (2014–1990)[b]	0.69 (0.52, 0.92)	0.0104
Agriculture[b]	0.67 (0.52, 0.88)	0.00345
Wind difference × temperature difference[b]	1.6 (1.1, 2.3)	0.0134

[a]*Binomial form.*
[b]*Standardized form.*

The spatial distribution of SNA and significant predictors included in the final model are shown in **Figure 1**. The effect of temperature difference interacting with wind difference at different levels in the occurrence of SNA is shown in **Figure 2**. This figure shows that when there was high wind difference, the probability of SNA was high, and the protective effect of temperature difference on SNA occurrence was minimized.

The resultant risk map for canine SNA shows that high-risk areas were concentrated close to the San Francisco Bay area, other coastal areas, and also in northern and central CA (**Figure 3**). The top 10 counties containing ZCTAs that exhibited the highest risk for canine SNA occurrence were Stanislaus, Santa Clara, Sonoma, Santa Cruz, Napa, Contra Costa, Placer, Monterey, Marin, and Sacramento. Residuals from the final logistic model were relatively randomly distributed within the study area (**Figure 4**).

DISCUSSION

To the best of authors' knowledge, this is one of the first studies to provide a canine SNA risk map and to analyze the association between environmental and climatic factors associated with canine SNA in California. Results provide valuable insights on climatic, environmental, and anthropogenic risk factors associated with the disease.

We compared results from both models (logistic and autologistic) by evaluating improvements in the intercept, regression coefficients, deviance residuals, and goodness of the fit (AUC). Since the autologistic model did not significantly change the results over the logistic model, we assumed that spatial autocorrelation was not playing an important role in this particular case. Moreover, residuals map for the final logistic model did not show values too far from 0, and the highest values were fairly dispersed. The majority of the regions with the highest deviance residual values (residual map values >2) are located mainly in the sierra and do not overlap with the highest risk areas (risk map values >0.4), located mainly in the coastal areas. Overall, the deviance residual values are low, indicating a good fit of the model, and are generally dispersed, which, in addition to the results of the autologistic regression model suggest no concerning spatial autocorrelation. Only values deviance residuals >2 may be indicative of lack of fit for those few counties in the sierra suggesting that our model seems to predict better coastal cases than cases in the sierra. Thus, the non-spatial logistic regression model was chosen for simplicity. Traffic density was the main factor associated with SNA (OR = 1.7), suggesting that pollution may predispose dogs to this disease. It is possible that the pathogenicity of *A. fumigatus* in dogs varies in the presence of different pollutants, or that pollution damages mucosal defense mechanisms. One recent study of the effect of urban air pollution on allergenicity of *A. fumigatus* spores in the laboratory found increased allergenicity in polluted urban environments during the first 12 h of exposure (46).

Wind and temperature were also significantly associated with SNA occurrence. Both wind and temperature differences between the beginning and the end of the study period (i.e., 1990 and 2014,

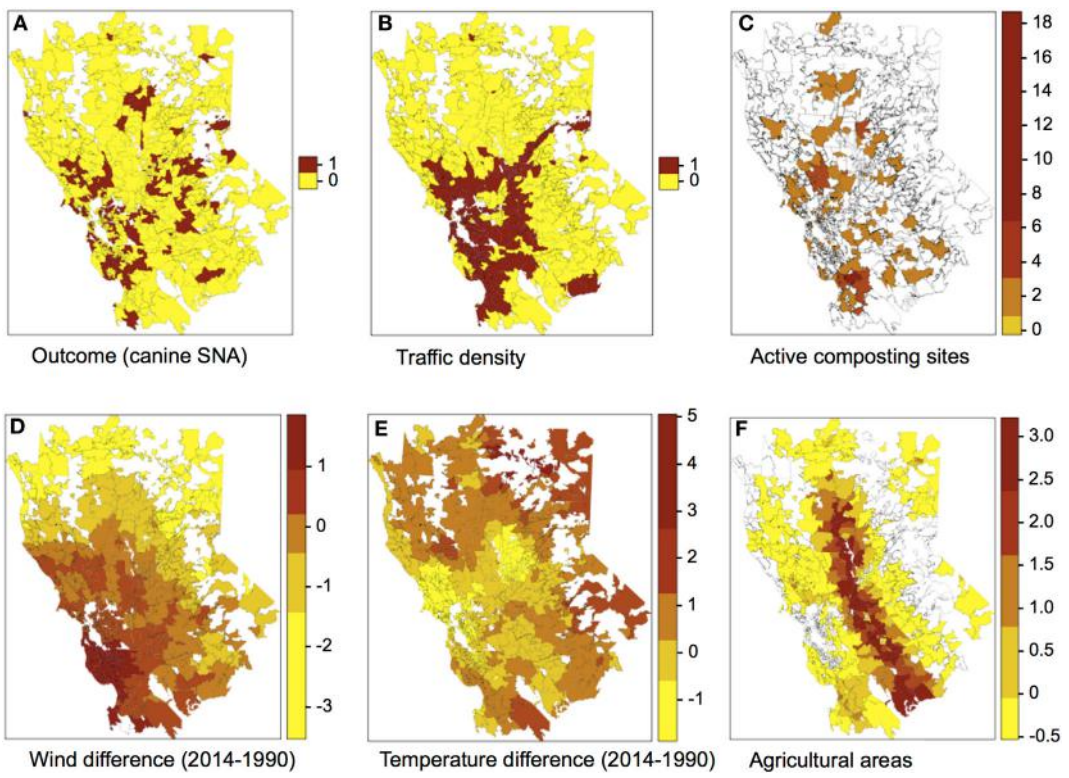


FIGURE 1 | Spatial distribution of the outcome **(A)** and predictors **(B–F)** included in the final multivariate logistic regression model. Categories for the colors of plots **(C–F)** were obtained using the Jenks algorithm (i.e., Natural breaks).

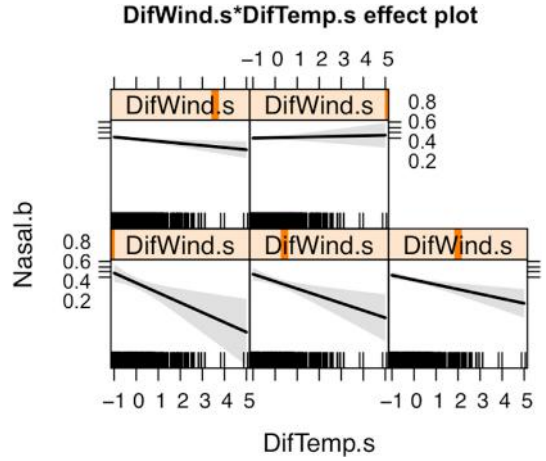


FIGURE 2 | Effect of the interaction between wind and temperature differences on the occurrence of canine sinonasal aspergillosis in Northern California. Vertical orange lines indicate the value of the wind difference considered for each plot.

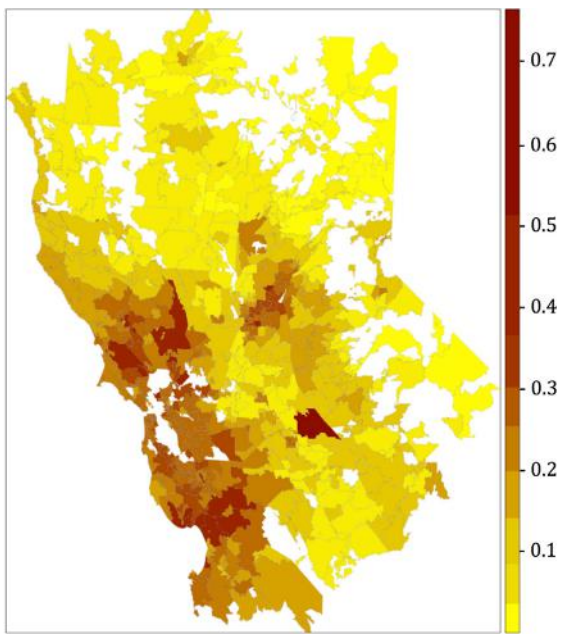


FIGURE 3 | Canine sinonasal aspergillosis risk map obtained after plotting the fitted values of the final multivariate logistic regression model. Categories for the colors were obtained using the Jenks algorithm (i.e., natural breaks).

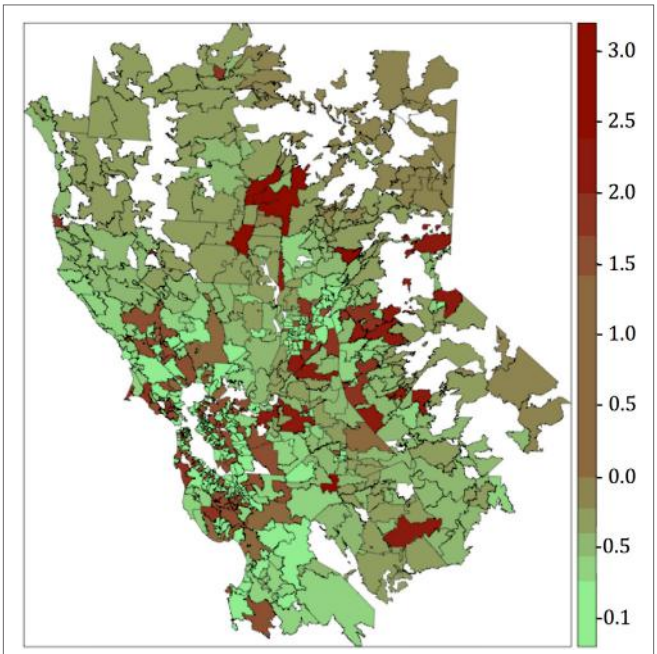


FIGURE 4 | Residuals map obtained after plotting the deviance residuals of the final multivariate logistic regression model. Categories for the colors were obtained using the Jenks algorithm (i.e., natural breaks).

respectively) were used as proxies of areas where climatic conditions are more drastically changing. Both wind and temperature measurements were higher in 2014 compared to 1990. The protective effect that temperature difference had on SNA occurrence was attenuated by wind (**Figure 2**). Wind may facilitate dispersion of spores and thus increase the risk of SNA in dogs. Areas with increased wind in the last years are mostly in urban and coastal settings, which tend to experience landscape modifications such as construction, leading to soil disturbance. Urban and coastal areas also appear to have higher humidity and traffic density compared to other areas. Similarly, this study suggests that dogs living in areas experiencing lower temperature differences are those at highest risk of SNA, although again this effect was influenced by wind. In general, these areas have lower elevations and higher temperatures (**Figure 1**). Active composting sites (for food, green waste, wood, biosolid, and agricultural waste) were significantly associated with an increased risk for SNA. *A. fumigatus* thrives in composting facilities due to the presence of organic matter (22, 37, 39). The thermotolerant nature of *A. fumigatus* allows this pathogen to grow in compost and crops at elevated temperatures (optimal growth at 37°C and maximum growth at 52°C) (47, 48).

Agricultural land use by ZCTA reduced the risk of canine SNA. This was somewhat unexpected because the fungus is frequently found in environments rich in nutrients and humidity and where the soil is usually disturbed (as it is usually the case of agricultural areas). However, one possible explanation for the negative association is fungicide application to crops. Unfortunately, detailed information about fungicide use in California was not available. Further studies should be conducted to clarify this effect.

Risk maps provide useful information for clinicians when evaluating animals with chronic nasal discharge that reside in high-risk areas and may help to accelerate accurate diagnosis, and in turn enable more prompt treatment and better prognosis. Furthermore, because *A. fumigatus* causes disease in both dogs and humans, and because dogs and humans share the same environments, our findings may also have relevance for identification of high-risk areas for human aspergillosis. Human invasive aspergillosis mostly affects immunocompromised patients and can be life-threatening (4, 38, 49, 50). Because

A. fumigatus spores can be found in air, dust, and water in hospital environments (11, 20, 36, 38, 41), hospitals located in high-risk areas might focus on improved ventilation systems to reduce risk of infection.

One of the limitations of this study was the availability of data from a single hospital, which can introduce selection bias. Because canine SNA requires specialized diagnostic equipment and treatments not commonly available in general veterinary practice, a diagnosis of canine SNA is often made at the VMTH or cases are referred in for treatment. In order to further reduce selection bias, the 145-mile catchment area around the VMTH was empirically delineated to represent areas of hospital caseload influence. This radius was assumed to be a reasonable distance for a pet owner to drive to seek treatment at the VMTH. Thirdly, the dog reference hospital population from the 25-year study period was taken into account for the SIR calculation; therefore, we adjusted the expected number of SNA cases observed by ZCTA with the population of dogs visiting VMTH from that ZCTA. Therefore, the use of SIR provides a better estimate of the SNA situation per region than the use of the number of cases *per se*. Lack of access to patient travel history was also a limitation of this study since canine SNA is a chronic disease, and some dogs may have become infected at a different residence prior to diagnosis. Similarly, a long study period was used because of the relatively low number of cases/year, and the relatively stable SIR over that 25-year period (mean: 1.34; SD: ±4.96). The variability of 4.96 is not very high, but it is important to consider. We tried to apply Bayesian zero-inflated Binomial and zero-inflated Poisson models that allow for overdispersion in INLA to model these data, but models did not converge. Authors acknowledge the limitation of merging cases over time and unfortunately it was not possible to include variables per year to minimize ecological fallacy due to the small sample size. However, changes over time were verified and variables such as difference in temperature, in wind, and in precipitation were included and used to account for changes over time and hopefully mitigate the impact of time on the study results.

Future studies that include data from more veterinary hospitals over a wider geographic range may provide more detail in regards to the spatiotemporal patterns of this disease and the varying impact that environmental and climatic conditions may have had in different regions over time.

Despite the limitations, this study provides preliminary insights into the spatial distribution and risk factors contributing to SNA occurrence in dogs in Northern California. These results may be useful to increase awareness and guide diagnosis and risk mitigation strategies in high-risk areas, ultimately opening new doors for further investigation of *A. fumigatus* infections not only in dogs, but also in humans.

AUTHOR CONTRIBUTIONS

All authors contributed to the design of the study. PV, JS, and MM contributed to the data gathering, cleaning, interpretation, and validation. MM and BM-L wrote the R codes, conducted the data analyses, and drafted the manuscript. All the authors contributed to the critical discussion of results and reviewed and edited the final manuscript.

ACKNOWLEDGMENTS

The authors thank the Center for Animal Disease Modeling and Surveillance (CADMS) and Master of Preventive Veterinary Medicine (MPVM) team for their support, Dr. Jaber Belkhiria for his assistance on software used for this research project, Dr. Philip Kass for kindly reviewing and providing valuable feedback for this paper, and all VMTH clinicians who participated in the care for the canine sinonasal aspergillosis cases in the 25 years studied. This work was part of an MPVM thesis defended in May 2016. Funds were provided by CADMS and the UC Davis Open Access Fund.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/fvets.2017.00104/full#supplementary-material.

REFERENCES

1. Day MJ. Canine sino-nasal aspergillosis: parallels with human disease. *Med Mycol* (2009) 47(Suppl 1):S315–23. doi:10.1080/13693780802056038
2. Tell L. Aspergillosis in mammals and birds: impact on veterinary medicine. *Med Mycol* (2005) 43(Suppl 1):S71–3. doi:10.1080/13693780400020089
3. Barrs V, Halliday C, Martin P, Wilson B, Krockenberger M, Gunew M, et al. Sinonasal and sino-orbital aspergillosis in 23 cats: aetiology, clinicopathological features and treatment outcomes. *Vet J* (2012) 191(1):58–64. doi:10.1016/j.tvjl.2011.02.009
4. Holding KJ, Dworkin MS, Wan P-CT, Hanson DL, Klevens RM, Jones JL, et al. Aspergillosis among people infected with human immunodeficiency virus: incidence and survival. *Clin Infect Dis* (2000) 31(5):1253–7. doi:10.1086/317452
5. Barrs VR, van Doorn TM, Houbraken J, Kidd SE, Martin P, Pinheiro MD, et al. *Aspergillus felis* sp. nov., an emerging agent of invasive aspergillosis in humans, cats, and dogs. *PLoS One* (2013) 8(6):e64871. doi:10.1371/journal.pone.0064871
6. Meler E, Dunn M, Lecuyer M. A retrospective study of canine persistent nasal disease: 80 cases (1998–2003). *Can Vet J* (2008) 49(1):71–6.
7. Akan M, Haziroğlu R, İlhan Z, Sareyyüpoğlu B, Tunca R. A case of aspergillosis in a broiler breeder flock. *Avian Dis* (2002) 46(2):497–501. doi:10.1637/0005-2086(2002)046[0497:ACOAIA]2.0.CO;2
8. Richard J, Cutlip R, Thurston J, Songer J. Response of turkey poults to aerosolized spores of *Aspergillus fumigatus* and aflatoxigenic and nonaflatoxigenic strains of *Aspergillus flavus*. *Avian Dis* (1981) 25:53–67. doi:10.2307/1589826
9. Zonderland J-L, Störk CK, Saunders JH, Hamaide AJ, Balligand MH, Clercx CM. Intranasal infusion of enilconazole for treatment of sinonasal aspergillosis in dogs. *J Am Vet Med Assoc* (2002) 221(10):1421–5. doi:10.2460/javma.2002.221.1421
10. Seyedmousavi S, Guillot J, Arné P, de Hoog GS, Mouton JW, Melchers WJ, et al. *Aspergillus* and aspergilloses in wild and domestic animals: a global health concern with parallels to human disease. *Med Mycol* (2015) 53(8):765–97. doi:10.1093/mmy/myv067
11. Schmitt H, Blevins A, Sobeck K, Armstrong D. *Aspergillus* species from hospital air and from patients. *Mycoses* (1989) 33(11–12):539–41.

12. Peeters D, Day M, Clercx C. An immunohistochemical study of canine nasal aspergillosis. *J Comp Pathol* (2005) 132(4):283–8. doi:10.1016/j.jcpa.2004.11.002
13. Benitah N. Canine nasal aspergillosis. *Clin Tech Small Anim Pract* (2006) 21(2):82–8. doi:10.1053/j.ctsap.2005.12.015
14. Tasker S, Knottenbelt C, Munro E, Stonehewer J, Simpson J, Mackin A. Aetiology and diagnosis of persistent nasal disease in the dog: a retrospective study of 42 cases. *J Small Anim Pract* (1999) 40(10):473–8. doi:10.1111/j.1748-5827.1999.tb02998.x
15. Windsor RC, Johnson LR, Herrgesell EJ, De Cock HE. Idiopathic lymphoplasmacytic rhinitis in dogs: 37 cases (1997–2002). *J Am Vet Med Assoc* (2004) 224(12):1952–3. doi:10.2460/javma.2004.224.1952
16. Mathews KG, Sharp NJ. Canine nasal aspergillosis-penicilliosis. *Infect Dis Dog Cat* (2006) 3:613–20.
17. Sharp N, Harvey C, Sullivan M. Canine nasal aspergillosis and penicilliosis. *Compend Contin Educ Pract Vet* (1991) 13:41–7.
18. Latge JP. *Aspergillus fumigatus* and aspergillosis. *Clin Microbiol Rev* (1999) 12(2):310–50.
19. Mullins J, Harvey R, Seaton A. Sources and incidence of airborne *Aspergillus fumigatus* (Fres). *Clin Exp Allergy* (1976) 6(3):209–17. doi:10.1111/j.1365-2222.1976.tb01899.x
20. Warris A, Klaassen CH, Meis JF, de Ruiter MT, de Valk HA, Abrahamsen TG, et al. Molecular epidemiology of *Aspergillus fumigatus* isolates recovered from water, air, and patients shows two clusters of genetically distinct strains. *J Clin Microbiol* (2003) 41(9):4101–6. doi:10.1128/JCM.41.9.4101-4106.2003
21. Curtis L, Cali S, Conroy L, Baker K, Ou C-H, Hershow R, et al. *Aspergillus* surveillance project at a large tertiary-care hospital. *J Hosp Infect* (2005) 59(3):188–96. doi:10.1016/j.jhin.2004.05.017
22. Haines J. *Aspergillus* in compost: straw man or fatal flaw? *Biocycle* (1995) 36(4):32–5.
23. Streifel A, Lauer J, Vesley D, Juni B, Rhame F. *Aspergillus fumigatus* and other thermotolerant fungi generated by hospital building demolition. *Appl Environ Microbiol* (1983) 46(2):375–8.
24. Ren P, Jankun T, Belanger K, Bracken M, Leaderer B. The relation between fungal propagules in indoor air and home characteristics. *Allergy* (2001) 56(5):419–24. doi:10.1034/j.1398-9995.2001.056005419.x
25. Mortellaro CM, Franca PD, Caretta G. *Aspergillus fumigatus*, the causative agent of infection of the frontal sinuses and nasal chambers of the dog. *Mycoses* (1989) 32(7):327–35. doi:10.1111/j.1439-0507.1989.tb02253.x
26. Sharp N, Harvey C, O'brien J. Treatment of canine nasal aspergillosis/penicilliosis with fluconazole (UK-49,858). *J Small Anim Pract* (1991) 32(10):513–6. doi:10.1111/j.1748-5827.1991.tb00868.x
27. Saunders JH, Zonderland JL, Clercx C, Gielen I, Snaps FR, Sullivan M, et al. Computed tomographic findings in 35 dogs with nasal aspergillosis. *Vet Radiol Ultrasound* (2002) 43(1):5–9. doi:10.1111/j.1740-8261.2002.tb00434.x
28. Saunders JH, Clercx C, Snaps FR, Sullivan M, Duchateau L, van Bree HJ, et al. Radiographic, magnetic resonance imaging, computed tomographic, and rhinoscopic features of nasal aspergillosis in dogs. *J Am Vet Med Assoc* (2004) 225(11):1703–12. doi:10.2460/javma.2004.225.1703
29. Saunders JH, Van Bree H. Comparison of radiography and computed tomography for the diagnosis of canine nasal aspergillosis. *Vet Radiol Ultrasound* (2003) 44(4):414–9. doi:10.1111/j.1740-8261.2003.tb00478.x
30. Pomrantz JS, Johnson LR, Nelson RW, Wisner ER. Comparison of serologic evaluation via agar gel immunodiffusion and fungal culture of tissue for diagnosis of nasal aspergillosis in dogs. *J Am Vet Med Assoc* (2007) 230(9):1319–23. doi:10.2460/javma.230.9.1319
31. De Lorenzi D, Bonfanti U, Masserdotti C, Caldin M, Furlanello T. Diagnosis of canine nasal aspergillosis by cytological examination: a comparison of four different collection techniques. *J Small Anim Pract* (2006) 47(6):316–9. doi:10.1111/j.1748-5827.2006.00153.x
32. Lane J, Warnock D. The diagnosis of *Aspergillus fumigatus* infection of the nasal chambers of the dog with particular reference to the value of the double diffusion test. *J Small Anim Pract* (1977) 18(3):169–77. doi:10.1111/j.1748-5827.1977.tb05867.x
33. Sullivan M. Rhinoscopy: a diagnostic aid? *J Small Anim Pract* (1987) 28(9):839–44. doi:10.1111/j.1748-5827.1987.tb01350.x
34. Denning DW, Pleuvry A, Cole DC. Global burden of allergic bronchopulmonary aspergillosis with asthma and its complication chronic pulmonary aspergillosis in adults. *Med Mycol* (2013) 51(4):361–70. doi:10.3109/13693786.2012.738312
35. Pappas PG, Alexander BD, Andes DR, Hadley S, Kauffman CA, Freifeld A, et al. Invasive fungal infections among organ transplant recipients: results of the Transplant-Associated Infection Surveillance Network (TRANSNET). *Clin Infect Dis* (2010) 50(8):1101–11. doi:10.1086/651262
36. Warris A, Voss A, Abrahamsen TG, Verweij PE. Contamination of hospital water with *Aspergillus fumigatus* and other molds. *Clin Infect Dis* (2002) 34(8):1059–60. doi:10.1086/339754
37. Ryckeboer J, Mergaert J, Vaes K, Klammer S, De Clercq D, Coosemans J, et al. A survey of bacteria and fungi occurring during composting and self-heating processes. *Ann Microbiol* (2003) 53(4):349–410.
38. Lutz BD, Jin J, Rinaldi MG, Wickes BL, Huycke MM. Outbreak of invasive *Aspergillus* infection in surgical patients, associated with a contaminated air-handling system. *Clin Infect Dis* (2003) 37(6):786–93. doi:10.1086/377537
39. Millner P, Marsh P, Snowden R, Parr J. Occurrence of *Aspergillus fumigatus* during composting of sewage sludge. *Appl Environ Microbiol* (1977) 34(6):765–72.
40. Millner PD, Bassett DA, Marsh PB. Dispersal of *Aspergillus fumigatus* from sewage sludge compost piles subjected to mechanical agitation in open air. *Appl Environ Microbiol* (1980) 39(5):1000–9.
41. Goodley J, Clayton Y, Hay R. Environmental sampling for aspergilli during building construction on a hospital site. *J Hosp Infect* (1994) 26(1):27–35. doi:10.1016/0195-6701(94)90076-0
42. US Census Bureau. *Census 2010, 5-Digit Zip Code Tabulation Areas (ZCTAs), Cartographic Boundary Files*. (2014). Available from: https://www.census.gov/geo/maps-data/data/cbf/cbf_zcta.html
43. Besag J. Spatial interaction and the statistical analysis of lattice systems. *J R Stat Soc Ser B (Methodol)* (1974) 36:192–236.
44. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (2016).
45. RStudio Team. *RStudio: Integrated Development for R. RStudio, Inc*. Boston, MA (2015). Available from: http://www.rstudio.org/
46. Lang-Yona N, Shuster-Meiseles T, Mazar Y, Yarden O, Rudich Y. Impact of urban air pollution on the allergenicity of *Aspergillus fumigatus* conidia: outdoor exposure study supported by laboratory experiments. *Sci Total Environ* (2016) 541:365–71. doi:10.1016/j.scitotenv.2015.09.058
47. Beffa T, Staib F, Lott Fischer J, Lyon P, Gumowski P, Marfenina O, et al. Mycological control and surveillance of biological waste and compost. *Med Mycol* (1998) 36(1):137–45.
48. Shehu K, Bello M. Effect of environmental factors on the growth of *Aspergillus* species associated with stored millet grains in Sokoto. *Niger J Basic Appl Sci* (2011) 19(2):218–23.
49. Patterson J, Peters J, Calhoon J, Levine S, Anzueto A, Al-Abdely H, et al. Investigation and control of aspergillosis and other filamentous fungal infections in solid organ transplant recipients. *Transpl Infect Dis* (2000) 2(1):22–8. doi:10.1034/j.1399-3062.2000.020105.x
50. Kontoyiannis D, Bodey G. Invasive aspergillosis in 2002: an update. *Eur J Clin Microbiol Infect Dis* (2002) 21(3):161–72. doi:10.1007/s10096-002-0699-z

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Veterinary Science

ORIGINAL RESEARCH
published: 10 November 2017
doi: 10.3389/fvets.2017.00191

Risk Factors for Culling, Sales and Deaths in New Zealand Dairy Goat Herds, 2000–2009

Milan Gautam[1], Mark A. Stevenson[2], Nicolas Lopez-Villalobos[1] and Victoria McLean[3]*

[1] Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand, [2] Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, VIC, Australia, [3] Dairy Goat Co-Operative, Hamilton, New Zealand

The aim of this study was to identify risk factors for culling, sales and deaths in intensively managed dairy goat herds in New Zealand. A data set provided by the New Zealand Dairy Goat Cooperative (n = 13,197 does) was analyzed using a Cox proportional hazard model. The outcome of interest was length of productive life (LPL), defined as the number of days from the date of second kidding to the date of removal from the herd or the date on which follow-up was terminated, whichever occurred first. Milk solids yield in the first lactation (MSL1) as a predictor of LPL was parameterized in the model as a penalized spline term. To account for MSL1 violating the proportional hazards assumption of the Cox model, LPL was divided into two intervals: T1 (less than or equal to 730 days from the date of second kidding) and T2 (greater than 730 days from the date of second kidding). MSL1 was then included in the model as a time-dependent covariate. A frailty term was included in the model to account for unmeasured, herd-level effects on LPL. During T1, the daily hazard of removal for does that produced 80 kg milk solids in the first lactation was 0.84 (95% CI 0.58–1.23) times the daily hazard of removal for does that produced 30 kg milk solids in the first lactation. During T2, the daily hazard of removal for does that produced 80 kg milk solids in the first lactation was 1.44 (95% CI 0.79–2.65) times the daily hazard of removal for does that produced 30 kg milk solids in the first lactation. We conclude that involuntary losses may be avoided if high MSL1 yielding does are preferentially managed from 2 years beyond the date of second kidding.

Keywords: epidemiology, dairy goats, length of productive life, survival analysis, Cox proportional hazards regression

OPEN ACCESS

Edited by:
Moh A. Alkhamis, Kuwait Institute for Scientific Research, Kuwait

Reviewed by:
Ane Nødtvedt, Norwegian University of Life Sciences, Norway
Catalina Picasso, University of Minnesota, United States

****Correspondence:***
Milan Gautam
m.gautam@massey.ac.nz

Specialty section:
This article was submitted to Veterinary Epidemiology and Economics, a section of the journal Frontiers in Veterinary Science

Received: *15 May 2017*
Accepted: *23 October 2017*
Published: *10 November 2017*

Citation:
Gautam M, Stevenson MA, Lopez-Villalobos N and McLean V (2017) Risk Factors for Culling, Sales and Deaths in New Zealand Dairy Goat Herds, 2000–2009. Front. Vet. Sci. 4:191. doi: 10.3389/fvets.2017.00191

INTRODUCTION

In farmed animal production systems (e.g., dairy, beef cattle, pig, and dairy goat farms) a long, productive life of individual production units is an essential prerequisite for economic efficiency (1). In dairy systems, longevity is defined as the interval between delivery of the first offspring and the date of removal from the herd (2). Increasing the longevity of dairy animals is desirable because it means that the cost of rearing replacements is amortized over a longer period of income production. Since longevity is a desirable quality in production animals (3), it is important to have an understanding of factors influencing the same. Very little work has been done in this area of the dairy goat industry, and an understanding of risk factors for culling, sales and deaths in dairy goats is limited.

In New Zealand, the number of dairy goat herds is small relative to the number of dairy cow herds, and a key industry focus is on the production of infant formula (4). Typically, does are housed indoors in open sided free stall barns and are fed fresh-cut pasture. Approximately two-thirds of the commercial dairy goat farms are concentrated in the Waikato region, in the upper North Island. Purebred and crossbred Saanens are the predominant breeds, but other breeds such as Toggenburgs and Alpines are common (4). At the time of writing, there were 69 herds registered with the New Zealand Dairy Goat Cooperative (NZDGC), a farmer-owned cooperative, each with around 700 milking does per herd, on average.

A better understanding of the various risk factors for removal can be used to enhance longevity in dairy animals. With this knowledge, it is possible to identify characteristics that can serve as early indicators of culling and, depending upon how strong the effect of a particular risk factor on removal is, it is possible to plan in advance the best time to remove an animal from the herd when it is still profitable to do so or at least incur minimal loss. Survival analysis is a commonly used technique to quantify longevity in domestic animals (5, 6). Using this technique, the association between risk factors and culling can be examined in relation to their effect on the length of productive life (LPL) instead of simply describing the relationship in terms of risk (5). In survival analysis, a quantity termed "hazard" is modeled instead of longevity itself (7). Hazard represents the instantaneous probability that an animal is removed at a given time, given that it is still present up to that time. Since it is the hazard that is modeled and not longevity, it is possible to use data from animals that have not yet been removed from the herd (as censored observations) as well as those that have been removed (7).

Although a number of studies have been carried out to identify risk factors for removal in dairy cows (5, 8, 9), the number of similar studies in dairy goats is limited (1, 10) and, to the best of our knowledge, none have been conducted in a New Zealand context. To address this knowledge gap, the aim of this study was to identify factors that influence the risk of removal in commercial dairy goat herds in New Zealand (11). This knowledge will allow managers of dairy goat herds take a more planned approach to culling: either to remove does at higher risk of removal at a time when it is economic to do so, or to preferentially manage profitable animals if it is known that they are at greater risk of removal compared to their herd mates.

MATERIALS AND METHODS

Study Population and Data Collection

The data for this study were obtained from the NZDGC. Since the total number of dairy goat herds in New Zealand is relatively small, we assumed the dairy goat herds affiliated with NZDGC provided an accurate reflection of commercial dairy goat farming in New Zealand. Although the complete data set was comprised of records for a total of 48,699 animals (including those with birth dates as early as August 1983 and production records up to December 2009), only those born on or after 1st January 2000 were used in the analyses presented in this paper. This restriction was applied because a large proportion of animals born prior to 1st January 2000 had missing observations, particularly those related to total lactation length and milk, fat, and protein yields.

Several exclusion criteria were applied to the NZDGC data (**Figure 1**). Bucks were excluded from the analyses. A doe had to complete her first lactation and then kid for a second time to be included in the data set so that the correct temporal sequence between first lactation milk solids yield (MSL1) and LPL was ensured. Finally, records were screened and limited to does having a first lactation length between 0 and 305 days and/or a first lactation total milk solids yield of less than or equal to 1,800 kg. Lactations of greater than 305 days and total lactation yields of more than 1,800 kg milk solids were deemed implausible. Finally, does for which the first lactation fat and protein yields

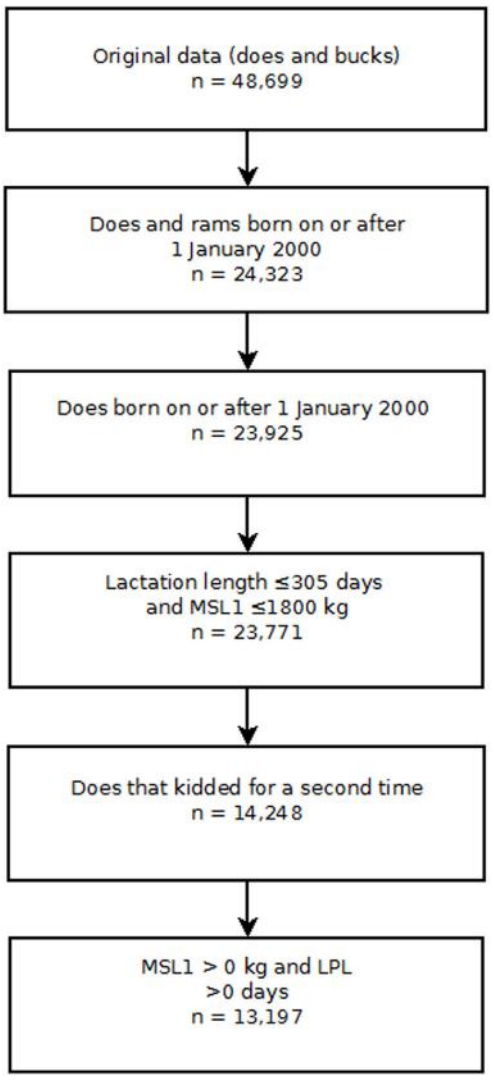


FIGURE 1 | Risk factors for culling, sales and deaths in New Zealand dairy goat herds, 2000–2009. Flow chart showing the exclusion criteria used to select individual doe records for analysis in this study. Key: MSL1 first lactation milk solids yield (kilograms); LPL, length of productive life.

were recorded as 0 were excluded from the analyses. Does were followed until 31st December 2009 or the date on which they were removed from the herd, whichever occurred first.

Herds registered with the NZDGC record data for individual animals including the date of birth, the unique animal identifier, breed, parity date(s), and the date and reasons for removal from the herd (if applicable). Herd managers record details of individual animals into paper diaries or, more rarely in the case of dairy goats, into dedicated herd health software. This information is then sent to the national milk recording authority, Livestock Improvement Corporation (LIC) who merge these details with test day milk yields measured at roughly 60-day intervals throughout the lactation. Animal biographical and production data recorded in the central database of LIC are then transferred to NZDGC in digital format. This information is used by NZDGC for genetic evaluation of individual animal (12). Estimated breeding values for milk, fat, protein, and milk solids (fat and protein) obtained from genetic evaluations are reported to the NZDGC and each herd manager receives an individual report with the genetic evaluation of his/her animals.

The outcome of interest in this study was LPL, defined as the difference in time (days) between the date of second kidding and the date of removal from the herd. In the context of this study, we use the term "removal" to refer to animals that leave the herd as either culled animals, sales, or deaths. For does that were still in the herd at the termination of the study (censored observations), LPL was quantified as the time between the date of second kidding and 31st December 2009.

Model Building

Selection of Explanatory Variables

The total yields of milk protein and milk fat from each animal in the first lactation were added to create a single variable called first lactation milk solids yield (MSL1).

Based on the reported breed composition of the sire and dam the breed of each animal was recorded in 16th for the following breeds: Saanen, Toggenburg, Nubian, Alpine, and "unknown." From these fractions (the total of which sum to one), the proportion of each breed was calculated. For instance, the breed composition of a doe with pedigree values 8, 4, 0, 0, 4 for Saanen, Toggenburg, Nubian, Alpine, and unknown (respectively) would be 50% Saanen, 25% Toggenburg, 0% Nubian, 0% Alpine, and 25% unknown. Given the several possible combinations of crossbreds, it was decided that the percentage of each breed would be forced into the model as a series of continuous variables to avoid any ambiguity created by breed defined as a categorical variable. The recorded parentage details for all does were not available. Where parentage details were not available, breed fractions were estimated by the herd manager.

Bivariate Analyses

Since all the explanatory variables in our study were continuously distributed, they were categorized into quartiles. The Kaplan–Meier technique (13) was then used to quantify LPL of does within each quartile. The log rank statistic was used to test the homogeneity of survivorship between quartile groups. Those explanatory variables that showed an association with LPL (that is, a difference in the Kaplan–Meier survival curves that was significant at $P < 0.20$) were selected for inclusion in the multivariate analyses.

Multivariable Analyses

Factors influencing LPL were quantified using a Cox proportional hazard model (14). Here, the hazard of removal at time t can be expressed as:

$$H(t,x) = h_0(t)\exp^{\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}}. \quad (1)$$

Equation 1 shows the hazard of an event at time t is the product of $h_0(t)$ and $\exp^{\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}}$. The first of these quantities, $h_0(t)$, is called the baseline hazard function and includes a time component t, representing how the hazard of removal changes as a function of time. The remaining quantity $exp^{\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}}$ is the exponential of the linear sum of a series of k explanatory variables. This quantity represents how the baseline hazard function is modified in response to a given set of explanatory variables. In contrast to the baseline hazard function, the set of explanatory variables does not involve a time component (15).

A key assumption of the Cox model is that of proportionality of hazards. According to this assumption, the effect of an explanatory variable on the outcome of interest does not change over time, i.e., the hazards for each level of an explanatory variable must be proportional at all times. In situations where this assumption is violated, modifications such as stratified analyses or inclusion of time-dependent covariates are necessary (16).

Model development was carried out using the contributed survival package (17) implemented in R version 3.3.3 (18). To start, a saturated Cox model was run including all explanatory variables identified as influencing LPL at the bivariate level. Explanatory variables that were not statistically significant were removed from the model one at a time, beginning with the least significant, until the estimated regression coefficients for all explanatory variables retained were significant at an alpha level of less than 0.05. Explanatory variables that were excluded at the initial screening stage were tested for inclusion in the final model and were retained in the model if their inclusion changed any of the estimated regression coefficients by more than 20%. Biologically plausible two-way interactions were between explanatory variables were assessed.

Checking the Scale of Continuous Covariates

A key assumption in including MSL1 into the model as a continuous variable was that the relationship between MSL1 and log hazard was linear. To test this assumption, MSL1 was categorized into quartiles and the regression coefficient for each quartile plotted as function of the midpoint of each quartile group. Since the line connecting the four midpoints was not linear, we concluded that MSL1 was not linear in its log hazard. Based on these findings, a penalized spline term was used to account for the non-linear association between MSL1 and LPL.

Testing the Proportional Hazard Assumption

To verify that the proportional hazards assumption of the Cox model was valid a plot of the scaled Schoenfeld residuals from the model as a function of time was constructed. In a model

where the proportional hazards assumption holds the Schoenfeld residuals should be scattered around 0. We calculated the Pearson product–moment correlation between the scaled Schoenfeld residuals and time and the hypothesis of no correlation between the two variables was assessed using a χ^2 test statistic. From these analyses, we concluded that MSL1 violated the proportional hazards assumption. To account for non-proportionality of hazards, we divided LPL into two intervals: less than or equal to 730 days (referred to as T1 in the remainder of this paper) and greater than 730 days (T2). The decision to use 730 days was semi-arbitrary and was selected because, being equivalent to 2 years, it approximated median LPL in this population. This division allowed us to quantify the effect of MSL1 separately for each period [less than or equal to 730 days (T1) and greater than 730 days (T2)]. The technique of dividing the time component into intervals to investigate the time-dependent effect of covariates is called a piecewise Cox proportional hazards model or a step function proportional hazards model.

Final Model

In addition to the terms to allow for the interaction between time and penalized MSL1, our final model included herd as a random effect, otherwise known as a frailty term.

RESULTS

The final data set was comprised of 23,771 does with a birth date greater than or equal to 1st January 2000. Of this group, 14,248 does completed their first lactation and kidded for the second time. Further screening of the production data and removal of implausible records reduced the final data set to comprised 13,197 does from 38 herds (**Figure 1**). Of this group, 5,386 animals were removed during the follow-up period and the remaining 7,811 animals that were recorded as being alive in the herd on 31st December 2009 were treated as censored observations. Descriptive statistics of the study population are presented in **Table 1**.

Inclusion of terms for breed in the Cox proportional hazards model was not statistically significant. Biologically plausible two-way interactions were tested and none were significant at an alpha level of 0.05.

As shown in **Table 2**, the interaction between MSL1 and time was significant for T1, but was not statistically significant for T2. During T1, the hazard of removal for does that produced 80 kg milk solids in the first lactation was 0.84 (95% CI 0.58–1.23) times the daily hazard of removal for does that produced 30 kg milk solids in the first lactation (**Figure 2**). During T2 (730 days after the date of second kidding), high producing MSL1 does had a higher daily hazard of removal compared to average producing herd mates: a doe producing 80 kg milk solids in the first lactation had 1.44 (95% CI 0.79–2.65) times the daily hazard of removal compared with does that produced 30 kg milk solids in the first lactation (**Figure 3**). These results show that relatively high levels of MSL1 production had no strong association with daily hazard of removal during the early phase of productive life, however, as LPL progressed, does with higher MSL1 yields were at greater risk of removal.

DISCUSSION

We used a piece-wise Cox proportional hazards model, to quantify the effect of MSL1 on LPL in dairy goats that completed their first lactation and kidded a second time. To the best of our knowledge, this is the first study of its kind to evaluate the effect of a time-dependent covariate on longevity in dairy goats.

Although the results presented in this study are based on data which were not originally collected for the purpose of this study, consent to use and analyze the data was obtained from NZDGC before the start of the study and results were presented to NZDGC stakeholders. A possible limitation of our study was selection bias in that the herds used for these analyses were those that participated in herd testing programs and were, therefore, likely to be a more intensively managed subset of dairy goat herds compared with the general population of New Zealand dairy goat herds. A second limitation was that we could not investigate the effect of specific diseases or disease categories on longevity. There were two reasons for this: (1) we had no reassurance that disease case definitions were used consistently over time and across each of the herds that took part in the study; and (2) does were removed for a wide range of reasons resulting in relatively low numbers of animals in each category. When studying factors influencing LPL in production animals, it is desirable to identify

TABLE 1 | Risk factors for culling, sales and deaths in New Zealand dairy goat herds, 2000–2009.

Outcome	*n*	Mean	SD	Median	Q1; Q3
L1 fat yield (kg)	13,197	16	8	16	10; 21
L1 protein yield (kg)	13,197	14	7	14	9; 18
L1 milk solids (kg)	13,197	30	15	29	19; 40
Age at first kidding (days)	13,197	580	421	390	369; 669
LPL (days)	5,386[a]	763	547	663	327; 1,084
Age at removal (days)	5,386[a]	1,644	596	1,500	1,142; 2,026
Number of lactations	5,386[a]	3	1.4	3	2; 4

Descriptive statistics of first lactation production outcomes, age at first kidding, LPL, age at removal and total number of lactations.
L1, lactation 1; LPL, length of productive life; Q1, first quantile; Q3, third quantile.
[a]*Uncensored does only.*

TABLE 2 | Risk factors for culling, sales and deaths in New Zealand dairy goat herds, 2000–2009.

Variable	Coefficient (SE)	Chi square	df	P
MSL1 × T1				
Linear	−0.0033 (0.0014)	5.31	1.0	0.021
Non-linear	–	1.68	3.0	0.650
MSL1 × T2				
Linear	0.0014 (0.0016)	1.00	1.0	0.360
Non-linear		3.05	3.0	0.030
Herd-level random effect	–	2,358.74	13.60	0.000

Regression coefficients of factors influencing risk of culling in dairy goats from the final piecewise Cox model.
MSL1, first lactation milk solids yield (kilogram); T1, 0–730 days from the date of second kidding; T2, greater than 730 days from the date of second kidding.

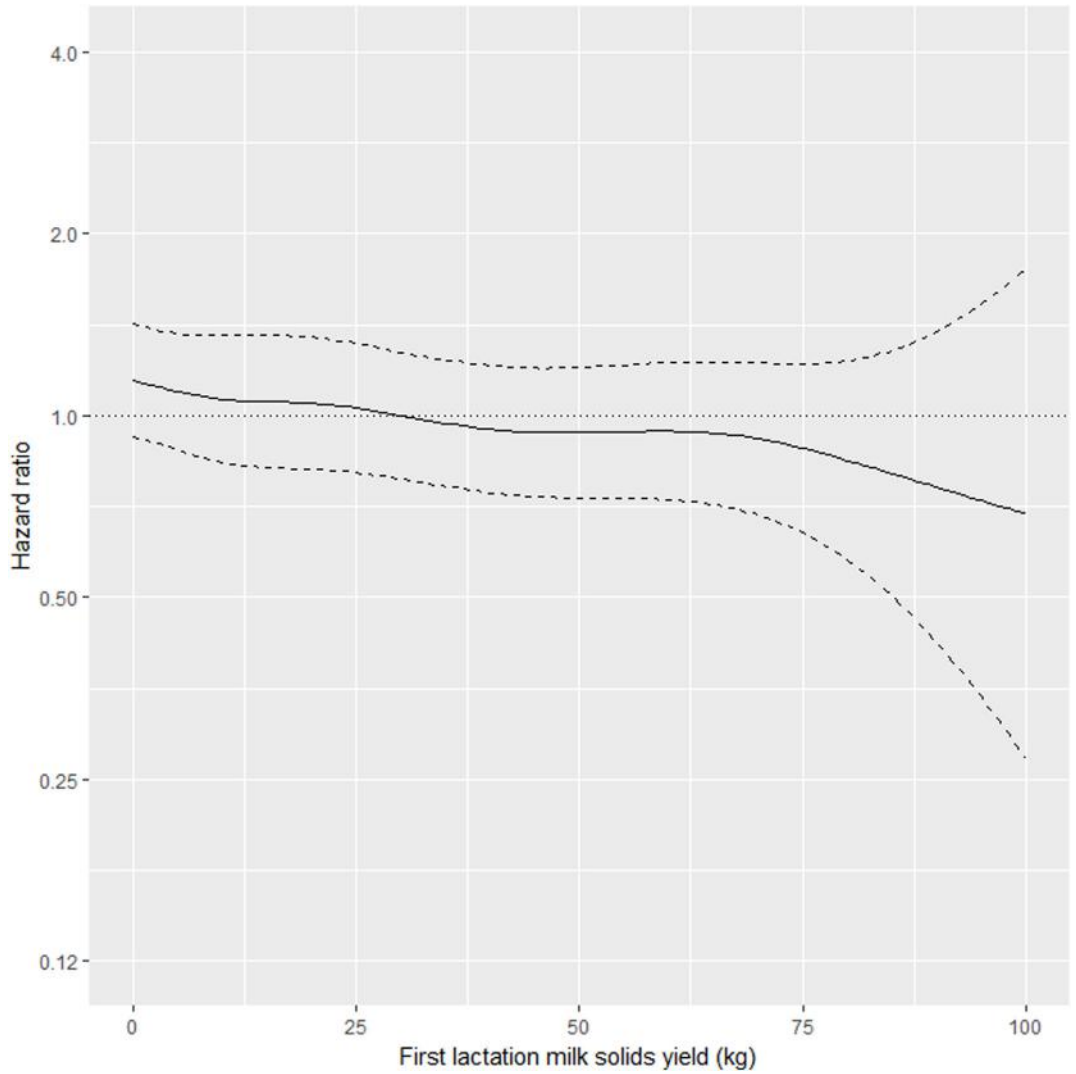


FIGURE 2 | Risk factors for culling, sales and deaths in New Zealand dairy goat herds, 2000–2009. Line plot showing, for the interval 0–730 days from the date of second kidding, the hazard ratio for removal as a function of first lactation milk solids yield (based on the model presented in **Table 2**). The dashed lines represent 95% confidence intervals around the point estimates of the hazard ratio. In the above plot, the reference category was a doe producing 30 kg milk solids in the first lactation. A doe producing 80 kg milk solids in the first lactation had 0.84 (95% CI 0.58–1.23) times the daily hazard of removal compared with a doe that produced 30 kg milk solids in the first lactation.

risk factors for specific removal reasons (e.g., reproductive failure, udder health, lameness) as opposed to considering all removals as a single group. Failure to do so is likely to mask some of the more subtle influences on longevity. As a prerequisite for being able to examine specific reasons for removal, it is necessary that removal reasons are recorded accurately and consistently across herds and over time.

Our results show that in the first 2 years after the date of second kidding, there was an inverse association between MSL1 yields and the daily hazard of removal (**Figure 2**). Does with higher MSL1 yields had lower daily hazards of removal compared with average producing herd mates. This trend reversed beyond 2 years from the date of second kidding (**Figure 3**) with high MSL1 yields having a higher daily hazard of removal compared with average producing herd mates. We believe these results provide useful information for the management of dairy goat herds. As high producers get older, herd managers need to take special steps to ensure that this group of animals is managed in such a way to minimize the impact of factors that could influence removal risk. For example, a herd manager might elect to run his/her high MSL1 producers as a separate mob and to provide preferential feeding, housing, and milking management.

A search of the literature did not identify any previous studies that investigated the association between first lactation milk solids yield and longevity in dairy goats. Even in dairy cattle, the number of studies that have examined the association between first lactation milk yield and longevity is limited (19–22). It has been shown that mean daily yield of milk in the first lactation of

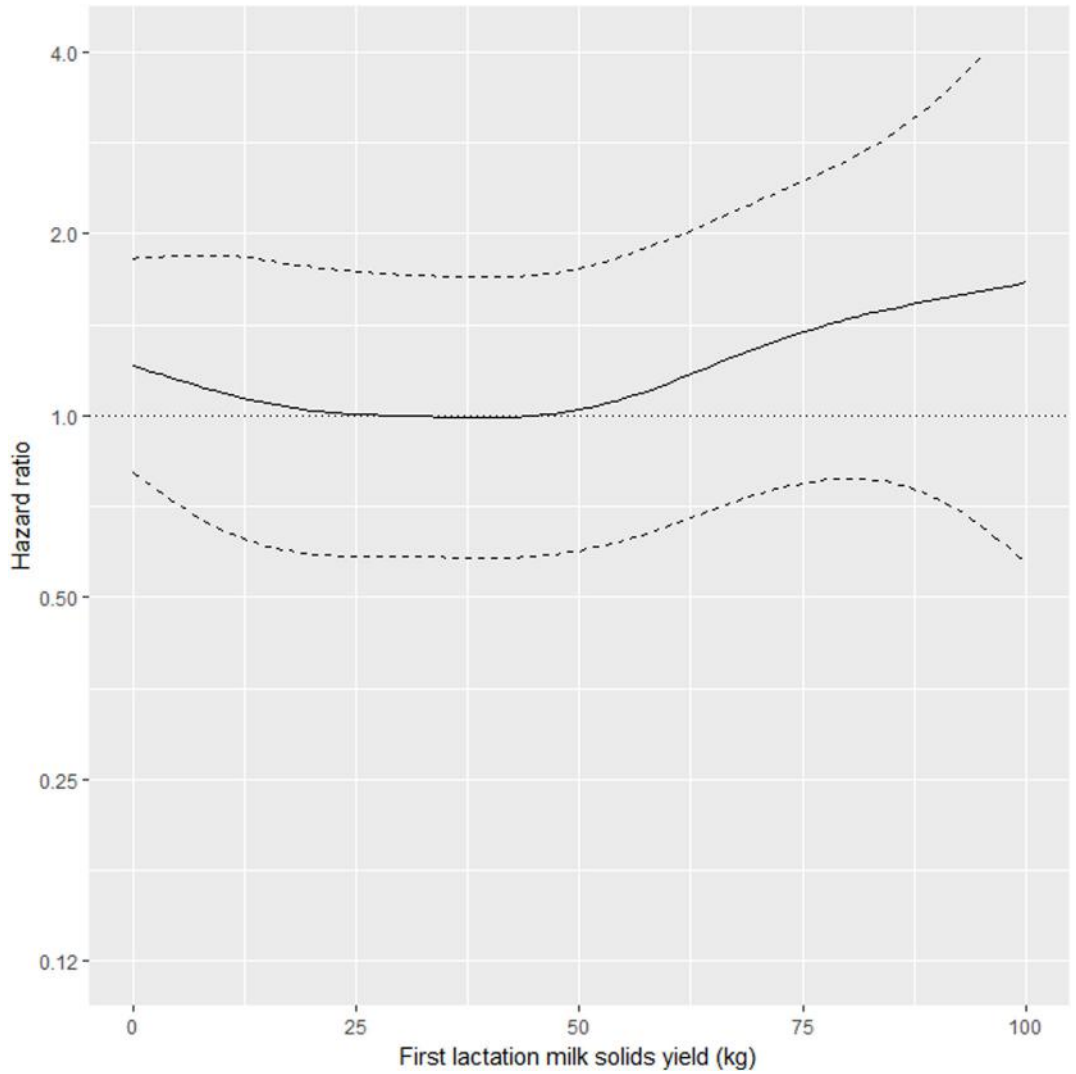


FIGURE 3 | Risk factors for culling, sales and deaths in New Zealand dairy goat herds, 2000–2009. Line plot showing, for the interval greater than 730 days from the date of second kidding, the hazard ratio for removal as a function of MSL1 (based on the model presented in **Table 2**). The dashed lines represent 95% confidence intervals around the point estimates of the hazard ratio. In the above plot, the reference category was a doe producing 30 kg milk solids in the first lactation. A doe producing 80 kg milk solids in the first lactation had 1.44 (95% CI 0.79–2.65) times the daily hazard of removal compared with a doe that produced 30 kg milk solids in the first lactation.

a cow is an early indicator of lifetime yield (21–23). While the total lifetime yield or daily milk yield in animals in subsequent lactations can be expected to be high in animals that produce more milk in the first lactation, overall reproductive performance decreases (22). Animals producing high amounts of milk in the first lactation are subject to a greater level of metabolic stress as a result of negative energy balance (20), which consequently leads to impaired fertility (24). Since we investigated the effect of MSL1 on LPL instead of first lactation milk yield and our study involved dairy goats, it is not possible to directly extrapolate the results of the above cow-based research to our study. Nevertheless, it is biologically plausible to assume that high yields of milk solids in first lactation would have a negative impact on the energy balance of dairy animals regardless of species. However, with good management, this negative effect may be unapparent for a reasonable period of time which, in our case, was approximately 2 years after the date of second kidding.

In this study, the effect of MSL1 on LPL was investigated using a model that included herd as a random effect (frailty) term. A frailty term is a continuous variable that quantifies the unobserved heterogeneity for groups of individuals such as those in families, classes, schools, or herds (25). Frailty terms are important because they provide a means for accounting for heterogeneity (i.e., "clustering") in outcome risk that arises from individuals within a cluster being more similar than individuals selected at random from the general population. Since variations in management practices among herds can be expected, the use of herd level effect as a frailty term is a standard practice

in epidemiological studies that quantify risk factors for given outcomes in domestic, farmed animal populations (26). The significance of the herd-level effect term in the model indicates that the hazard of removal as a function of LPL varied across herds. We propose that studies comparing herds with upper quartile frailty terms with those with lower quartile frailty terms may be useful to identify specific herd-level factors that are influential determinants of LPL. For example, a cross-sectional questionnaire survey can be designed to investigate various aspects of management such as nutrition, veterinary care, breeding practices, and milking practices in these two categories of farms and the data used to analyze differences between "low risk" and "high risk" herds in terms of survival.

In general, where heterogeneity is an unavoidable feature of the population under investigation, researchers should take into account the existence of dissimilarities among groups to avoid errors during analysis. By failing to acknowledge such heterogeneity, a researcher is more likely to make Type I error, which means he/she is likely to report a false association between explanatory and outcome variables when there is none. Interestingly, the protective effect of high MSL1 on the hazard of removal during T1 was evident only after the effect of herd was accounted-for in the model as a frailty term. When herd-level effects were not controlled-for, high MSL1 in L1 was positively associated with an increase in the risk of removal.

Several studies conducted on dairy cows have studied animal traits affecting LPL. Since longevity usually refers to the time between the first parity of an animal and its removal from the herd, it is not possible to get a direct measure of longevity for all animals, particularly those that are younger (6). However, with the use of survival analysis, such issues can be accounted-for because the technique uses information from all animals used in the study regardless of their culling status at the end of the study. Since we were interested to find out if MSL1 was associated with longevity, we defined longevity as the number of days between the date of second kidding and the date of removal from the herd. In this way, we could be sure that the explanatory variable (MSL1) preceded the study outcome (LPL), ensuring the correct temporal sequence between cause and effect.

CONCLUSION

This study identified a time varying effect of MSL1 on removal in New Zealand dairy goats. We found that does with high MSL1 yields had a lower risk of removal during the first 2 years following the second kidding compared with compared with their average producing herd mates. Beyond 2 years following the second kidding, does with high MSL1 yields had a relatively high hazard of removal compared with their average producing herd mates. We conclude that involuntary losses may be avoided if high MSL1 yielding does are preferentially managed from 2 years beyond the date of second kidding.

The data and analyses presented in this paper are based on the first author's thesis presented as partial fulfillment of the requirements for the degree of Master of Veterinary Studies at Massey University, New Zealand.

AUTHOR CONTRIBUTIONS

Study conception and design and critical revision: MG, MS, NL-V, and VM. Acquisition of data: NL-V and VM. Analysis and interpretation of data, and drafting of manuscript: MG and MS.

REFERENCES

1. Pérez-Razo M, Sánchez F, Torres-Hernández G, Becerril-Pérez C, Gallegos-Sánchez J, González-Cosío F, et al. Risk factors associated with dairy goats stayability. *Livest Prod Sci* (2004) 89:139–46. doi:10.1016/j.livprodsci.2004.02.008
2. Essl A. Longevity in dairy cattle breeding: a review. *Livest Prod Sci* (1998) 57:79–89. doi:10.1016/S0301-6226(98)00160-2
3. Jovanovac S, Raguž N, Sölkner J, Mészáros G. Genetic evaluation for longevity of Croatian Simmental bulls using a piecewise Weibull model. *Arch Anim Breed* (2013) 56:89–101. doi:10.7482/0003-9438-56-009
4. Solis-Ramirez J, Lopez-Villalobos N, Blair HT. Dairy goat production systems in Waikato, New Zealand. *Proceedings of the New Zealand Society of Animal Production*. Invercargill (2011). p. 86–91.
5. Stevenson M, Lean I. Risk factors for culling and deaths in eight dairy herds. *Aust Vet J* (1998) 76:489–94. doi:10.1111/j.1751-0813.1998.tb10188.x
6. Szabó F, Dákay I. Estimation of some productive and reproductive effects on longevity of beef cows using survival analysis. *Livest Sci* (2009) 122:271–5. doi:10.1016/j.livsci.2008.09.024
7. Forabosco F. *Breeding for Longevity in Italian Chianina Cattle [Doctor of Philosophy Dissertation]*. Wageningen, The Netherlands: Department of Animal Science, University of Wageningen (2005).
8. Seegers H, Beaudeau F, Fourichon C, Bareille N. Reason for culling in French Holstein cows. *Prev Vet Med* (1998) 36:257–71. doi:10.1016/S0167-5877(98)00093-2
9. Bell M, Wall E, Russell G, Roberts D, Simm G. Risk factors for culling in Holstein-Friesian dairy cows. *Vet Rec* (2010) 167:238–40. doi:10.1136/vr.c4267
10. Malher X, Seegers H, Beaudeau F. Culling and mortality in large dairy goat herds managed under intensive conditions in western France. *Livest Prod Sci* (2001) 71:75–86. doi:10.1016/S0301-6226(01)00242-1
11. Gautam M. *Epidemiological Study of Removals in New Zealand Dairy Goat Herds [Master of Veterinary Science Dissertation]*. Palmerston North, New Zealand: Institute of Veterinary, Animal and Biological Sciences, Massey University (2012).
12. Singireddy SR, Lopez-Villalobos N, Garrick DJ. Across-breed genetic evaluation of New Zealand dairy goats. *Proceedings of the New Zealand Society of Animal Production*. Lincoln: New Zealand Society of Animal Production (1997). p. 43–5.
13. Efron B. The efficiency of Cox's likelihood function for censored data. *J Am Stat Assoc* (1977) 76:312–9. doi:10.1080/01621459.1981.10477650
14. Cox D. Regression models and life tables. *J R Stat Soc* (1972) 34(B):187–220.
15. Kleinbaum D, Klein M. *Survival Analysis: A Self Learning Text*. New York, USA: Springer-Verlag (2012).
16. Ata N, Sözer M. Cox regression models with nonproportional hazards applied to lung cancer survival data. *Hacettepe J Math Stat* (2007) 36:157–67.
17. Therneau T, Grambsch P. *Modeling Survival Data: Extending the Cox Model*. New York: Springer (2000).
18. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (2017).
19. Robertson A, Barker JSF. The correlation between first lactation milk production and longevity in dairy cattle. *Anim Prod* (1966) 8:241–52. doi:10.1017/S0003356100034619
20. Pasman E, Otte M, Esslemont R. Influences of milk yield, fertility and health in the first lactation on the length of productive life of dairy cows in Great Britain. *Prev Vet Med* (1995) 24:55–63. doi:10.1016/0167-5877(94)00457-T
21. Haworth G, Tranter W, Chuck J, Cheng Z, Wathes D. Relationships between age at first calving and first lactation milk yield, and lifetime productivity and longevity in dairy cows. *Vet Rec* (2008) 162:1–6. doi:10.1136/vr.162.20.643

22. Sawa A, Krężel-Czopek S. Effect of first lactation milk yield on efficiency of cows in herds with different production levels. *Arch Anim Breed* (2009) 52:7–14.
23. Jairath L, Hayes J, Cue R. Correlations between first lactation and lifetime performance traits of Canadian Holsteins. *J Dairy Sci* (1995) 78:438–48. doi:10.3168/jds.S0022-0302(95)76653-X
24. Pryce J, Royal M, Garnsworthy P, Mao I. Fertility in the high-producing dairy cow. *Livest Prod Sci* (2004) 86:125–35. doi:10.1016/S0301-6226(03)00145-3
25. Wienke A. *Frailty Models. MPIDR Working Paper WP 2003-032*. Max Planck Institute for Demographic Research (2003). Available from: https://www.demogr.mpg.de/papers/working/wp-2003-032.pdf
26. Dohoo I, Martin S, Stryhn H. *Veterinary Epidemiologic Research*. Prince Edward Island, Canada: AVC Inc Charlottetown (2009).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer CP and handling editor declared their shared affiliation.

frontiers in Veterinary Science

TECHNOLOGY REPORT
published: 17 July 2017
doi: 10.3389/fvets.2017.00105

Novel Methods in Disease Biogeography: A Case Study with Heterosporosis

Luis E. Escobar[1,2,3], Huijie Qiao[4], Christine Lee[1] and Nicholas B. D. Phelps[1,2]*

[1] Minnesota Aquatic Invasive Species Research Center, University of Minnesota, St. Paul, MN, United States, [2] Department of Fisheries, Wildlife, and Conservation Biology, University of Minnesota, St. Paul, MN, United States, [3] Escuela de Estudios de Postgrado, Facultad de Medicina Veterinaria y Zootecnia, Universidad de San Carlos de Guatemala, Guatemala, Guatemala, [4] Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

Disease biogeography is currently a promising field to complement epidemiology, and ecological niche modeling theory and methods are a key component. Therefore, applying the concepts and tools from ecological niche modeling to disease biogeography and epidemiology will provide biologically sound and analytically robust descriptive and predictive analyses of disease distributions. As a case study, we explored the ecologically important fish disease Heterosporosis, a relatively poorly understood disease caused by the intracellular microsporidian parasite *Heterosporis sutherlandae.* We explored two novel ecological niche modeling methods, the minimum-volume ellipsoid (MVE) and the Marble algorithm, which were used to reconstruct the fundamental and the realized ecological niche of *H. sutherlandae*, respectively. Additionally, we assessed how the management of occurrence reports can impact the output of the models. Ecological niche models were able to reconstruct a proxy of the fundamental and realized niche for this aquatic parasite, identifying specific areas suitable for Heterosporosis. We found that the conceptual and methodological advances in ecological niche modeling provide accessible tools to update the current practices of spatial epidemiology. However, careful data curation and a detailed understanding of the algorithm employed are critical for a clear definition of the assumptions implicit in the modeling process and to ensure biologically sound forecasts. In this paper, we show how sensitive MVE is to the input data, while Marble algorithm may provide detailed forecasts with a minimum of parameters. We showed that exploring algorithms of different natures such as environmental clusters, climatic envelopes, and logistic regressions (e.g., Marble, MVE, and Maxent) provide different scenarios of potential distribution. Thus, no single algorithm should be used for disease mapping. Instead, different algorithms should be employed for a more informed and complete understanding of the pathogen or parasite in question.

Keywords: disease biogeography, risk map, ecological niche modeling, minimum-volume ellipsoid, heterosporosis

OPEN ACCESS

Edited by:
Victoria J. Brookes, University of Sydney, Australia

Reviewed by:
Hans-Hermann Thulke, Helmholtz-Zentrum für Umweltforschung (UFZ), Germany
Lina Mur, Kansas State University, United States

****Correspondence:***
Luis E. Escobar
ecoguate2003@gmail.com

Specialty section:
This article was submitted to Veterinary Epidemiology and Economics, a section of the journal Frontiers in Veterinary Science

Received: *31 January 2017*
Accepted: *19 June 2017*
Published: *17 July 2017*

Citation:
Escobar LE, Qiao H, Lee C and Phelps NBD (2017) Novel Methods in Disease Biogeography: A Case Study with Heterosporosis. Front. Vet. Sci. 4:105. doi: 10.3389/fvets.2017.00105

INTRODUCTION

Disease biogeography is the study of the geographic distribution of infectious diseases (1). It is a powerful approach for mapping disease events, which can inform decision-makers, managers, researchers, and animal and public health specialists (2, 3). Disease biogeography has been proposed as a promising field that can help understand why diseases emerge in one site, but not in another

(descriptive analyses), and also provides information to identify suitable areas where outbreaks could occur in the future (predictive analysis) (1).

Conceptual Bases

According to the assumption of disease biogeography, diseases are not distributed at random across the landscape, instead occur in non-random tractable and quantifiable landscape or environmental conditions. Disease biogeography incorporates the concept of the ecological niche as a crucial element to understand the environmental requirements of a disease transmission system as well as the geographic distribution of the species involved in the system (1, 2). Disease biogeographers use the conceptual bases and methods from the field of ecological niche modeling to make disease biogeography more quantitative (3, 4). Ecological niche modeling links field reports with environmental variables, allowing for development of the descriptive and predictive analyses required by disease biogeography. When ecological niche modeling is used for spatial epidemiology, it varies in complexity, ranging from simple "black-box" approaches (focusing on infected individuals only to reconstruct the conditions where the disease may persist) to more complex hierarchical ecological niche models (including several components of the disease system, e.g., intermediate host, reservoir, vector) (2). Black-box ecological niche models are usually employed for rare diseases where data for susceptible individuals, reservoirs, and vectors is scarce (3). Complex ecological niche models can be developed when more information is available, such as seasonality, density of vectors and reservoirs, and immunity of susceptible hosts, allowing to identify with more detail the different levels of disease transmission risk across areas, periods, and populations (1).

Theoretically, species' niches can be described as Fundamental Niche (N_F) and Realized Niche [N_R (5, 6); **Figure 1**]. The N_F would resemble the abiotic conditions not modifiable by the species and that are necessary by the species to survive and, most importantly, to maintain populations in the long term without the need for immigration. The N_R is represented by the portion of the N_F that is actually occupied by the species (2). N_F and N_R are usually estimated in ecological niche modeling based on field observations also termed *occurrences* and the environmental conditions in a region, here termed *background*. In the field of ecological niche modeling, considerable efforts have been made to develop methods and environmental variables to determine the N_F and N_R of species under the assumption that $occurrences \subseteq N_R \subseteq N_F \subseteq background$. Ecological niche modeling estimations are therefore developed in environmental dimensions to be later projected to geography in the form of maps of areas occupied and potentially occupied by the species in question (**Figure 1**).

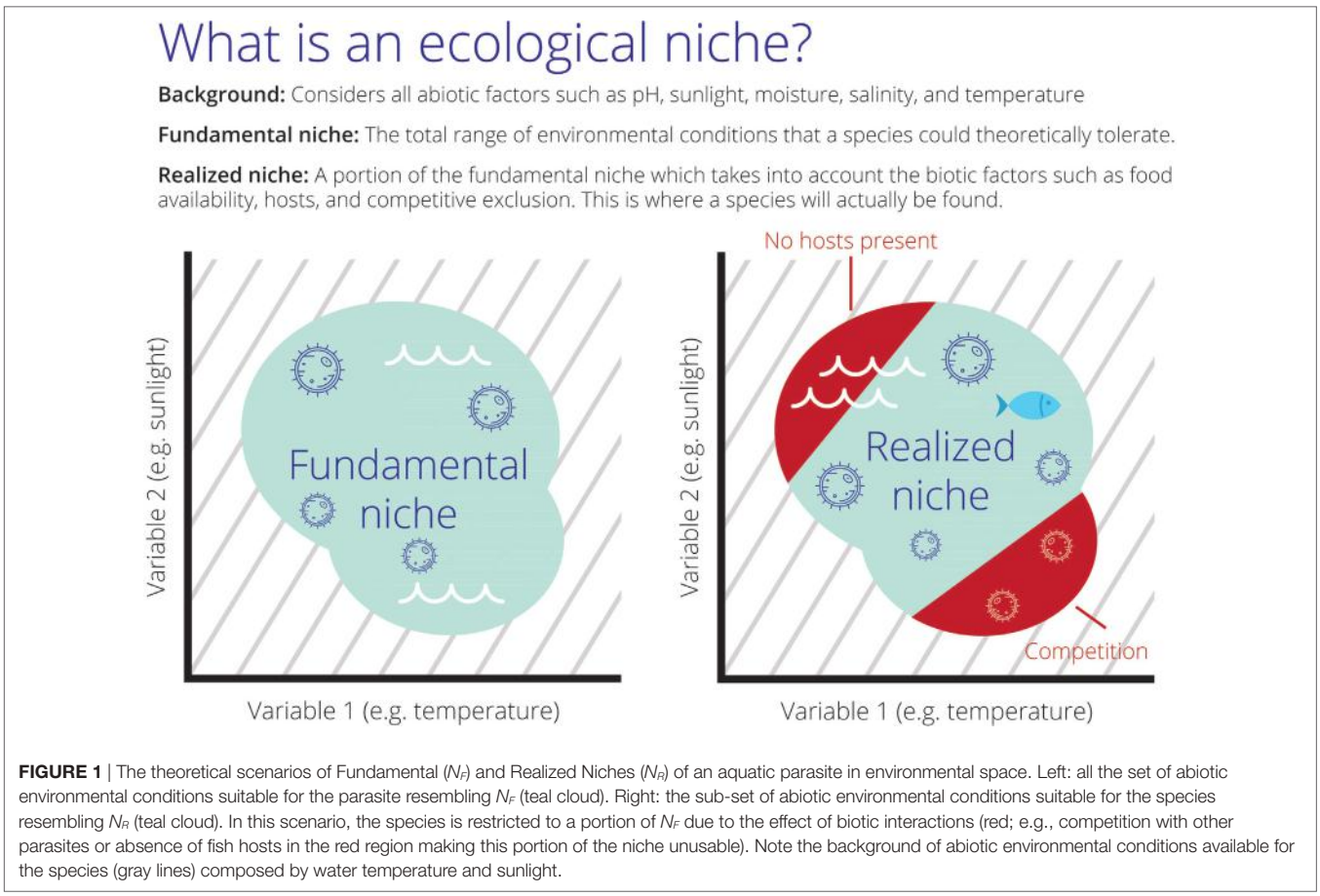


FIGURE 1 | The theoretical scenarios of Fundamental (N_F) and Realized Niches (N_R) of an aquatic parasite in environmental space. Left: all the set of abiotic environmental conditions suitable for the parasite resembling N_F (teal cloud). Right: the sub-set of abiotic environmental conditions suitable for the species resembling N_R (teal cloud). In this scenario, the species is restricted to a portion of N_F due to the effect of biotic interactions (red; e.g., competition with other parasites or absence of fish hosts in the red region making this portion of the niche unusable). Note the background of abiotic environmental conditions available for the species (gray lines) composed by water temperature and sunlight.

Applications in Epidemiology

While biogeographic methods have gained attention in the epidemiology of terrestrial ecosystems (3), they have been barely explored in the epidemiology of aquatic organisms (7). Examples of biogeographic analyses applied to infectious aquatic diseases include forecasts of *Gyrodactylus salaris* an ectoparasite of salmon (8), *Vibrio cholera* in coastal waters (9), and Viral Hemorrhagic Septicemia virus in the Great Lakes (10). Descriptive biogeographic analyses are useful to understand the natural history of novel infectious diseases, poorly known diseases, or diseases barely explored in the field (11–13). Predictive analyses are useful to anticipate risk in areas where the diseases has not yet been reported, and to guide active surveillance and research (14). A poorly understood infectious disease of epidemiological importance is Heterosporosis which infects fish in the Great Lakes region. Heterosporosis is caused by the microsporidian parasite *Heterosporis sutherlandae* and is known to infect at least eight fish species of economic and ecological importance (15). This disease was first confirmed in 2000 in Leech Lake and Catfish Lake in Minnesota and Wisconsin and has since been reported in waterbodies in Minnesota ($n = 26$), Wisconsin ($n = 16$), Michigan ($n = 2$) in the USA and Lake Ontario (15). The obligate intracellular parasites proliferate inside skeletal muscle cells (**Figure 2A**), eventually leading to liquefaction of the muscle tissue. Advanced stages of the disease likely result in indirect parasite-induced mortality due to decreased overall fitness, inability to capture prey or escape predation, and increased host stress (**Figure 2B**). The transmission of *H. sutherlandae* is thought to be horizontal, through the consumption of infected prey or contact with mature spores shed into the water column. Consequently, the overland transport of infected fish or water are likely risk factors for the spread of this pathogen. The possibility does exist for vertical transmission, similar to other microsporidian species infecting fish (16).

With Heterosporosis as a case study, we explored the use of next generation biogeography tools to evaluate how these tools and approaches can help (i) understand the ecology of a rare infectious disease and (ii) forecast the geographic areas where future investigation is necessary. This contribution aims to use the most state-of-the-art algorithms and variables available in order to incorporate disease biogeography in the toolkit of modern epidemiology.

METHODS

Occurrences

We obtained Heterosporosis-positive occurrence locations from Miller (17) and Phelps et al. (15), who in turn received

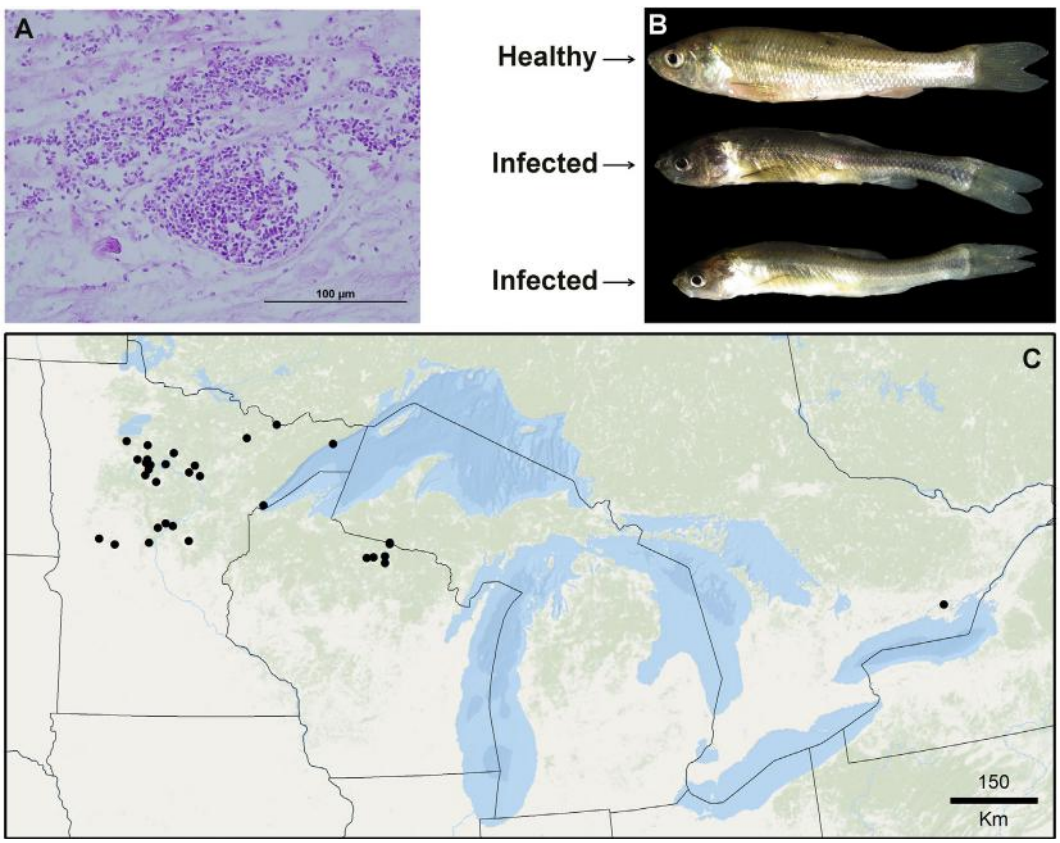


FIGURE 2 | Species used in this exploration. **(A)** Necrotic muscle tissue of the fish Fathead minnows (*Pimephales promelas*) infected with large aggregations of spores from the parasite *Heterosporis sutherlandae*. **(B)** Fathead minnows experimentally challenged with *H. sutherlandae*. **(C)** Heterosporosis-positive occurrences (black points) across the Great Lakes region used for this study. Lines denote administrative boundaries.

the reports from natural resource management agencies (i.e., Minnesota Department of Natural Resources, Wisconsin Department of Natural Resources, and U.S. Fish and Wildlife Service). Reports were confirmed by gross lesions and histopathology, and in some cases by PCR and sequencing. Anecdotal reports not verified in the laboratory were not included in this study. Lake centroids were used to determine latitude and longitude locations, and duplicate coordinates were removed. To explore the effect of data curation in the model's performance, models were developed using all the final occurrences available and a subset of resampled occurrences without environmental outliers (see below).

Fundamental Niche (N_F)

The N_F was estimated in a large model calibration region including: all the occurrences and the filtered occurrences. Specifically, we focused on the Laurentian Great Lakes region of North America (41.4° and 49.3°N and −97.8° and −74.8°W), a bi-national Canadian–American region with portions of the American states of Ohio, Illinoi, Indiana, Minnesota, Wisconsin, Michigan, Pennsylvania, New York, and the Canadian province of Ontario (**Figure 2C**). We used climatic variables from this calibration region to construct a *background* of environmental conditions in which the N_F was estimated (18) resembling the landscape and terrestrial environmental drivers where parasites and hosts co-occur. We used climate data from the CliMond repository (19), selecting the first 35 bioclimatic variables with original measurable information on annual, weekly, and seasonal temperature, soil moisture, radiation, and precipitation (**Table 1**), as these variables are a proxy to reconstruct ecoregions and present-day faunistic distributions (20). These variables are a summary of climatic conditions between 1961 and 1990 in the form of rasters at ~19 km spatial resolution. A principal component analysis was developed using NicheA software 3.0 (21) to reduce dimensionality and correlation between variables, retaining the first three components as they contained 83.85% of the information from the original set of variables. These three components composed the environmental *background* that summarized the environmental patterns in the area with reduced spatial and temporal autocorrelation and were used in posterior analyses. The background developed was then used by the ecological niche model algorithms to identify the relationship of parasite occurrences with this environmental background. Once this relationship is established, models search for this combination of conditions across the entire study area to define locations suitable and unsuitable for the parasite.

To mitigate uncertainty implicit in occurrences, we employed a method modified from Van Aelst and Rousseeuw (22) as filter to remove potential errors in occurrences. This filtering method is robust for outlier detection: we estimated minimum ellipsoids around occurrences displayed in environmental space and removed 5% [i.e., $\alpha = 0.05$ (3, 23)] of occurrences with the most marginal environmental values, as these outlier values could be associated with occurrence errors [e.g., misidentification; see, Ref. (24)]. The script for occurrences filtering by detection of the outliers has been included as Supplementary Material S1. We then estimated the N_F using NicheA with the remaining filtered

TABLE 1 | Environmental variables used to construct the background.

Fundamental niche	Realized niche
Annual mean temperature (°C)	Mean value of the monthly MODIS enhanced vegetation index (EVI) time series data (index)
Mean diurnal temperature range [mean(period max-min)] (°C)	SD of the monthly MODIS EVI time series data (index)
Isothermality (Bio02 ÷ Bio07)	Mean value the 8-day MODIS day-time land surface temperature (LST) time series data (°C)
Temperature seasonality (C of V)	SD of the 8-day MODIS day-time LST time series data (°C)
Max temperature of warmest week (°C)	Minimum value of the 8-day MODIS day-time LST time series data (°C)
Min temperature of coldest week (°C)	Maximum value of the 8-day MODIS day-time LST time series data (°C)
Temperature annual range (Bio05-Bio06) (°C)	Mean value the 8-day MODIS night-time LST time series data (°C)
Mean temperature of wettest quarter (°C)	SD of the 8-day MODIS night-time LST time series data (°C)
Mean temperature of driest quarter (°C)	Minimum value of the 8-day MODIS night-time LST time series data (°C)
Mean temperature of warmest quarter (°C)	Maximum value of the 8-day MODIS night-time LST time series data (°C)
Mean temperature of coldest quarter (°C)	Mean value of the 8-day MODIS day-time LST time series data for December/January (°C)
Annual precipitation (mm)	Mean value of the 8-day MODIS day-time LST time series data for February/March (°C)
Precipitation of wettest week (mm)	Mean value of the 8-day MODIS day-time LST time series data for April/May (°C)
Precipitation of driest week (mm)	Mean value of the 8-day MODIS day-time LST time series data for June/July (°C)
Precipitation seasonality (C of V)	Mean value of the 8-day MODIS day-time LST time series data for August/September (°C)
Precipitation of wettest quarter (mm)	Mean value of the 8-day MODIS day-time LST time series data for October/November (°C)
Precipitation of driest quarter (mm)	
Precipitation of warmest quarter (mm)	
Precipitation of coldest quarter (mm)	
Annual mean radiation (W m^{-2})	
Highest weekly radiation (W m^{-2})	
Lowest weekly radiation (W m^{-2})	
Radiation seasonality (C of V)	
Radiation of wettest quarter (W m^{-2})	
Radiation of driest quarter (W m^{-2})	
Radiation of warmest quarter (W m^{-2})	
Radiation of coldest quarter (W m^{-2})	
Annual mean moisture index	
Highest weekly moisture index	
Lowest weekly moisture index	
Moisture index seasonality (C of V)	
Mean moisture index of wettest quarter	
Mean moisture index of driest quarter	
Mean moisture index of warmest quarter	
Mean moisture index of coldest quarter	

Fundamental niche: variables based on climatic data at ~19 km spatial resolution. Realized Niche: variables based on MODIS data at ~1 km spatial resolution.

occurrences. The N_F was calculated as the minimum-volume ellipsoid (MVE) from the occurrences in a three-dimensional environmental scenario composed by the first three components from the original environmental variables, described elsewhere (21, 22). Basically, occurrences are displayed and analyzed in three environmental dimensions instead of two geographic dimensions (i.e., latitude and longitude). NicheA estimates the centroid point of the occurrences' cloud, which will be the center of the ellipsoid. Then, the Euclidean distance is estimated between the center of the ellipsoid and the most external occurrences. The two most external occurrences are the coordinate axes of the ellipsoid and in tandem with the Euclidean distances are used as parameters for a standard tri-axial ellipsoid equation (22). This ellipsoid was then used to simulate Gaussian response curves of the species to the environmental data employed to resemble ecological theories of species responses to environmental conditions (5, 25–27). To visualize the impacts of occurrences curation in estimations, a second model was developed as described above, but without occurrences filtered, i.e., using all the reports available to us.

Realized Niche (N_R)

The N_R was estimated in a reduced calibration region, including only areas falling inside the N_F model (**Figure 1**). In these sub-regions, we used 16 remotely sensed variables summarizing land surface temperature (LST) and primary productivity (28). Specifically, we used MODIS data at ~1 km spatial resolution, including day and night-time values of LST, and primary productivity in the form of enhanced vegetation index (EVI; **Table 1**) available from the WorldGrids repository (28).[1] These variables were also reduced in number and correlation via a principal component analysis that summarized >89.21% of the overall information from the original variables in the first three components.

We used the Marble algorithm to estimate the N_R. Marble is a novel algorithm that identifies clusters of occurrences in *n*-dimensional environmental spaces as has been described elsewhere (29). Briefly, Marble is based on the generalized density-based clustering algorithm that determines the position of occurrences in the multidimensional environmental space [see, Ref. (30)] and identifies clusters of occurrences of arbitrary shape but also is able to identify noise in the form of non-clustered occurrences in the environmental space [see, **Figure 6** in Ref. (29)]. The default parameters are the automatic estimation of the radii according to the number and position of occurrences allowing the inclusion of at least 99% of occurrences in the clusters. Due to the ability of the Marble algorithm to prioritize groups of occurrences and exclude isolated occurrences, the algorithm generates ecological niche models from consistent clusters only, with reduced interpolation and extrapolation. This approach results in models of metamorphosed shapes in the environmental space (29). The script employed in this study to develop Marble models in R has been included as Supplementary Material S2. We employed the occurrences and MODIS data that were inside the areas predicted by the N_F model. The N_F and N_R were then projected to the geographic space to identify areas suitable as predicted by the models.

Finally, to highlight the predictions of MVE and Marble vs. a classic ecological niche modeling method, we developed a series of models using Maxent algorithm (32). Maxent is a type of logistic regression (33) and is currently a standard method to estimate species' ecological niches (34). Maxent models included the estimation of the N_F based on climate data and N_R based on remote sensing data. The N_F and N_R were estimated using the original occurrences and filtered occurrences as described before. Models were calibrated using default settings in Maxent 3.3.3k (34).

All models were compared using a cumulative binomial distribution test using two sets of occurrences, one for model calibration and one for model evaluation, as in Peterson et al. (24). The R script used here for automated data split is included as Supplementary Material S3. Evaluation occurrences were not used during model calibration and instead were used to test the ability of the model to predict independent data using evaluation points as trials, evaluation points predicted correctly as successes, and the proportion of area predicted suitable as the probability of a success (23). The method used to develop this evaluation is included as Supplementary Material S4 to facilitate replication.

RESULTS

Once duplicates and environmental outliers were removed, 32 single occurrences remained and were used for modeling. The data curation process in the environmental space allowed us to identify several environmental outlier occurrences; one was removed based on our threshold defined *a priori* (**Figure 3**). The MVE estimated from this set of filtered occurrences, as a proxy of the N_F, revealed that the species was not occurring in all environmental conditions available in the model calibration region, instead, it occurred in consistent, tractable climatic conditions (**Figures 4** and **5**). When the N_F was projected from the environmental space to the geographic space, suitable areas were identified across North central Minnesota, northern areas of Wisconsin, and a small portion of western Michigan (**Figure 4**). Once the N_R of the parasite was estimated in these areas, we found suitability in specific areas of these states with high detail that allowed the identification of lakes that could be suitable for Heterosporosis (**Figure 4**). The Marble algorithm estimated fine scale suitability as a proxy of the N_R, based on a cloud of occurrences that excluded three isolated marginal occurrences detected outside of a main cluster (**Figure 3**). This generated a model of suitability based on the occurrences occupying the most tractable and consistent environmental conditions.

Once models were calibrated using all the data available, including the climatic outlier (**Figure 3**), the ecological niche models predicted broader areas suitable for Heterosporosis across the Great Lakes basin, resulting in 406% increase in areas predicted for this N_F model compared with the N_F without outliers (**Figure 6**). Changes in N_F estimations generated changes in the range of environmental values predicted suitable for the parasite (**Figure 5**). Changes in the range of environmental

[1] http://worldgrids.org.

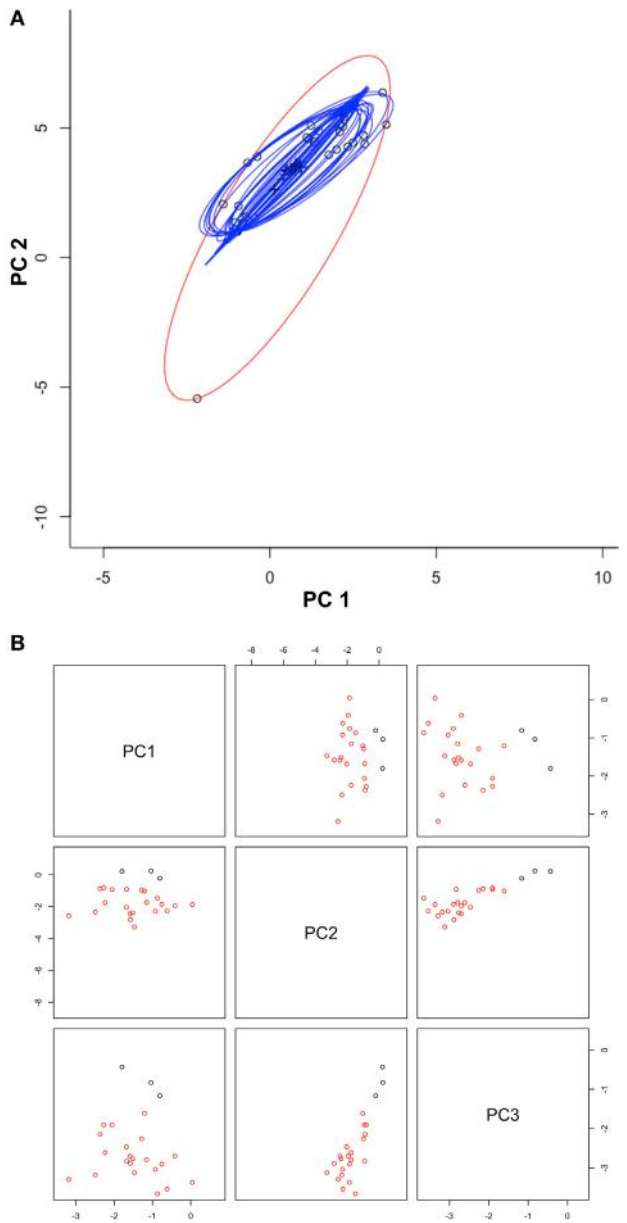


FIGURE 3 | Automated occurrences curation process. **(A)** Occurrences (black circles) were displayed in a two-dimensional environmental space of principal components one (PC1) and two (PC2) from the original climate data. Ellipsoids were estimated using the full occurrences (red ellipsoid) and then reducing one occurrence at a time (blue ellipsoids), to filter occurrences via outlier elimination. Note that using 100% of the points resulted in the detection of an outlier (black circle in edge of the red ellipsoid). **(B)** The first three PC from MODIS data were used to display the distribution of filtered occurrences (red circles) and also occurrences detected clusters (black circles). Note that outlier occurrences in term of climate were also outliers in terms of MODIS data (black points). The script for outlier detection is included as Supplementary Material S1.

tolerances occurred in the highest limit for some variables, while others showed shifts in the lowest limits. For some variables (e.g., maximum temperature, precipitation of wettest week, SD EVI, and maximum day-time LST), the impact of the outlier in the range of environmental tolerances was minimal, while others had more dramatic impacts in the range estimated (e.g., annual mean and minimum temperature, annual precipitation, and precipitation of the driest week; **Figure 5**).

Maxent models generated predictions comparable to those of Marble in the regions of Minnesota and Wisconsin. However, Maxent predictions were restricted to areas surrounding the occurrences when the entire data set was employed, showing low effect of outliers during model calibration as compared to MVE models (**Figure 6** vs. **Figure 7**). Using independent calibration and evaluation occurrences during model evaluations, all models showed prediction better than by chance in all the scenarios (Supplementary Material S5). The outputs, however, varied between algorithms. For example, we found that estimations of N_F was overfitted in Maxent, while MVE provided more generalized predictions when the model was calibrated using all the data available (**Figure 6** vs. **Figure 7A**).

DISCUSSION

Ecological niche models for Heterosporosis allowed the identification of suitable areas beyond the current locations with reports of the parasite, providing information about sites where the parasite could potentially occur based on suitable environmental conditions (4). MVE and Marble, the two novel algorithms employed in the modeling process, generated suitability surfaces in the form of binary maps showing areas with environmental conditions similar to those with Heterosporosis records (**Figures 4** and **6**). This binary modeling output format avoids continuous suitability surfaces of difficult biological interpretation (3). The models based on filtered occurrences without environmental outliers generated models with the best fit as expressed by the similarity of environmental conditions occupied by the occurrences vs. the conditions predicted by the MVE. That is to say, failure to remove outlier occurrences may have severe consequences in the areas predicted suitable by some ecological niche model algorithms (35), including MVE (see **Figure 4** vs. **Figure 6**). For example, removing outlier occurrences generated models with more detailed identification of regions suitable for Heterosporosis, thus, making forecasts a more useful tool to guide active epidemiological surveillance in specific constrained areas.

We found that the inclusion of environmental outliers also had a dramatic impact on the predictions in both the geographic and the environmental space. In this case, this was particularly true for the N_F models based on the MVE algorithm. For example, models calibrated with the environmental outlier generated predictions with high extrapolation for the higher values of predicted suitability, including annual mean and minimum temperature and annual precipitation and precipitation of driest week. For other variables, such as precipitation of wettest week, the outlier generated extrapolation in the lower values (**Figure 5**). We found, however, that in other variables the inclusion or not of the outlier occurrence was less dramatic (e.g., maximum temperature, SD of EVI, day-time LST values for the annual maximum and minimum, and the mean values for December and January, and for June and July; **Figure 5**). The Marble algorithm was less sensitive

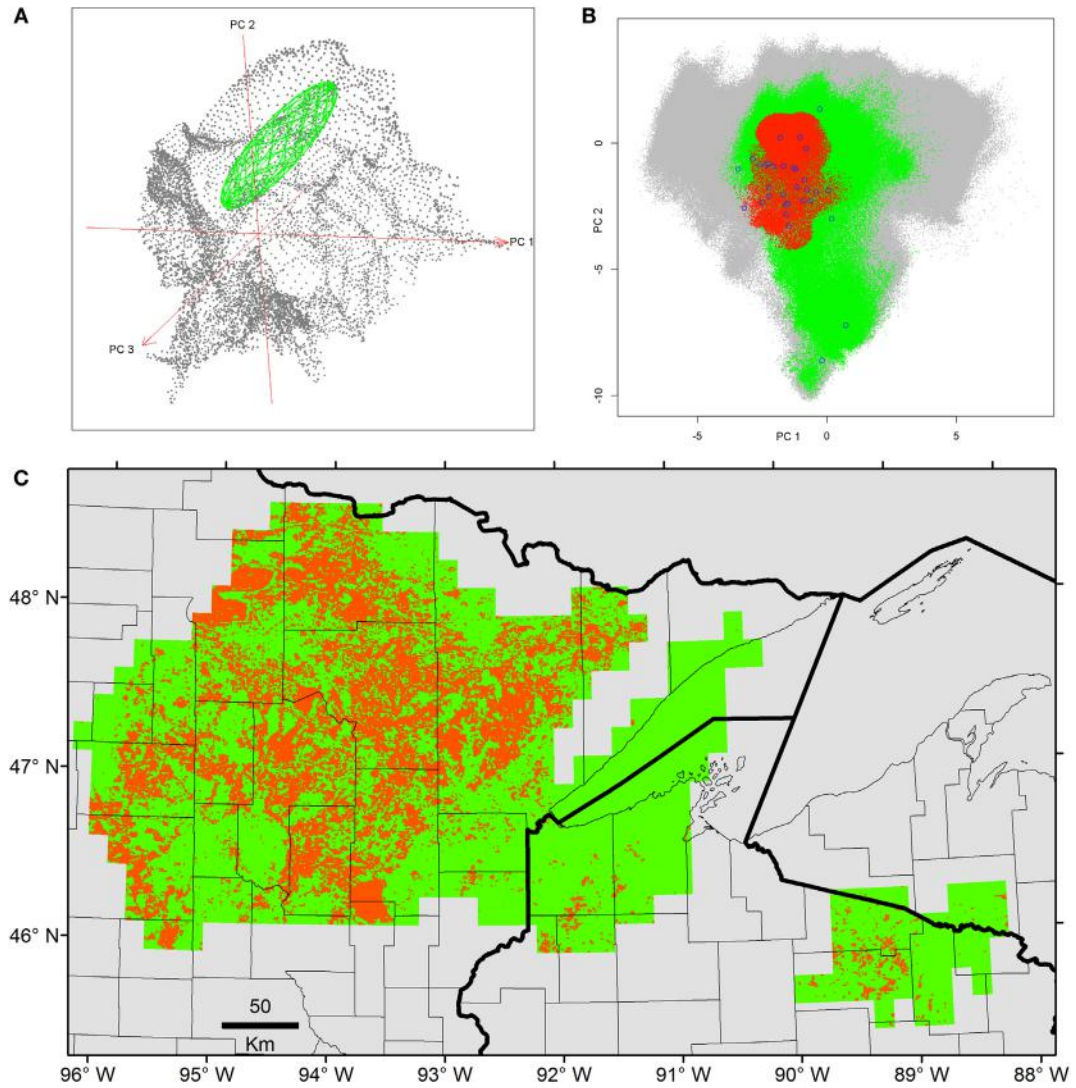


FIGURE 4 | Ecological niche models from filtered Heterosporosis data. **(A)** The Fundamental Niche (N_F; green ellipsoid) was estimated based on the minimum-volume ellipsoid formula in NicheA, using as background (gray points) the first three principal components (PC) of climate (red axes). **(B)** The Realized Niche (N_R; red) as estimated inside the conditions predicted suitable by the N_F (green) across the *background* constructed with the PC of the MODIS data (gray). **(C)** The N_F (green) and the N_R (red) were projected to the geography. In this case, the axes are longitude and latitude.

since this method automatically accounts for occurrences outside environmental clusters (**Figure 3**), i.e., noise detection (30).

Fundamental Niche (N_F)

According to ecological theories, the N_F of an organism should have an ellipsoidal form (21). This assumption is supported by experimental data showing Gaussian responses of species to abiotic environmental variables (26, 27, 36–39). The MVE estimated from the occurrences in environmental dimensions was able to generate response curves resembling normal distributions as the theory suggested (**Figure 5**), allowing us to have a proxy of the environmental tolerances of the species according to the data available to us. This suggests that NicheA could be a promising tool to simulate how species occupy environmental conditions based on field records; however, this would require high quality records. Erroneous records could tremendously impact the range of values used to estimate the ellipsoids (30), and in turn, the areas predicted suitable (**Figure 5**). To mitigate the inclusion of errors from the set of occurrences (40), we propose to employ an automated data curation system developed in environmental dimensions (**Figure 3**).

In addition to occurrence filtering, the estimation of MVEs is a protocol that requires a series of steps including a PCA analysis, displaying occurrences in the environmental space, calculations of ellipsoids, and projection of the final model to the geographic space. To facilitate this process, the workflow of the analyses developed here is included as Supplementary Material S6 to be executed in NicheA (21) and includes data to replicate this

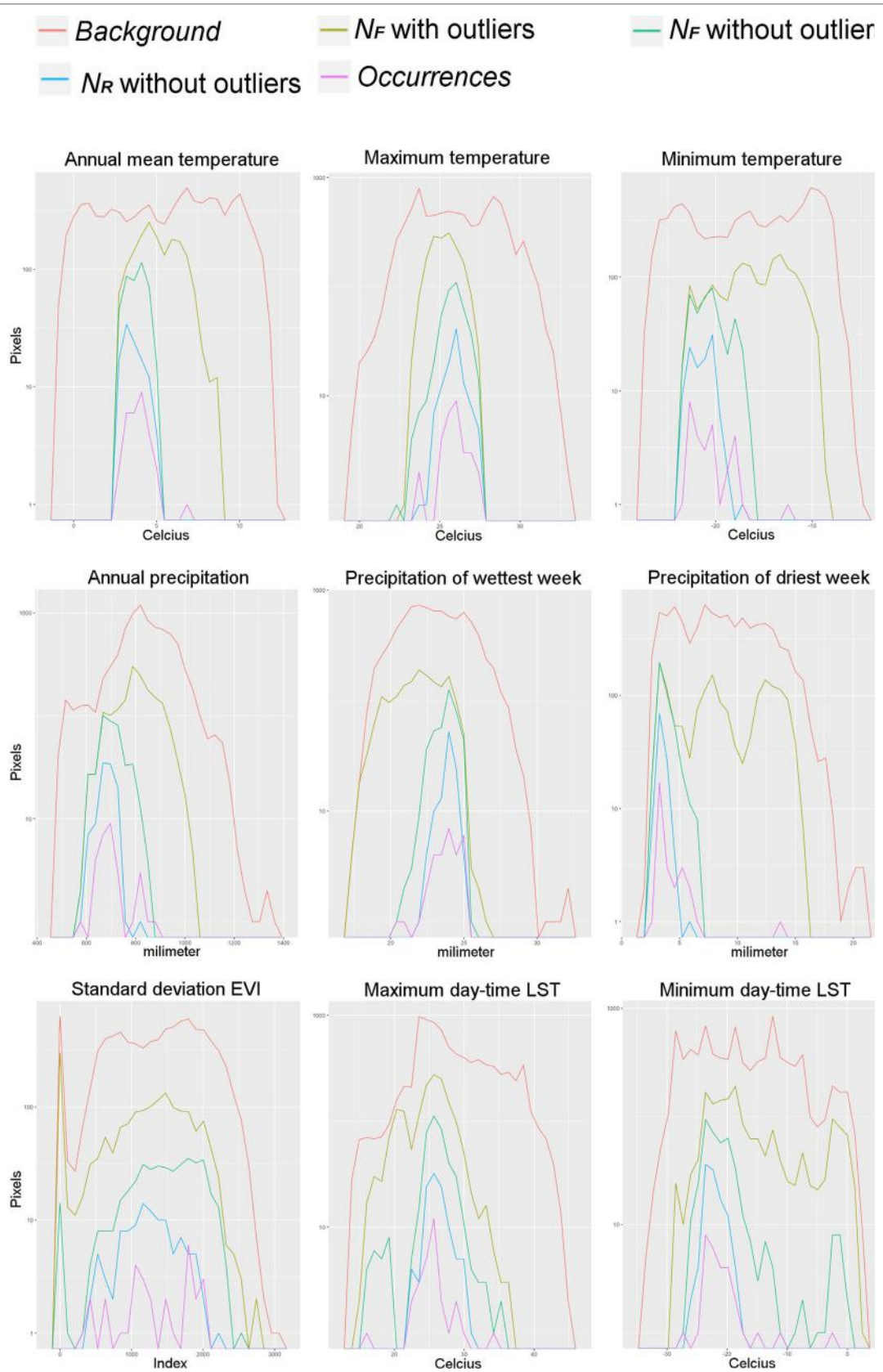


FIGURE 5 | Continued

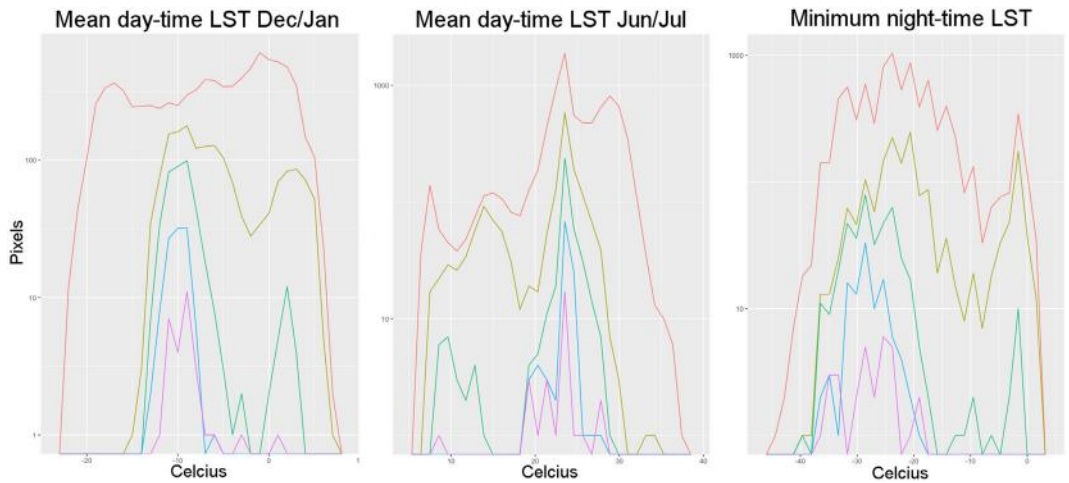


FIGURE 5 | Example of predictions represented in terms of single environmental variables. Pixels values of each environmental variable were counted across the study area representing the *background* (red line), the pixels predicted suitable by the Fundamental Niche (N_F) models based on a minimum-volume ellipsoid including all occurrences, i.e., with the environmental outlier occurrences (olive line), and with outliers removed (green line), and the estimation of environments occupied as predicted by the Realized Niche (N_R) model from the Marble algorithm (blue line). The *occurrences* employed for model calibration are also displayed (pink line). Count of pixels in log value for better visualization. Note that including all the occurrences without filtering generates high extrapolation of the N_F (i.e., broader range from the N_F estimations; olive line) compared with the models based on filtered occurrences (i.e., green line).

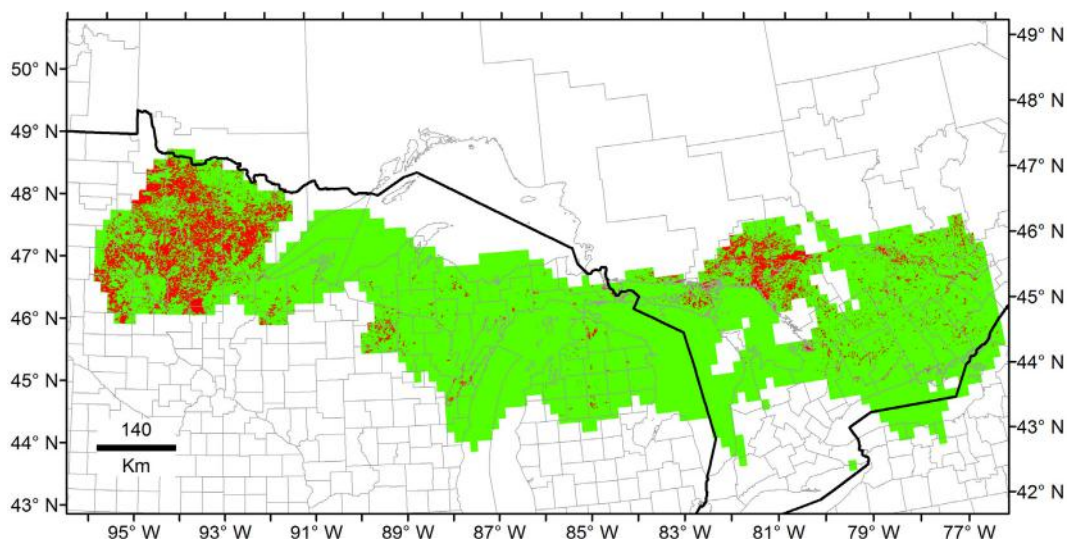


FIGURE 6 | Example of ecological niche models from unfiltered Heterosporosis data. The Fundamental Niche (N_F; green) was estimated based on the minimum-volume ellipsoid formula in NicheA, using all the occurrences including outliers and, as background, the first three principal components (PC) of climate. Then, the Realized Niche (N_R; red) was estimated inside the Fundamental Niche using marble algorithm based on the PC of the MODIS data (similar to **Figure 4** but with unfiltered occurrences).

workflow (Supplementary Material S6). Step-by-step instructions to estimate N_F of any species can also be found in the website of NicheA.[2]

Realized Niche (N_R)

While the N_F aims to estimate environmental tolerances, algorithms to estimate N_R, as the case with Marble, are meant to identify in environmental space the most "immediate" environmental conditions that are suitable to the species. In other words, models aiming to estimate the N_R are expected to overfit to the occurrences used for model calibration, resulting in a reduced interpolation and extrapolation. To our knowledge, this is the first application of Marble in epidemiology, and in turn in modeling diseases in fish. We showed that Marble is a promising algorithm to estimate realized niches, which in turn estimates areas that are suitable in high detail, avoiding the inclusion of environmental conditions beyond those currently used by the species.

[2]http://nichea.sourceforge.net/.

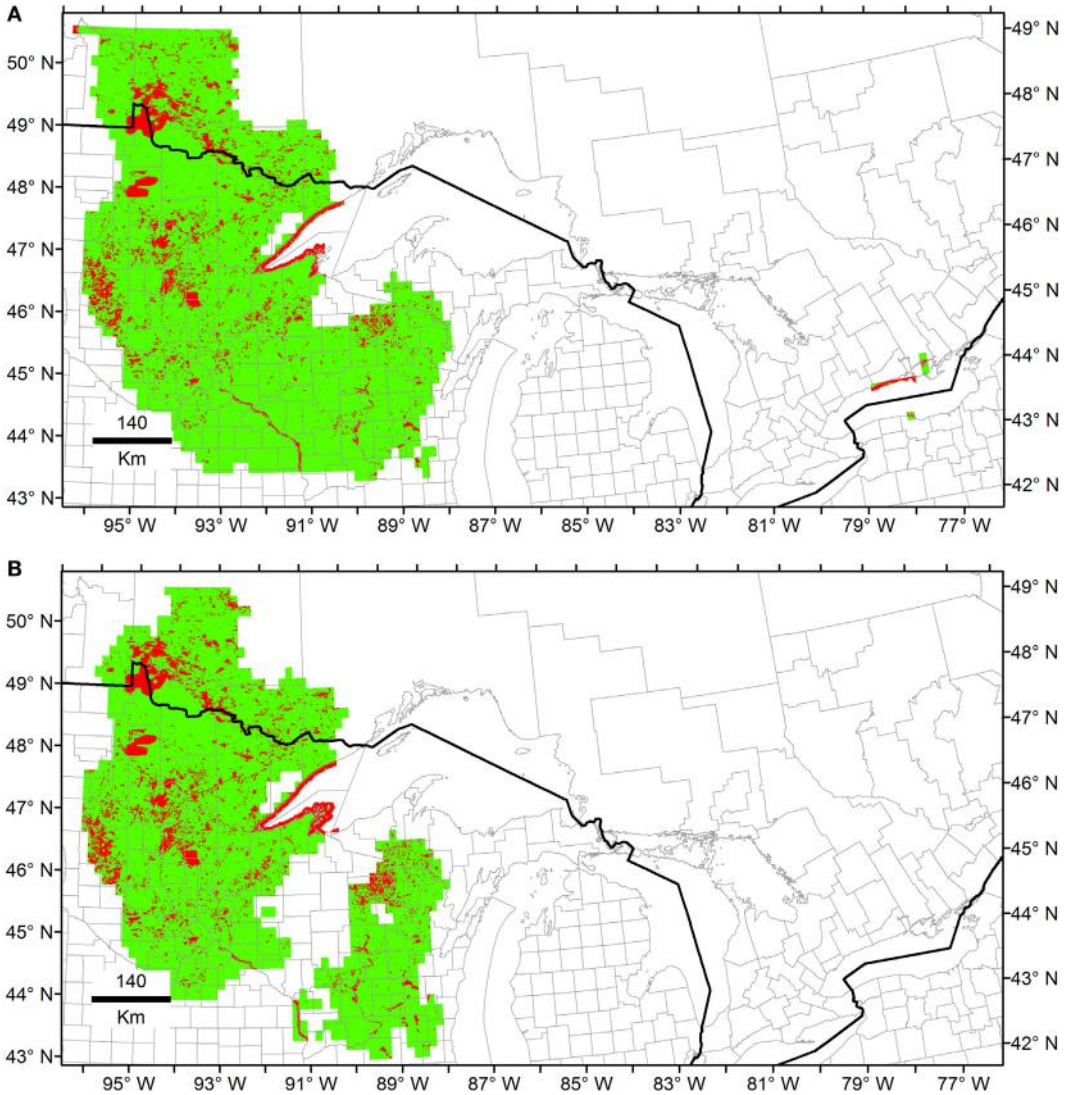


FIGURE 7 | Ecological niche models from Heterosporosis data using Maxent. The Fundamental Niche (N_F; green) was estimated using as background the first three principal components (PC) of climate. Then, the Realized Niche (N_R; red) was estimated inside the Fundamental Niche based on the PC of the MODIS data. **(A)** Models using all the occurrences available. **(B)** Models based on filtered data without outliers.

Novel vs. Classic Methods

We explored two novel methods to estimate species niches based on (i) algorithms resembling ecological theories (i.e., MVE and Marble) and (ii) algorithms resembling the data (i.e., Maxent). All models showed that predictions of independent occurrences were better than random in all model scenarios. However, it was evident that the machine learning structure of Maxent provides a high fit of the model with the data available (33). If assumptions are more relaxed and the data and information of the species are limited, MVE can be a good solution as this algorithm is less complex than Maxent (requires less parameters during calibration). This predictive behavior was replicated during N_R estimations: Marble provided generalized estimations with broad areas predicted suitable for the parasite and Maxent provided more conservative estimations principally in sites surrounding reports. We note that both modeling approaches, (i) algorithms resembling ecological theories (i.e., MVE and Marble) and (ii) algorithms resembling the data, are not wrong. In fact, both approaches develop niche estimations based on different assumptions: algorithms resembling ecological theories may overestimate the areas suitable due to the high levels of interpolation (31) aiming to reconstruct niche shapes as supported species physiology (21), while machine learning algorithms may have increased sensitivity to the data due to reduced extrapolation and interpolation to gain model fit. We argue that both approaches have pros and cons, one can prefer a simple model generalizing the niche estimation to gain knowledge or one can prefer a model with limited overestimation to obtain predictions

dictated by the data. Under both scenarios, the study question and assumptions will vary. For example, one can assume that Heterosporosis is still on its path to occupy the full ecological niche (i.e., ecological equilibrium) and model over estimations reducing the overfit of models to the data would be desirable. To mitigate uncertainties during model selection, two main frameworks could be considered in ecological niche modeling, one in which several algorithms are explored to capture consensus and variability (31), and one in which a single algorithm is explored under a detailed parameterization and assumptions based on abundant data and a considerable knowledge of the species in question (41).

Further Research

Current methods for disease mapping in epidemiology are dominated by distance-based analyses restricted to geography (e.g., spatial clusters), neglecting the importance of the landscape heterogeneity (42). However, recent literature in epidemiology has attempted to consider the climate and/or the landscape configuration when mapping disease transmission risk (1). While these attempts have important benefits in terms of the information generated and biological realism in the maps produced, most of these studies still lack a biogeographic framework to design the study and interpret the results. Indeed, click-and-run tools to generate ecological niche models are common in the scientific literature with studies of poor study design, but more strikingly without justification of the model parameters, assumptions, variables, occurrences, and study areas selected, even when such factors have been largely recognized as crucial in ecological niche modeling (4, 33–35, 43, 44).

Our study case focused on a fish parasite; thus, the model was calibrated using exclusively infected fish, resulting in a "black-box" approach as a proxy for all the species acting in the Heterosporosis system: the parasite and the susceptible hosts (2). Future studies are necessary at finer scales in the areas identified here as suitable for the parasite to include fish density, fish community assemblages, and other competitive parasites limiting the occurrence of Heterosporosis at a local level.

We assumed that N_F could be reconstructed using environmental data at coarse resolution, while N_R would require environmental variables at finer grain. These assumptions may be a limitation to the areas predicted by the models and should be a crucial point during the study design of models developed for spatial epidemiology. Beyond resolution, models could be impacted by the assumptions on the response of species to the environmental values absent in the occurrence data available. An important assumption is environmental interpolation. MVE has high interpolation of values predicting suitable all the environmental conditions falling inside the range of values estimated from the available occurrences. Thus, MVE would be less sensitive to sampling bias but would be sensitive to outliers. Maxent and Marble have limited interpolation with overfit to the data available, resulting in suitable conditions resembling the data. Thus, these algorithms are more sensitive to sampling bias (e.g., oversampling close to the roads or only during summer conditions) but are less sensitive to outliers. A good practice would be a careful selection of algorithms with the abilities to answer the research question, i.e., estimation of the potential distribution (N_F) or current distribution of the disease (N_R), considering the weaknesses in the environmental data (e.g., resolution) and occurrence data (e.g., bias).

Final Remarks

Several ecological niche modeling tools exist to map infectious diseases, but easy-to-use tools are preferred even if most users do not understand how the algorithms work (45). For instance, Maxent, an easy-to-use ecological niche modeling software, has suffered abuse in its application to epidemiology in a series of "recipe-like" studies with Maxent assumptions that may not be appropriated to the particular study questions (1, 3, 46–48). In biogeography, ecological niche modelers have cautioned the development of models with poor study design (3, 40, 46, 49, 50), which may lead to incorrect assumptions and interpretations. The algorithm selection and study design is particularly crucial in applications of ecological niche modeling to epidemiology, considering that modeling outputs could be used by public health intelligence and animal health policy makers.

We propose novel ecological niche modeling methods that can help understand the biogeography of an aquatic infectious disease, identify areas at risk for disease transmission, and can complement current methods. First, we highlight the importance of data curation and show a method for outlier removal in environmental dimensions based on *a priori* assumptions. Also, the ecological niche modeling algorithms proposed require low parameterization as they are based on the position (MVE) and density (Marble) of occurrences in an environmental space (22, 30), but also require a series of biological assumptions to make the outputs interpretable [e.g., Fundamental Niches of an ellipsoidal shape (21)]. We found that exploring algorithms of different analytical nature such as those aiming to fit environmental clusters, climatic envelopes, and logistic regressions (e.g., Marble, MVE, and Maxent) provided different scenarios of the potential distribution of Heterosporosis. Thus, no single algorithm should be used for disease mapping as this may result in an incomplete panorama of forecasts. We argue that different algorithms are necessary to achieve more informed predictions of the potential distribution of pathogen or parasites of public health or veterinary concern.

AUTHOR CONTRIBUTIONS

LE conceived and designed the study, collected and analyzed the data, and wrote the paper. HQ analyzed the data and co-wrote the paper. CL co-wrote the paper. NP collected the data and co-wrote the paper. All authors approved the final version of this manuscript.

ACKNOWLEDGMENTS

Authors thank Megan Tomamichel for providing important advice on the disease system.

FUNDING

This study was supported by the National Key R&D Program of China (2017YFC1200603), the Minnesota Environment and Natural Resources Trust Fund, the Minnesota Aquatic Invasive Species Research Center, and the Clean Water Land and Legacy Fund.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/fvets.2017.00105/full#supplementary-material.

SUPPLEMENTARY MATERIAL S1 | R script for outliers detection in environmental space (DOCX).

SUPPLEMENTARY MATERIAL S2 | R script to develop ecological niche models using Marble (DOCX).

SUPPLEMENTARY MATERIAL S3 | R script to split the occurrence data for model evaluation as in Ref. (24) (TXT).

SUPPLEMENTARY MATERIAL S4 | Spreadsheet formula to develop model evaluation (XLSX).

SUPPLEMENTARY MATERIAL S5 | Model evaluation results (DOCX).

SUPPLEMENTARY MATERIAL S6 | NicheA workflow and data to replicate analyses (XML).

REFERENCES

1. Escobar LE, Craft ME. Advances and limitations of disease biogeography using ecological niche modeling. *Front Microbiol* (2016) 7:1174. doi:10.3389/fmicb.2016.01174
2. Peterson AT. Biogeography of diseases: a framework for analysis. *Naturwissenschaften* (2008) 95:483–91. doi:10.1007/s00114-008-0352-5
3. Peterson AT. *Mapping Disease Transmission Risk: Enriching Models Using Biology and Ecology*. Baltimore: Johns Hopkins University Press (2014).
4. Peterson AT, Soberón J, Pearson RG, Anderson RP, Martínez-Meyer E, Nakamura M, et al. *Ecological Niches and Geographic Distributions*. New Jersey: Princeton University Press (2011).
5. Hutchinson GE. Concluding remarks. *Cold Spring Harb Symp Quant Biol* (1957) 22:415–27. doi:10.1101/SQB.1957.022.01.039
6. Soberón J. Grinnellian and Eltonian niches and geographic distributions of species. *Ecol Lett* (2007) 10:1115–23. doi:10.1111/j.1461-0248.2007.01107.x
7. Thrush MA, Murray AG, Brun E, Wallace S, Peeler EJ. The application of risk and disease modelling to emerging freshwater diseases in wild aquatic animals. *Freshw Biol* (2011) 56:658–75. doi:10.1111/j.1365-2427.2010.02549.x
8. Morris D. *Development of a Risk Evaluation System for the Establishment of Gyrodactylus salaris in Scottish River Systems*. Stirling: Scottish Aquaculture Research Forum (2011).
9. Escobar LE, Ryan SJ, Stewart-Ibarra AM, Finkelstein JL, King CA, Qiao H, et al. A global map of suitability for coastal *Vibrio cholerae* under current and future climate conditions. *Acta Trop* (2015) 149:202–11. doi:10.1016/j.actatropica.2015.05.028
10. Escobar LE, Kurath G, Escobar-Dodero J, Craft ME, Phelps NBD. Potential distribution of the viral haemorrhagic septicaemia virus in the Great Lakes region. *J Fish Dis* (2017) 40:11–28. doi:10.1111/jfd.12490
11. Estrada-Peña A, Ostfeld RS, Peterson AT, Poulin R, de la Fuente J. Effects of environmental change on zoonotic disease risk: an ecological primer. *Trends Parasitol* (2014) 30:205–14. doi:10.1016/j.pt.2014.02.003
12. Monroe BP, Nakazawa YJ, Reynolds MG, Carroll DS. Estimating the geographic distribution of human Tanapox and potential reservoirs using ecological niche modeling. *Int J Health Geogr* (2014) 13:34. doi:10.1186/1476-072X-13-34
13. Peterson AT, Lash RR, Carroll DS, Johnson KM. Geographic potential for outbreaks of Marburg hemorrhagic fever. *Am J Trop Med Hyg* (2006) 75:9–15.
14. Dicko AH, Lancelot R, Seck MT, Guerrini L, Sall B, Lo M, et al. Using species distribution models to optimize vector control in the framework of the tsetse eradication campaign in Senegal. *Proc Natl Acad Sci U S A* (2014) 111:10149–54. doi:10.1073/pnas.1407773111
15. Phelps NBD, Mor SK, Armién AG, Pelican KM, Goyal SM. Description of the microsporidian parasite, *Heterosporis sutherlandae* n. sp., infecting fish in the Great Lakes Region, USA. *PLoS One* (2015) 10:e0132027. doi:10.1371/journal.pone.0132027
16. Phelps NBD, Goodwin AE. Vertical transmission of *Ovipleistophora ovariae* (Microspora) within the eggs of the golden shiner. *J Aquat Anim Health* (2008) 20:45–53. doi:10.1577/H07-029.1
17. Miller P. *Diagnosis, Prevalence, and Prevetion of the Spread of the Parasite Heterosporis sp. (Microsporidia: Pleistophoridae) in Yellow Perch (Perca flavescens) and Other Freshwater Fish in Northern Minnesota, Wisconsin, and Lake Ontario*. Wisconsin: University of Wisconsin (2009).
18. Soberón J, Peterson AT. Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodivers Inf* (2005) 2:1–10.
19. Kriticos DJ, Webber BL, Leriche A, Ota N, Macadam I, Bathols J, et al. CliMond: global high-resolution historical and future scenario climate surfaces for bioclimatic modelling. *Methods Ecol Evol* (2012) 3:53–64. doi:10.1111/j.2041-210X.2011.00134.x
20. Ficetola GF, Mazel F, Thuiller W. Global determinants of zoogeographical boundaries. *Nat Ecol Evol* (2017) 1:89. doi:10.1038/s41559-017-0089
21. Qiao H, Peterson AT, Campbell LP, Soberón J, Ji L, Escobar LE. NicheA: creating virtual species and ecological niches in multivariate environmental scenarios. *Ecography* (2016) 39:805–13. doi:10.1111/ecog.01961
22. Van Aelst S, Rousseeuw P. Minimum volume ellipsoid. *Wiley Interdiscip Rev Comput Stat* (2009) 1:71–82. doi:10.1002/wics.19
23. Peterson ATT. Niche modeling: model evaluation. *Biodivers Inf* (2012) 8:41. doi:10.17161/bi.v8i1.4300
24. Peterson AT, Papes M, Soberón J. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecol Modell* (2008) 213:63–72. doi:10.1016/j.ecolmodel.2007.11.008
25. Austin MP, Cunningham RB, Fleming PM. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. *Vegetation* (1989) 55:11–27. doi:10.1007/BF00039976
26. Birch LC. Experimental background to the study of the distribution and abundance of insects: III. The relation between innate capacity for increase and survival of different species of beetles living together on the same food. *Evolution* (1953) 7:136–44. doi:10.2307/2405749
27. Hooper HL, Connon R, Callaghan A, Fryer G, Yarwood-Buchanan S, Biggs J, et al. The ecological niche of *Daphnia magna* characterized using population growth rate. *Ecology* (2008) 89:1015–22. doi:10.1890/07-0559.1
28. Hengl T, Kilibarda M, Carvalho-Ribeiro ED, Reuter HI. Worldgrids — a public repository and a WPS for global environmental layers. *WorldGrids*. (2015). Available from: http://worldgrids.org/doku.php?id=about&rev=1427534899
29. Qiao H, Lin C, Jiang Z, Ji L. Marble algorithm: a solution to estimating ecological niches from presence-only records. *Sci Rep* (2015) 5:14232. doi:10.1038/srep14232
30. Sander J, Ester M, Kriegel HP, Xu X. Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. *Data Min Knowl Discov* (1998) 194:169–94. doi:10.1023/A:1009745219419
31. Qiao H, Soberón J, Peterson AT. No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. *Methods Ecol Evol* (2015) 6:1126–36. doi:10.1111/2041-210X.12397
32. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. *Ecol Modell* (2006) 190:231–59. doi:10.1016/j.ecolmodel.2005.03.026
33. Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ. A statistical explanation of Maxent for ecologists. *Divers Distrib* (2011) 17:43–57. doi:10.1111/j.1472-4642.2010.00725.x
34. Merow C, Smith MJ, Silander JA. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* (2013) 36:1058–69. doi:10.1111/j.1600-0587.2013.07872.x

35. Boria RA, Olson LE, Goodman SM, Anderson RP. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecol Modell* (2014) 275:73–7. doi:10.1016/j.ecolmodel.2013.12.012
36. Angilletta MJ. *Thermal Adaptation: A Theoretical and Empirical Synthesis*. Oxford: Open University Press (2009).
37. Birch LC. Experimental background to the study of distribution and abundance of insects: I. The influence of temperature, moisture and food on the innate capacity for increase of three grain beetles. *Ecology* (1953) 34:698–711. doi:10.1017/CBO9781107415324.004
38. Rehfeldt GE, Ying CC, Spittlehouse DL, Hamilton DA. Genetic responses to climate in *Pinus contorta*: niche breath, climate change, and reforestation. *Ecol Monogr* (1999) 69:375–407. doi:10.2307/2657162
39. Soberón J, Nakamura M. Niches and distributional areas: concepts, methods, and assumptions. *Proc Natl Acad Sci U S A* (2009) 106:19644–50. doi:10.1073/pnas.0901637106
40. Peterson AT, Moses LM, Bausch DG. Mapping transmission risk of Lassa Fever in West Africa: the importance of quality control, sampling bias, and error weighting. *PLoS One* (2014) 9:e100711. doi:10.1371/journal.pone.0100711
41. Holt RD. Bringing the Hutchinsonian niche into the 21st century: ecological and evolutionary perspectives. *Proc Natl Acad Sci U S A* (2009) 106:19659–65. doi:10.1073/pnas.0905137106
42. Auchincloss AH, Gebreab SY, Mair C, Diez Roux AV. A review of spatial methods in epidemiology, 2000-2010. *Annu Rev Public Health* (2012) 33:107–22. doi:10.1146/annurev-publhealth-031811-124655
43. Barve N, Barve V, Jiménez-Valverde A, Lira-Noriega A, Maher SP, Peterson AT, et al. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecol Modell* (2011) 222:1810–9. doi:10.1016/j.ecolmodel.2011.02.011
44. Warren DL, Seifert SN. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecol Appl* (2011) 21:335–42. doi:10.1890/10-1171.1
45. Joppa LN, McInerny G, Harper R, Salido L, Takeda K, O'Hara K, et al. Troubling trends in scientific software use. *Science* (2013) 340:814–5. doi:10.1126/science.1231535
46. Anderson RP. Modeling niches and distributions: it's not just "click, click, click". *Biogeografía* (2015) 8:11–27.
47. Escobar LE. Modelos de nicho ecológico en salud pública: Cinco preguntas cruciales. *Pan Am J Public Health* (2016) 40:98.
48. Escobar LE, Peterson AT. Spatial epidemiology of bat-borne rabies in Colombia. *Pan Am J Public Health* (2013) 34:135–6.
49. Lash RR, Carroll DS, Hughes CM, Nakazawa Y, Karem K, Damon IK, et al. Effects of georeferencing effort on mapping monkeypox case distributions and transmission risk. *Int J Health Geogr* (2012) 11:23. doi:10.1186/1476-072X-11-23
50. Peterson AT, Nakazawa Y. Environmental data sets matter in ecological niche modelling: an example with *Solenopsis invicta* and *Solenopsis richteri*. *Glob Ecol Biogeogr* (2007) 17:135–44. doi:10.1111/j.1466-8238.2007.00347.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Veterinary Science

ORIGINAL RESEARCH
published: 17 October 2017
doi: 10.3389/fvets.2017.00172

Inferring the Ecological Niche of *Toxoplasma gondii* and *Bartonella* spp. in Wild Felids

Luis E. Escobar[1,2,3], Scott Carver[4†], Daniel Romero-Alvarez[5†], Sue VandeWoude[6], Kevin R. Crooks[7], Michael R. Lappin[8] and Meggan E. Craft[1]*

[1] Department of Veterinary Population Medicine, University of Minnesota, Minneapolis, MN, United States, [2] Department of Fisheries, Wildlife and Conservation Biology, University of Minnesota, St. Paul, MN, United States, [3] Department of Fish and Wildlife Conservation, Virginia Tech, Blacksburg, VA, United States, [4] School of Biological Sciences, University of Tasmania, Hobart, TAS, Australia, [5] Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS, United States, [6] Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, CO, United States, [7] Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, CO, United States, [8] Department of Clinical Sciences, Colorado State University, Fort Collins, CO, United States

OPEN ACCESS

Edited by:
Victoria J. Brookes, University of Sydney, Australia

Reviewed by:
Gerardo Acosta-Jamett, Universidad Austral de Chile, Chile
Anke Wiethoelter, University of Melbourne, Australia

***Correspondence:**
Luis E. Escobar
lescobar@umn.edu, escobar1@vt.edu

[†]*These authors have contributed equally to this work.*

Specialty section:
This article was submitted to Veterinary Epidemiology and Economics, a section of the journal Frontiers in Veterinary Science

Received: *31 March 2017*
Accepted: *28 September 2017*
Published: *17 October 2017*

Citation:
Escobar LE, Carver S, Romero-Alvarez D, VandeWoude S, Crooks KR, Lappin MR and Craft ME (2017) Inferring the Ecological Niche of Toxoplasma gondii and Bartonella spp. in Wild Felids. Front. Vet. Sci. 4:172. doi: 10.3389/fvets.2017.00172

Traditional epidemiological studies of disease in animal populations often focus on directly transmitted pathogens. One reason pathogens with complex lifecycles are understudied could be due to challenges associated with detection in vectors and the environment. Ecological niche modeling (ENM) is a methodological approach that overcomes some of the detection challenges often seen with vector or environmentally dependent pathogens. We test this approach using a unique dataset of two pathogens in wild felids across North America: *Toxoplasma gondii* and *Bartonella* spp. in bobcats (*Lynx rufus*) and puma (*Puma concolor*). We found three main patterns. First, *T. gondii* showed a broader use of environmental conditions than did *Bartonella* spp. Also, ecological niche models, and Normalized Difference Vegetation Index satellite imagery, were useful even when applied to wide-ranging hosts. Finally, ENM results from one region could be applied to other regions, thus transferring information across different landscapes. With this research, we detail the uncertainty of epidemiological risk models across novel environments, thereby advancing tools available for epidemiological decision-making. We propose that ENM could be a valuable tool for enabling understanding of transmission risk, contributing to more focused prevention and control options for infectious diseases.

Keywords: ***Bartonella*** **spp., environmental transmission,** ***Lynx rufus*****, niche,** ***Puma concolor*****,** ***Toxoplasma gondii***

INTRODUCTION

Traditional epidemiological studies of disease in animal populations are dominated by intraspecific transmission of contact-dependent (directly transmitted) parasites or pathogens (1, 2). However, many important parasites have complex life cycles that include vectors or environmental stages, and we often know much less about these types of parasites (3). For parasites or pathogens transmitted *via* vectors or the environment, it is especially important to understand not only the relationships between the host and pathogen, but also the environmental niche—the environmental conditions in which the pathogen persists in the long term (4). In practice, understanding

the environmental niche of many pathogens can be difficult to achieve due to challenges associated with detecting pathogens in vectors and the environment (e.g., sparsely distributed pathogens in vectors, in soil, on plant matter, or in water). As an alternative, capturing and sampling wildlife hosts is more effective. Innovative methodological approaches that overcome some of these environmental challenges are therefore needed and would be valuable for enabling understanding of transmission risk, thereby contributing to more focused prevention and control options.

Current approaches to map pathogens often include conducting a cluster analysis and spatial interpolations of disease cases in a specific area, thereby creating a tentative risk map for pathogen exposure (5, 6). However, a limitation of these classic approaches is the questionable value for forecasting risk in novel areas beyond those with ongoing surveillance. That is to say, geographic interpolations and cluster analyses do not consider environmental features and only reflect the sampling effort (7). Environmental (or ecological) niche modeling (ENM) is the practice of reconstructing a species' environmental determinants (8). These methods can be useful in creating predictive maps that can forecast pathogen presence in novel regions (9). Ecological niche modeling is established for species distribution modeling and is gaining attraction in the field of veterinary epidemiology (7).

Recent research has demonstrated the utility of ENM to predict disease in distant novel areas (8), but it remains rare for these predictive models to be validated using independent data, which is particularly true for models of pathogens in wildlife. We tackle this problem using a unique dataset of two pathogens, *Toxoplasma gondii* and *Bartonella* spp., isolated from wild felids across North America. Specifically, we analyzed samples from bobcats (*Lynx rufus*) and puma (*Puma concolor*), two secretive carnivores that are widespread in North America and are adaptable to a wide array of habitats where they are exposed to pathogens acquired from their environment (10–12). *T. gondii* is an intracellular protozoan parasite found in warm-blooded animals, including birds and mammals, and is transmitted *via* consumption of sporulated oocysts in feces, water, and soil or bradyzoites in tissues of prey species (13); in these wild felids, *T. gondii* is likely transmitted *via* consumption of infected prey such as rodents, lagomorphs, and cervids (10). The *Bartonella* genus includes gram negative anaerobic facultative intracellular bacteria species that cause an array of diseases affecting mammals; contact with arthropod vectors, particularly fleas, is the primary route of transmission of *Bartonella henselae*, *Bartonella koehlerae*, and *Bartonella clarridgeiae* (hereafter *Bartonella* spp.) (14, 15). Both, *T. gondii* and *Bartonella* spp., require other organisms to persist; thus, here we define them as micro-parasites or simply parasites (7).

This study has two primary tasks. First, we evaluate if ENM can characterize the potential distribution of parasites with complex lifecycles found in felid host species. This is particularly important when there is limited knowledge about the environmental niche of the pathogens, such as in this study. Second, we examine if ENM results from one region can be applied to other novel regions. We emphasize important novelties from this study: (i) this study utilizes remote sensing data that captures the habitat heterogeneity across study sites with high detail; (ii) this environmental heterogeneity is explicitly incorporated into risk maps produced by ENM; and (iii) we detail the uncertainty of epidemiological risk models across novel environments, thereby advancing epidemiological decision-making tools.

MATERIALS AND METHODS

Our dataset included 467 felids serologically positive for *T. gondii* and/or *Bartonella* spp. from Florida, Colorado, and California. Of these exposed felids, 328 were positive to *T. gondii* parasites and 234 to *Bartonella* spp.; occurrence records contained each animal's capture location and exposure status (16). These data were coupled with landscape information from satellite imagery to develop ENMs and create a risk map for each pathogen.

Occurrences

Occurrences of *T. gondii* and *Bartonella* spp. were recorded in ongoing research featuring an unusually large collection of wild felid serosurvey data from three different study areas: Florida, Colorado, and California (10, 12, 17, 18). The study areas were chosen as part of a previous study to represent a range of sites important for puma and bobcat conservation and were also representative of a wide degree of anthropogenic impacts (i.e., habitat fragmentation, urbanization, and agriculture) across North America. The Californian study region is a highly urbanized landscape characterized by a warm dry Mediterranean climate with vegetation communities dominated by coastal California sage scrub, chaparral, riparian and coastal oak woodlands, and annual grasslands. Colorado region was delimited by two polygons resembling sampling in rural and exurban areas with cold semi-arid climates and vegetation characterized by coniferous woodlands and forests primarily interspersed with aspens. The two regions in Colorado represent an area proximate to human development and a more natural area with agricultural surroundings. The Florida region is a mixture of urban, exurban, and agricultural areas spanning humid subtropical and tropical savanna climates with vegetation communities consisting of pine flatwoods, south Florida rockland, cypress domes and strands, dwarf cypress, prairies, mixed hardwood swamps, hardwood hammocks, freshwater swamps, and mangroves.

At each region, individual felids were captured, their location recorded, and samples for pathogen screening were collected according to protocols previously described (10, 12, 18). Wild felids were anesthetized using various tranquilizers/sedatives (19, 20), sampled, and released. Thoracic fluid was collected from hunter-killed animals instead of serum for a subset of bobcats from Colorado (11). Blood and serum samples were initially stored in ethylenediaminetetraacetic acid and serum-separating tubes. Samples were either refrigerated at 4°C or kept on ice until return from the field where they were temporarily frozen at −20°C, and later transferred to −80°C until screening for pathogen exposure. All procedures were performed after appropriate

Institutional Animal Care and Use Committee approvals were obtained.

Exposure to *T. gondii* and *Bartonella* spp. in puma (*P. concolor*) and bobcats (*L. rufus*) was estimated by measuring serum antibodies at the Specialized Infectious Disease Laboratory (Colorado State University) according to protocols previously described (10, 12). Serological samples were considered positive for *T. gondii* if they were positive to IgM or IgG. Samples were considered positive to *Bartonella* spp. if immunofluorescence antibody assay (IFA) tests detecting antibodies against *B. henselae* and *B. clarridgeiae* were positive (21–23); this was also confirmed independently by performing PCR on matched blood samples (12). For each study area and species, samples were generally collected over a 2- to 3-year intensive study period, and cumulatively the majority of samples across all sites were collected between 2001 and 2012 (12, 16). Puma and bobcat from Florida were not tested for *Bartonella* spp. (12). For the purpose of reducing overfit of models to the data, duplicate pathogen records from the same location (i.e., those from different individuals captured at the same location, but both exposed to the same pathogen) were restricted to single occurrence records for analyses.

Model Calibration Area

The area selected for ENM calibration has a direct effect on the model results (24), resulting in models area-dependent. Thus, the calibration area must hypothesize the occurrence potential and the sampling effort of the organism in question (25). Based on Poo-Muñoz et al. (26), we used the average distance among available occurrences for *T. gondii* and *Bartonella* spp. to generate a buffer around occurrences in each region. The buffered area was used as model calibration region (26), assuming that this region provided a proxy of the landscape conditions contained across the sampled areas (8). Total areas considered for each selected regions are as follow: ~52,500 km^2 for terrestrial area of California, ~105,400 km^2 for Colorado divided in two polygons (**Figure 1** left: ~64,700 km^2 and right: ~40,700 km^2), and ~43,000 km^2 for terrestrial areas of Florida (**Figure 1**).

Environmental Variables

Capturing fine-scale features of the landscape to understand the occurrence of pathogens is challenging and usually restricted to small study areas (27). A valuable alternative to landscape characterization is the use of satellite-derived remote sensing imagery. All objects emit radiation, at different intensities and wavelengths (28). This radiation can be characterized using satellite imagery from, for example, the MODerate-resolution Imaging Spectroradiometer sensor in the Terra satellite (29). These images offer low cost broad spatial coverage environmental information in the form of vegetation indexes (27), such as the Normalized Difference Vegetation Index (NDVI). The NDVI has proven representative of photosynthetic activity, biomass, net primary production, soil features, precipitations and humidity, and terrestrial landscapes in general. Thus, NDVI values have been associated with the distributional ecology and population dynamics of plants, invertebrates, birds, amphibians, ungulates, primates, carnivores, rodents, and reptiles in natural ecosystems; NDVI also provides information on changes in land use and soil humidity (27, 30).

Normalized Difference Vegetation Index data collected at 250 m spatial resolution at 16-day composites during 2005 in raster format were available from the Global Land Cover Facility (29). The resulting 21 original NDVI layers were reduced in number and collinearity *via* a principal component analysis (PCA) using ArcGIS 10.3 (31). We obtained new uncorrelated principal components (PC) with their respective descriptive values (e.g., correlation coefficients, eigenvalues, and eigenvectors). For the niche modeling procedure, we selected the PC summarizing at least 90% of the overall variance to capture a considerable amount of information from the original NDVI variables. The first three components were then utilized as axes to generate a three-dimensional environmental space as a proxy of Hutchinson's duality to extract the environmental information of the geography (32) and were used to display occurrences in environmental terms. This environmental space was developed using NicheA 3.0 software (33), available at http://nichea.sourceforge.net/.

Ecological Niche Modeling

We used Maxent 3.3.3k to generate the ecological niche models. Maxent is a machine learning tool developed to forecast species distributions with incomplete data (34). Maxent estimates the most uniform distribution of species occurrences compared with the available environmental background in the study area given constraints derived from the environmental data (35). Maxent also uses a regularization coefficient to increase or reduce the fit of the models to the available data, with a default value of 1 (36). We tested 20 regularization coefficients to find the best fit for our model. We used Akaike information criterion values corrected by sample sizes (AICc) to discriminate among models (37). This evaluation was developed using ENMTools 1.4.4 software (38). Specific settings in the final Maxent model included 100 bootstrap replicates with random seed and logistic output. The average of replicates in continuous format was converted to a binary format using a threshold value of $E = 5\%$; this threshold aims to remove 5% of the calibration occurrences with the lowest logistic value (8).

Occurrence data were split into the three buffered study regions (i.e., California, Colorado, and Florida). Models were calibrated with all the occurrences in two regions, models were then transferred (neither clamping nor extrapolation allowed in Maxent) to the remaining region (39), and were then evaluated with the occurrences from such region (40). For example, we calibrated models using occurrence data for *T. gondii* from two regions (e.g., California and Florida) and evaluated predictions with occurrence data in the third region (e.g., Colorado). For *Bartonella* spp., due to the lack of occurrence records in Florida, we used one site (i.e., Colorado) to predict the other (i.e., California) and *vice versa*. This split configuration assured a fair evaluation of the models by using data independent from that used during model calibration. Maxent predictions were tested between the three study regions using partial receiver operating characteristic (Partial ROC) (41), a metric developed for ecological niche models to assess the correct prediction

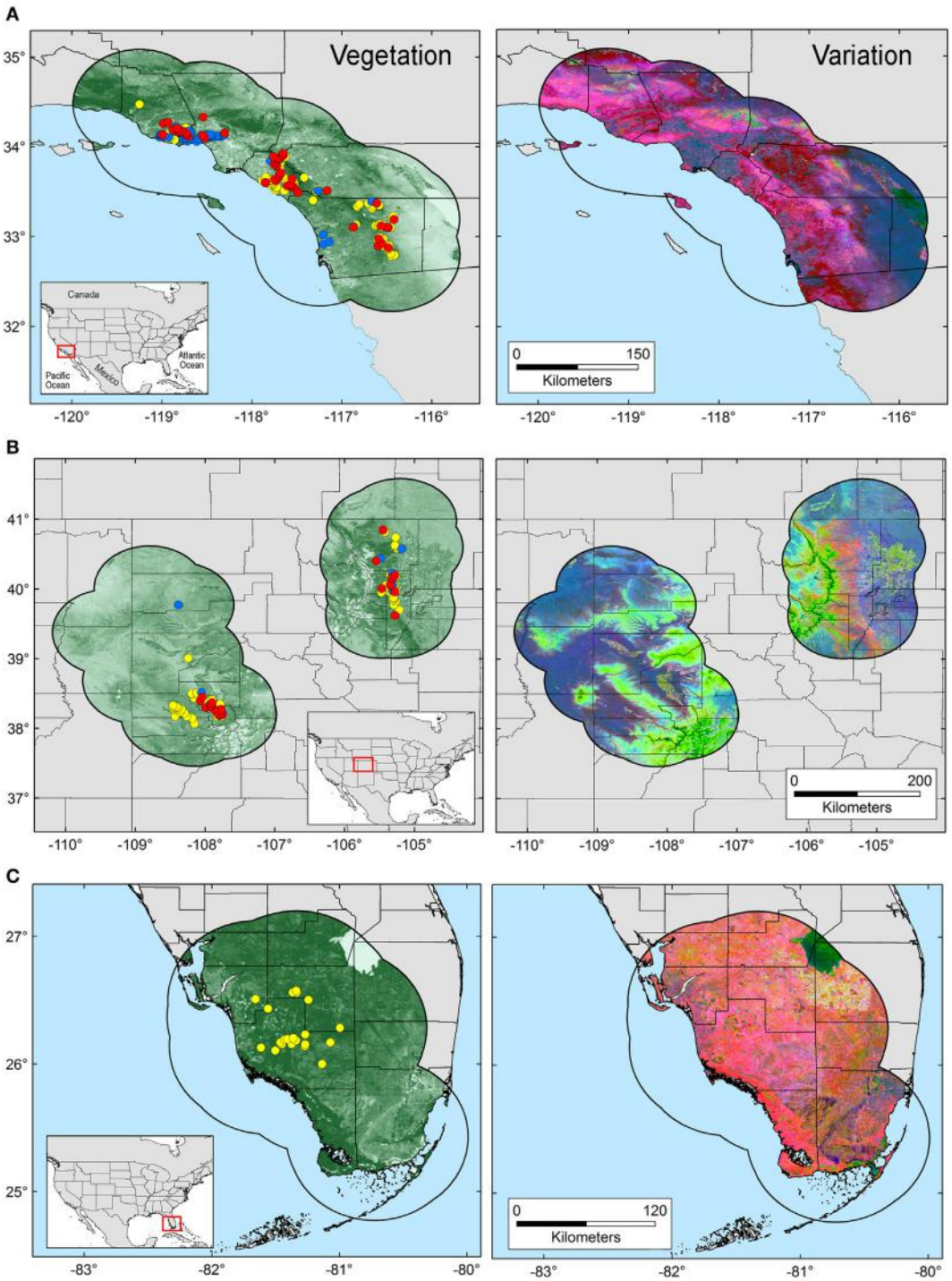


FIGURE 1 | Study areas and environmental variables employed in this study. Model calibration areas were defined in California **(A)**, Colorado **(B)**, and Florida **(C)** based on a buffer zone estimated from the average distance among occurrences. Left: occurrences for *Toxoplasma gondii* (yellow points), *Bartonella* spp. (blue points), and co-infections (red points) are displayed on a surface resembling landscape vegetation in the form of Normalized Difference Vegetation Index (NDVI) data. Right: original NDVI data were transformed to uncorrelated variables *via* principal component analysis. The variability across the study areas is summarized in principal components 1, 2, and 3, represented by the colors red, green, and blue, respectively.

of independent evaluation occurrences and the proportion of area predicted suitable, against a null model (42). Partial ROC analyses were conducted using the Partial ROC metric (41, 43); parameters included 5% of omission, $\alpha < 0.05$, 50% of random occurrences used for model testing, and 100 bootstrap iterations (41). Partial ROC estimates area under the curve (AUC)

ratio values ranging between 0 and 2, with values above 1 (null model) resembling predictions better than by random expectations that are considered statistically significant (42).

RESULTS

Environmental variables showed heterogeneous landscapes in spatial terms and collinearity among NDVI variables in temporal terms (**Figure 1**; Table S1 in Supplementary Material). For example, NDVI values in the summer (e.g., Julian days 193 and 209 in July in Table S1) showed low correlation with greenness with data from winter (e.g., days 1 and 17 in January in Table S1 in Supplementary Material). However, consecutive 16-day NDVI comparisons showed high correlation (e.g., Julian days 1 and 17, 17 and 33, and so on), with correlation coefficients ranging between 0.74 and 0.83 for comparisons between consecutive 21 layers (Table S1 in Supplementary Material). The first ten PC accumulated 90.77% of the overall information contained in the original 21 NDVI variables and were used for modeling (Table S2 in Supplementary Material). The first three components showed high environmental variability inside and between study areas, contained most of the information (80.37%) from the NDVI variables (Tables S2 and S3 in Supplementary Material), and showed differences in vegetation cover composition in California, Colorado, and Florida (**Figure 1**, right). Further, these three components were used to display the distribution of species in a three-dimensional virtual representation of the environmental space (**Figure 2**); here, the environmental distribution of both *T. gondii* and *Bartonella* spp. showed high overlap, despite the broader use of environmental conditions by *T. gondii* (**Figure 2**).

In all, 328 samples were positive for *T. gondii,* including 129 bobcats and 199 pumas across California, Colorado, and Florida (**Figure 1**). Two hundred thirty-four samples were positive for *Bartonella* spp. in 196 bobcats and 38 pumas from California and Colorado (**Figure 1**; **Table 1**). Models were calibrated using 291 single occurrence records for *T. gondii* and 189 occurrences for *Bartonella* spp. Models for both species required regularization coefficients other than the default value of 1 to have the best fit and lowest AICc: *T. gondii* required a regularization coefficient

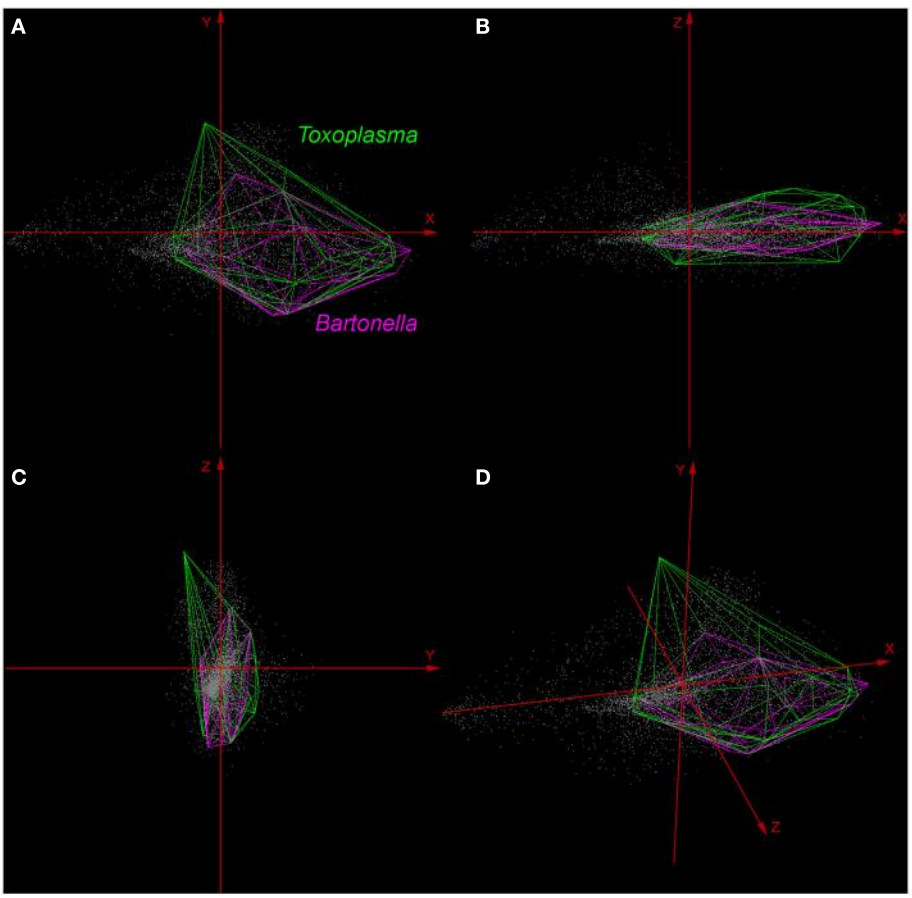


FIGURE 2 | Distribution of *Toxoplasma gondii* and *Bartonella* spp. in a three-dimensional representation of the environmental space. All the available occurrences of *T. gondii* (green polyhedron) and *Bartonella* spp. (pink polyhedron) were displayed based on environmental values (gray points) available in California, Colorado, and Florida. Axes (red arrows) were constructed using principal components (PC) 1 (*X* axis), PC 2 (*Y* axis), and PC 3 (*Z* axis), which are the same variables represented in the right side of **Figure 1**. **(A)** View of the occupied niches based on PC 1 and 2. **(B)** View of niches based on PC 1 and 3. **(C)** View of niches using PC 2 and 3. **(D)** Three-dimensional view of species distributions based on PC 1, 2, and 3.

TABLE 1 | Positive cases by host (bobcat and puma), parasite (*Toxoplasma gondii* and *Bartonella* spp.), and region.

Infection	Host	California	Colorado	Florida	Total
T. gondii	Bobcat	51	10	5	66
	Puma	69	82	16	167
	Total	120	92	21	233
Bartonella spp.	Bobcat	102	31	N/A	133
	Puma	5	1	N/A	6
	Total	107	32	N/A	139
Co-infections	Bobcat	46	17	N/A	63
	Puma	22	10	N/A	32
	Total	68	27	N/A	95

Samples from Florida were not tested for Bartonella spp.

of 1.2, while the *Bartonella* spp. model required a regularization coefficient of 1.3 (Table S4 in Supplementary Material). Once calibrated, model evaluations showed that predictions between states were significantly better than a random model (AUC ratios above 1, $p < 0.05$) when data of *T. gondii* from California and Colorado were used to predict the location of this parasite in Florida (mean AUC ratio = 1.056, SD = 0.044), from Colorado and Florida to California (mean AUC ratio = 1.089, SD = 0.066), and when data from California and Florida were used to predict occurrences in Colorado (mean AUC ratio = 1.247, SD = 0.085) (Figure S1 in Supplementary Material). *Bartonella* spp. models calibrated in California were significantly predictive of the occurrence of this parasite in Colorado with AUC ratios above 1; similarly, models calibrated in Colorado significantly predicted *Bartonella* spp. in California (mean AUC ratio = 1.107, SD = 0.029) (Figure S1 in Supplementary Material).

The *T. gondii* model identified suitable areas for this parasite, but also showed heterogeneity in uncertainty estimations across areas (i.e., predictions ranged from low uncertainty to high uncertainty in each area), with SD ranging between 8.13×10^{-6} (lowest) to 0.42 (highest; **Figure 3**). Binary models for *T. gondii* showed high proportion of suitability mainly in California (43.7% of the area) as compared with Colorado (35.8%) and Florida (20.5%); these predictions came with some variation in certainty (**Figure 3A**). Models also predicted isolated and limited suitability for both regions in Colorado, also with some variation evident in uncertainty, although these models were more confident in the places where *T. gondii* is unlikely to occur (**Figure 3B**). Florida showed wide suitability for this parasite across all the study areas (but with high uncertainty in suitability), except for consistent unsuitable predictions in Lake Okeechobee region (**Figure 3C**). Models for *Bartonella* spp. had a similar variation in predictions of suitable areas of pathogen occurrence, and uncertainty in predictions (SD ranging from 2.77×10^{-6} to 0.39; **Figure 4**). Our *Bartonella* spp. models predicted extensive suitability throughout California, with high certainty in unsuitable areas for *Bartonella* spp. occurrence (**Figure 4A**). It is notable that the area of uncertainty for *Bartonella* spp. in California was greater than for *T. gondii* (see **Figures 3A** vs. **4A**). In Colorado, models predicted low *Bartonella* spp. suitability in the study area to the west with high certainty, but higher suitability to the east (**Figure 4B**). Even when no pathogen records were available to us for *Bartonella* spp. in Florida, our model predicted suitable conditions in specific sites across this region (**Figure 4C**) and with less uncertainty than *T. gondii*. In general terms, however, *Bartonella* spp. was predicted to be less widespread, as compared with *T. gondii* in Colorado and Florida.

DISCUSSION

Here, we illustrate the utility of a cutting-edge analytical tool that can be used to advance the understanding of the epidemiology of pathogens with complex lifecycles. Our modeling framework attempted to reconstruct the occupied niche of the parasites in question [*sensu* (8)]—the subset of the environmental space occupied by the species in the area studied. That is to say, the host species included in the study have broad home ranges [puma ~48.6 km^2, bobcat ~30.7 km^2 (44)] and occur through the Americas from Canada to Patagonia (puma) or across North America (bobcats), a typical characteristic of Felidae (45, 46). Thus, our representation of patterns of suitable areas for parasites is a high-resolution site and time specific reconstruction of risk. We found that although exposure to both *T. gondii* and *Bartonella* spp. was generally widespread in the study areas (**Figure 1**), *T. gondii* showed a broader distribution across environmental conditions than did *Bartonella* spp. (**Figure 2**), suggesting a broader niche for *T. gondii*. Although *Bartonella* spp. was not tested in the Florida samples, our niche model experiments suggest suitability in diverse areas of this state. We found that our models were most accurate in predicting areas where these parasites were least likely to occur. Specifically, the uncertainty, expressed as variability found in our Maxent predictions was the smallest for areas predicted unsuitable for *T. gondii* and *Bartonella* spp. (**Figures 3** and **4**).

Ecological niche models, and NDVI satellite imagery, proved to be useful to characterize the potential distribution of the selected pathogens at the landscape level, generating distribution maps for *T. gondii* and *Bartonella* spp. from exposure in wild felids. NDVI captures with high accuracy information of soil features, temperature conditions, and changes of humidity and precipitations as expressed in the structure of local vegetation (27); thus, allowing to capture the environmental signature of *Bartonella*-positive reservoirs associated with increments on precipitation, as is the case for some *Bartonella* species (47, 48). Environmental variables showed collinearity, and thus, using PC instead of the original NDVI variables mitigated Maxent overfit by reducing correlation and number of parameters employed by the model. The PCA allowed us to capture landscape variation, which was evident when the first three PC were displayed for each study area (**Figure 1**, right), suggesting that NDVI is a powerful tool for epidemiological studies aiming to forecast disease transmission risk at a habitat level (i.e., 250 m spatial resolution).

We had predictive success when applying ENM from one region to other, even though there were marked environmental differences among regions (**Figure 1**). Nevertheless, although all predictions among regions were significant, not all of our sites were equal in predictive abilities (Figure S1 in Supplementary Material). This highlights the key role that environment similarity can play between calibration and projection areas in Maxent.

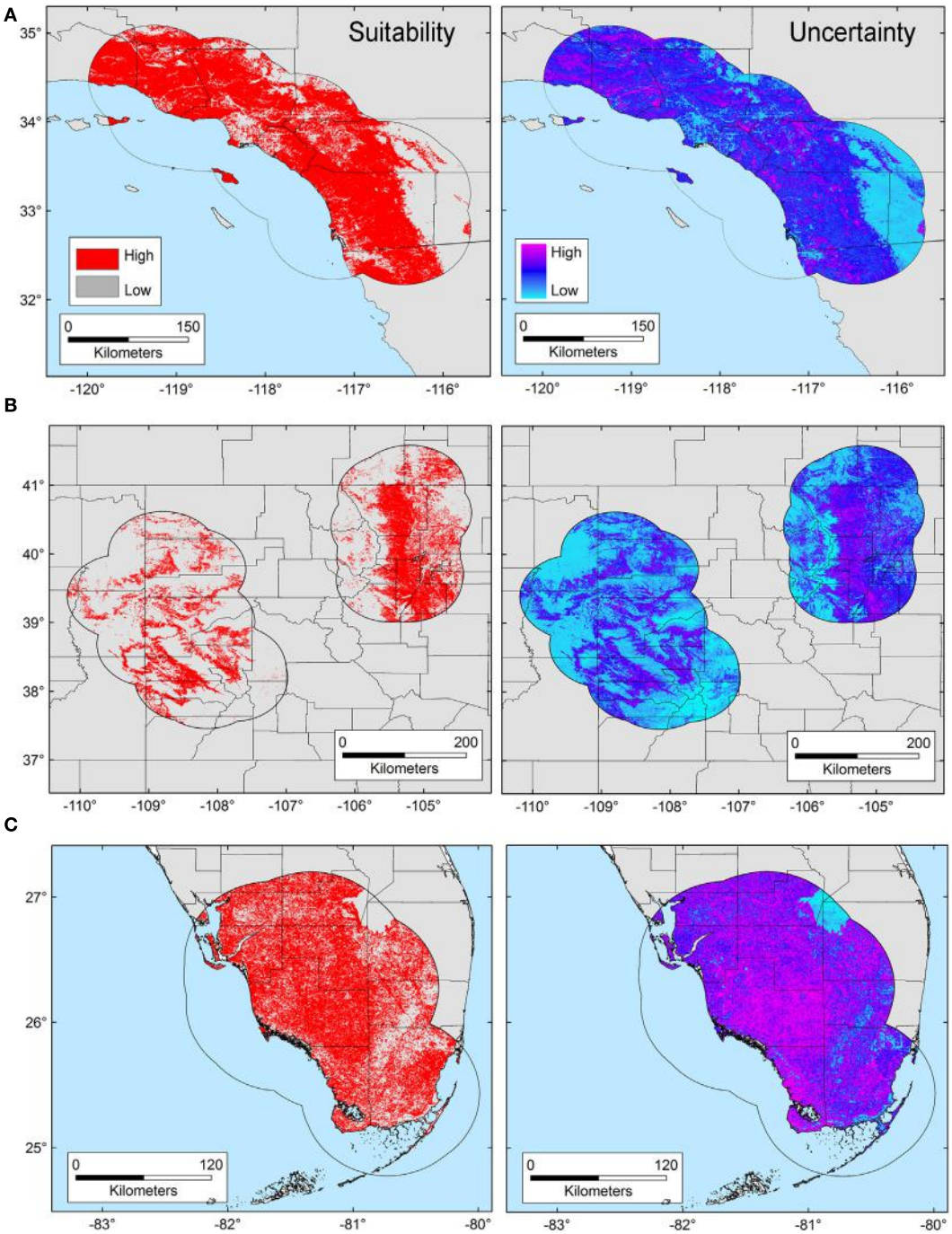


FIGURE 3 | Ecological niche model of *Toxoplasma gondii*. Binary maps of *T. gondii* suitability (red) were developed for areas in California **(A)**, Colorado **(B)**, and Florida [**(C)**; left panel]. Uncertainty estimations based on the suitability differences among models (right panel) show areas of low (cyan) and high (pink) uncertainty as follows: California **(A)** from 3.31×10^{-5} to 0.42, Colorado **(B)** from 9.96×10^{-6} to 0.32, and Florida **(C)** from 8.13×10^{-6} to 0.3.

Potentially this supports the idea that Maxent predictions are more consistently suited for transference to similar environmental conditions (39). Further, we also show that NDVI environmental data are robust for reconstructing the environmental conditions suitable for pathogens, similar to more routine approaches using climate variables.

ENM applied to environmental dependent pathogens facilitates the identification of habitats of risk where collection of information has been lacking maybe due to limited sampling effort or other factors related to the detection of pathogens. It implies an advancement in understanding the distribution of pathogens beyond the use of data of their vectors or reservoirs.

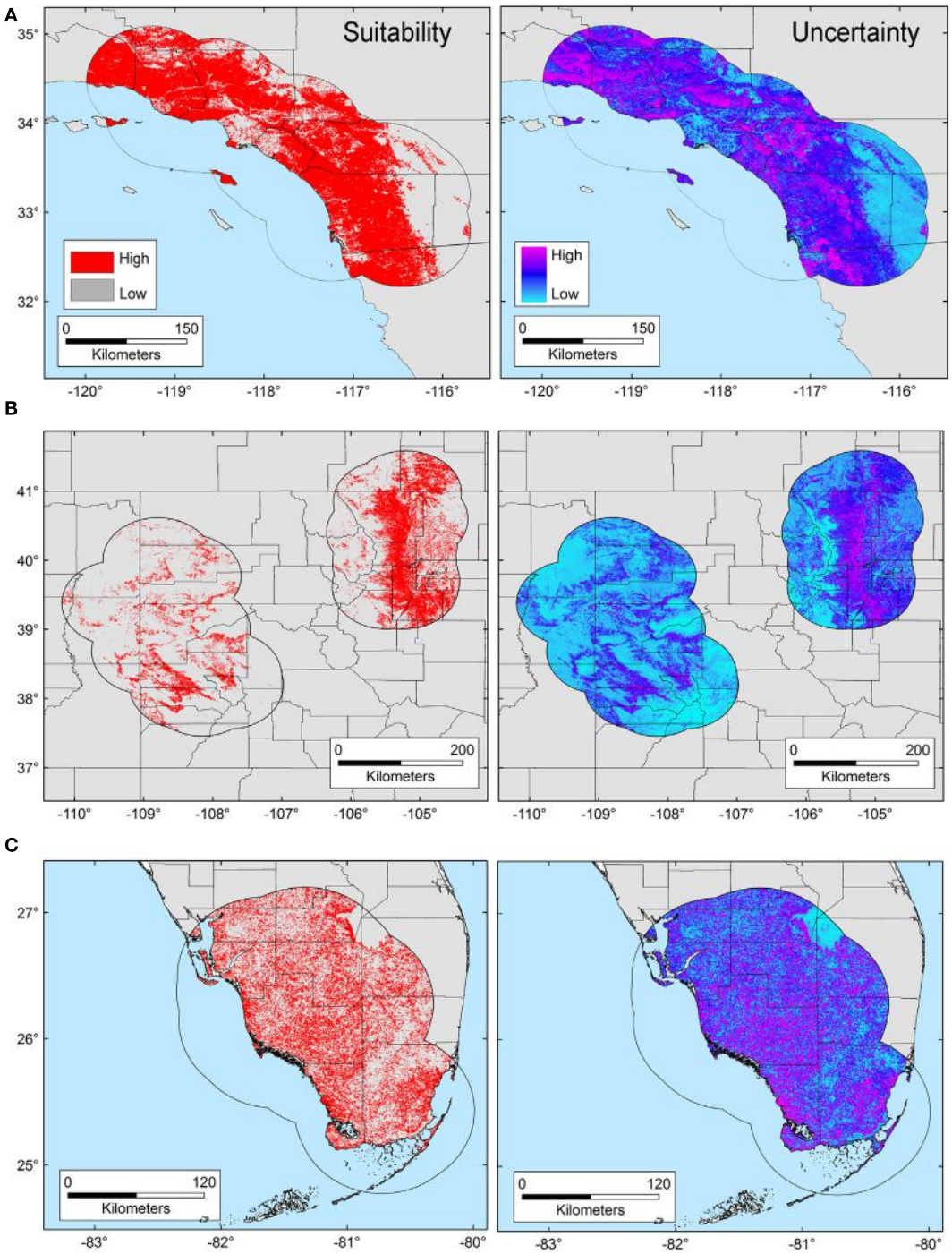


FIGURE 4 | Ecological niche model of *Bartonella* spp. Binary maps of *Bartonella* spp. suitability (red) were developed for areas in California **(A)**, Colorado **(B)**, and Florida [**(C)**; left panel]. Uncertainty estimations based on the suitability differences among models (right panel) show areas of low (cyan) and high (pink) uncertainty as follows: California **(A)** from 5.77×10^{-6} to 0.37, Colorado **(B)** from 2.75×10^{-6} to 0.39, and Florida **(C)** from 8.59×10^{-6} to 0.39.

For example, *T. gondii* oocysts can be viable in the environment for up to 18 months (49), or potentially more importantly for these large felids, in their prey (rodents, lagomorphs, and cervids). *Bartonella* spp. easily survive in fleas whose abundance is associated with increasing humidity (50), and with microclimate conditions indirectly represented by NDVI, which could determine the distribution of *Bartonella* spp. between wildlife and domestic reservoirs (10). The ENM framework used here, including freely available vegetation data, with the presence-background Maxent algorithm, has the potential to be used to explore other

environmental dependent pathogens. Our suitability maps of *T. gondii* and *Bartonella* spp. suggest that risk may exist in broad areas in the three states studied. Potential transmission may occur in the areas predicted suitable if hosts, the pathogen, and the vectors converge (**Figures 3** and **4**).

Despite the evident benefits, our approach is a simplification of two complex parasite systems. We based our interpretation of the pathogens' niche from infected wild felids (i.e., bobcats and puma) in their sylvatic habitats, but may be missing other pieces of the epidemiological triangle. For instance, the distribution of intermediate hosts for *T. gondii* (such as rodents, lagomorphs, and cervids) (13) was not included in our models, nor was the presence of domestic cats (another definitive host) owing to insufficient data across all study areas. For *Bartonella* spp., we did not account for presence of vectors (e.g., fleas) (10), and thus, even when we anticipate suitable conditions for the parasite occurrence, suitable conditions for vectors could limit the occurrence of *Bartonella* spp. in certain areas. Moreover, we modeled *Bartonella* spp. at genus level under the assumption of niche conservatism, which proposes that species phylogenetically close will share ecological niche characteristics, and that intraspecific differentiation of niches is challenging (51, 52). Although our diagnostics tests have proven effective for these wild felids, there could exist a small number of false-positive and false-negative results, we assumed that this proportion would not change the general patterns of the findings.

Ecological niche models of both parasite species based on hosts from wild areas revealed that our models were a proxy of the sylvatic cycle of both parasites; however, these pathogens might also occur in urban areas, which are not often frequented by puma or bobcats. *T. gondii* can also occur in urban environments given its adaptability and host generalization as a result of its broad ecological niche (53). Nonetheless, Lélu et al. (54) suggest *T. gondii* is likely to be less prevalent in urban areas owing to reduced transmission through the food chain, a conclusion supported by our work in these study areas where domestic cats are restricted to urban areas, wild felids avoid urban areas, and *T. gondii* has a higher prevalence in wild felids (12). Conversely, our previous research shows a strong positive relationship between urbanization and exposure to *Bartonella* spp. (12), suggesting that these bacteria persist in stable homogeneous urban landscapes. Future research should include the urban component in the distribution of *T. gondii* and *Bartonella* spp. parasites for a broader characterization of the ecological potential of both parasites in natural and impervious surfaces in North America.

Previous studies have demonstrated niches of pathogens independent of potential reservoir distributions (55), thus showing that the modeling of pathogens-only provides accurate forecasts of disease transmission risk. These cutting-edge available tools of disease modeling are worthy of exploration to generate further fine-scale hypotheses to advance our knowledge of the environmental component of infectious disease transmission chains (9). Although the occurrences of the two pathogens were explored in wildlife, they are also zoonotic, so the results of this study have implications for human, as well as domestic and other wild animals' health. NDVI and longitudinal epidemiological studies can help address questions not only about the prevalence of *Bartonella* spp. and *T. gondii* in the environment, but also can allow us to identify suitable habitats for their presence, and in turn, forecast into the future as these methods can incorporate the effects of land use change to understand the ecology of infectious diseases, particularly environmentally dependent forms, before outbreaks occur.

AUTHOR CONTRIBUTIONS

LE and DR-A designed the experiments, developed the analyses, and co-wrote the manuscript. SC and MC designed the study and co-wrote the manuscript. SV, KC, and ML provided the data and co-wrote the manuscript.

ACKNOWLEDGMENTS

LE thanks Andres Perez and Kim VanderWaal for the logistic support for DR-A to develop part of the analyses.

FUNDING

This research was funded by National Science Foundation's Ecology and Evolution of Infectious Diseases Research Program (NSF EF-0723676 and NSF DEB-1413925). MC was funded by the University of Minnesota's Office of the Vice President for Research and Academic Health Center Seed Grant. LE and DR-A thank the MiniGrant MF-0010-15 from the Institute of the Environment of the University of Minnesota, which supported the internship of DR-A in Minnesota.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/article/10.3389/fvets.2017.00172/full#supplementary-material.

REFERENCES

1. Craft ME, Caillaud D. Network models: an underutilized tool in wildlife epidemiology? *Interdiscip Perspect Infect Dis* (2011) 2011:1–12. doi:10.1155/2011/676949
2. Hudson PJ, Rizzoli A, Grenfell BT. *The Ecology of Wildlife Diseases*. New York: Oxford University Press (2002).
3. Godfrey SS. Networks and the ecology of parasite transmission: a framework for wildlife parasitology. *Int J Parasitol Parasites Wildl* (2013) 2:235–45. doi:10.1016/j.ijppaw.2013.09.001
4. Peterson AT. Biogeography of diseases: a framework for analysis. *Naturwissenschaften* (2008) 95:483–91. doi:10.1007/s00114-008-0352-5
5. Auchincloss AH, Gebreab SY, Mair C, Roux AVD. A review of spatial methods in epidemiology, 2000 – 2010. *Annu Rev Public Heal* (2012) 33:107–22. doi:10.1146/annurev-publhealth-031811-124655.A
6. Carpenter TE. The spatial epidemiologic (r)evolution: a look back in time and forward to the future. *Spat Spatiotemporal Epidemiol* (2011) 2:119–24. doi:10.1016/j.sste.2011.07.002
7. Escobar LE, Craft ME. Advances and limitations of disease biogeography using ecological niche modeling. *Front Microbiol* (2016) 7:1174. doi:10.3389/fmicb.2016.01174
8. Peterson AT, Soberón J, Pearson RG, Anderson RP, Martínez-Meyer E, Nakamura M, et al. *Ecological Niches and Geographic Distributions*. New Jersey: Princeton University Press (2011).

9. Peterson AT. *Mapping Disease Transmission Risk: Enriching Models Using Biology and Ecology*. Baltimore: Johns Hopkins University Press (2014).
10. Bevins SN, Carver S, Boydston EE, Lyren LM, Alldredge M, Logan KA, et al. Three pathogens in sympatric populations of pumas, bobcats, and domestic cats: implications for infectious disease transmission. *PLoS One* (2012) 7:e31403. doi:10.1371/journal.pone.0031403
11. Carver S, Scorza AV, Bevins SN, Riley SPD, Crooks KR, VandeWoude S, et al. Zoonotic parasites of bobcats around human landscapes. *J Clin Microbiol* (2012) 50:3080–3. doi:10.1128/JCM.01558-12
12. Carver S, Bevins SN, Lappin MR, Boydston EE, Lyren LM, Alldredge M, et al. Pathogen exposure varies widely among sympatric populations of wild and domestic felids across the United States. *Ecol Appl* (2016) 26:367–81. doi:10.1890/15-0445
13. Dubey JP, Jones J. *Toxoplasma gondii* infection in humans and animals in the United States. *Int J Parasitol* (2008) 38:1257–78. doi:10.1016/j.ijpara.2008.03.007
14. Breitschwerdt EB. Feline bartonellosis and cat scratch disease. *Vet Immunol Immunopathol* (2008) 123:167–71. doi:10.1016/j.vetimm.2008.01.025
15. Chomel BB, Kasten RW, Henn JB, Molia S. *Bartonella* infection in domestic cats and wild felids. *Ann N Y Acad Sci* (2006) 1078:410–5. doi:10.1196/annals.1374.080
16. Gilbertson ML, Carver S, VandeWoude S, Crooks KR, Lappin MR, Craft ME. Is pathogen exposure spatially autocorrelated? Patterns of pathogens in puma and bobcat. *Ecosphere* (2016) 7:1–27. doi:10.1002/ecs2.1558
17. Lagana DM, Lee JS, Lewis JS, Bevins SN, Carver S, Sweanor LL, et al. Characterization of regionally associated feline immunodeficiency virus (FIV) in bobcats (*Lynx rufus*). *J Wildl Dis* (2013) 49:718–22. doi:10.7589/2012-10-243
18. Troyer RM, Beatty JA, Stutzman-Rodriguez KR, Carver S, Lozano CC, Lee JS, et al. Novel Gammaherpes viruses in North American domestic cats, bobcats, and pumas: identification, prevalence, and risk factors. *J Virol* (2014) 88:3914–24. doi:10.1128/JVI.03405-13
19. Logan A, Sweanor LL. *Desert Puma: Evolutionary Ecology and Conservation of an Enduring Carnivore*. Washington, DC: Island Press (2001).
20. Riley SPD, Foley J, Chomel BB. Exposure to feline and canine pathogens in bobcats and gray foxes in urban and rural zones of a national park in California. *J Wildl Dis* (2004) 40:11–22. doi:10.7589/0090-3558-40.1.11
21. Jensen WA, Lappin MR, Kamkar S, Reagan WJ. Use of a polymerase chain reaction assay to detect and differentiate two strains of *Haemobartonella felis* in naturally infected cats. *Am J Vet Res* (2001) 62:604–8. doi:10.2460/ajvr.2001.62.604
22. Lappin MR, Jacobson ER, Kollias GV, Powell CC, Stover J. Comparison of serologic assays for the diagnosis of Toxoplasmosis in nondomestic felids. *J Zoo Wildl Med* (1991) 2:169–74.
23. Lappin MR, Powell CC. Comparison of latex agglutination, indirect hemagglutination, and ELISA techniques for the detection of *Toxoplasma gondii*-specific antibodies in the serum of cats. *J Vet Intern Med* (1991) 5:299–301. doi:10.1111/j.1939-1676.1991.tb03137.x
24. Barve N, Barve V, Jiménez-Valverde A, Lira-Noriega A, Maher SP, Peterson AT, et al. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecol Modell* (2011) 222:1810–9. doi:10.1016/j.ecolmodel.2011.02.011
25. Soberón J, Peterson AT. Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiv Inform* (2005) 2:1–10. doi:10.17161/bi.v2i0.4
26. Poo-Muñoz DA, Escobar LE, Peterson AT, Astorga F, Organ JF, Medina-Vogel G. *Galictis cuja* (Mammalia): an update of current knowledge and geographic distribution. *Iheringia Série Zool* (2014) 104:341–6. doi:10.1590/1678-476620141043341346
27. Pettorelli N. *The Normalized Difference Vegetation Index*. Oxford: Oxford University Press (2013).
28. Horning N, Robinson J, Sterling E, Turner W, Spector S. *Remote Sensing for Ecology and Conservation*. New York: Oxford University Press (2010).
29. Carroll ML, DiMiceli CM, Sohlberg RA, Townshend JRG. *250m MODIS Normalized Difference Vegetation Index, 250ndvi28920033435, Collection 4*. (2009). Available from: http://glcf.umd.edu/data/ndvi/
30. Pettorelli N, Vik JO, Mysterud A, Gaillard JM, Tucker CJ, Stenseth NC. Using the satellite-derived NDVI to assess ecological responses to environmental change. *Trends Ecol Evol* (2005) 20:503–10. doi:10.1016/j.tree.2005.05.011
31. ESRI. *ArcGIS Desktop: Release 10.3*. Redlands, CA: Environmental Systems Research Institute (2015).
32. Colwell R, Rangel T. Hutchinson's duality: the once and future niche. *Proc Natl Acad Sci U S A* (2009) 106:19651–8. doi:10.1073/pnas.0901650106
33. Qiao H, Peterson AT, Campbell LP, Soberón J, Ji L, Escobar LE. NicheA: creating virtual species and ecological niches in multivariate environmental scenarios. *Ecography* (2016) 39:1–9. doi:10.1111/ecog.01961
34. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. *Ecol Modell* (2006) 190:231–59. doi:10.1016/j.ecolmodel.2005.03.026
35. Merow C, Smith MJ, Silander JA. A practical guide to Maxent for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* (2013) 36:1058–69. doi:10.1111/j.1600-0587.2013.07872.x
36. Phillips SJ, Dudík M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* (2008) 31:161–75. doi:10.1111/j.2007.0906-7590.05203.x
37. Radosavljevic A, Anderson RP. Making better Maxent models of species distributions: complexity, overfitting and evaluation. *J Biogeogr* (2014) 41:629–43. doi:10.1111/jbi.12227
38. Warren DL, Seifert SN. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecol Appl* (2011) 21:335–42. doi:10.1890/10-1171.1
39. Owens HL, Campbell LP, Dornak LL, Saupe EE, Barve N, Soberón J, et al. Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. *Ecol Modell* (2013) 263:10–8. doi:10.1016/j.ecolmodel.2013.04.011
40. Escobar LE, Peterson AT, Favi M, Yung V, Pons DJ, Medina-Vogel G. Ecology and geography of transmission of two bat-borne rabies lineages in Chile. *PLoS Negl Trop Dis* (2013) 7:e2577. doi:10.1371/journal.pntd.0002577
41. Peterson AT. Niche modeling: model evaluation. *Biodiv Inform* (2012) 8:41. doi:10.17161/bi.v8i1.4300
42. Peterson AT, Papeş M, Soberón J. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecol Modell* (2008) 213:63–72. doi:10.1016/j.ecolmodel.2007.11.008
43. Barve N. *Tool for Partial-ROC*. (2008). Available from: http://kuscholarworks.ku.edu/dspace/handle/1808/10059
44. Gittleman JL, Harvey PH. Carnivore home-range size, metabolic needs and Ecology. *Behav Ecol Sociobiol* (1982) 10:57–63. doi:10.1007/BF00296396
45. Kelly M, Morin D, Lopez-Gonzalez CA. *Lynx Rufus*. The IUCN Red List of Threatened Species (2016). e.T12521A50655874. doi:10.2305/IUCN.UK.2016-1.RLTS.T12521A50655874.en
46. Nielsen C, Thompson D, Kelly M, Lopez-Gonzalez CA. *Puma Concolor*. The IUCN Red List of Threatened Species (2015). e.T18868A97216466. doi:10.2305/IUCN.UK.2015-4.RLTS.T18868A50663436.en
47. Beldomenico PM, Chomel BB, Foley JE, Sacks BN, Baldi CJ, Kasten RW, et al. Environmental factors associated with *Bartonella vinsonii* subsp. *berkhoffii* seropositivity in free-ranging coyotes from northern California. *Vector Borne Zoonotic Dis* (2005) 5:110–9. doi:10.1089/vbz.2005.5.110
48. Jiyipong T, Morand S, Jittapalapong S, Rolain J-M. *Bartonella* spp. infections in rodents of Cambodia, Lao PDR, and Thailand: identifying risky habitats. *Vector Borne Zoonotic Dis* (2015) 15:48–55. doi:10.1089/vbz.2014.1621
49. VanWormer E, Fritz H, Shapiro K, Mazet JAK, Conrad PA. Molecules to modeling: *Toxoplasma gondii* oocysts at the human–animal–environment interface. *Comp Immunol Microbiol Infect Dis* (2013) 36:217–31. doi:10.1016/j.cimid.2012.10.006
50. Jameson P, Greene C, Regnery R, Dryden M, Marks A, Brown J, et al. Prevalence of *Bartonella henselae* antibodies in pet cats throughout regions of North America. *J Infect Dis* (1995) 172:1145–9. doi:10.1093/infdis/172.4.1145
51. Peterson AT, Soberón J, Sánchez-Cordero V. Conservatism of ecological niches in evolutionary time. *Science* (1999) 285:1265–7. doi:10.1126/science.285.5431.1265
52. Tocchio LJ, Gurgel-Gonçalves R, Escobar LE, Peterson AT. Niche similarities among white-eared opossums (Mammalia, Didelphidae): is ecological niche modelling relevant to setting species limits? *Zool Scr* (2014) 44:1–10. doi:10.1111/zsc.12082
53. Flegr J, Prandota J, Sovičková M, Israili ZH. Toxoplasmosis – a global threat. Correlation of latent toxoplasmosis with specific disease burden in a set of 88 countries. *PLoS One* (2014) 9:e90203. doi:10.1371/journal.pone.0090203
54. Lélu M, Langlais M, Poulle M-L, Gilot-Fromont E. Transmission dynamics of *Toxoplasma gondii* along an urban–rural gradient. *Theor Popul Biol* (2010) 78:139–47. doi:10.1016/j.tpb.2010.05.005

55. Maher SP, Ellis C, Gage KL, Enscore RE, Peterson AT. Range-wide determinants of plague distribution in North America. *Am J Trop Med Hyg* (2010) 83:736–42. doi:10.4269/ajtmh.2010.10-0042

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Veterinary Science

ORIGINAL RESEARCH
published: 19 January 2017
doi: 10.3389/fvets.2017.00002

Using Machine Learning to Predict Swine Movements within a Regional Program to Improve Control of Infectious Diseases in the US

Pablo Valdes-Donoso[1,2], Kimberly VanderWaal[1], Lovell S. Jarvis[2], Spencer R. Wayne[3] and Andres M. Perez[1]*

[1] Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota, St. Paul, MN, USA, [2] Department of Agricultural and Resource Economics, University of California Davis, Davis, CA, USA, [3] Veterinary Services Pipestone, Pipestone, MN, USA

Between-farm animal movement is one of the most important factors influencing the spread of infectious diseases in food animals, including in the US swine industry. Understanding the structural network of contacts in a food animal industry is prerequisite to planning for efficient production strategies and for effective disease control measures. Unfortunately, data regarding between-farm animal movements in the US are not systematically collected and thus, such information is often unavailable. In this paper, we develop a procedure to replicate the structure of a network, making use of partial data available, and subsequently use the model developed to predict animal movements among sites in 34 Minnesota counties. First, we summarized two networks of swine producing facilities in Minnesota, then we used a machine learning technique referred to as random forest, an ensemble of independent classification trees, to estimate the probability of pig movements between farms and/or markets sites located in two counties in Minnesota. The model was calibrated and tested by comparing predicted data and observed data in those two counties for which data were available. Finally, the model was used to predict animal movements in sites located across 34 Minnesota counties. Variables that were important in predicting pig movements included between-site distance, ownership, and production type of the sending and receiving farms and/or markets. Using a weighted-kernel approach to describe spatial variation in the centrality measures of the predicted network, we showed that the south-central region of the study area exhibited high aggregation of predicted pig movements. Our results show an overlap with the distribution of outbreaks of porcine reproductive and respiratory syndrome, which is believed to be transmitted, at least in part, though animal movements. While the correspondence of movements and disease is not a causal test, it suggests that the predicted network may approximate actual movements. Accordingly, the predictions provided here might help to design and implement control strategies in the region. Additionally, the methodology here may be used to estimate contact networks for other livestock systems when only incomplete information regarding animal movements is available.

Keywords: swine industry, pig movements, regional control programs, Minnesota, random forest, social network analysis

OPEN ACCESS

Edited by:
Salome Dürr,
University of Bern, Switzerland

Reviewed by:
Marco De Nardi,
Safoso, Switzerland
Hartmut H. K. Lentz,
Friedrich Loeffler Institute, Germany
Vitaly Belik,
Freie Universität Berlin, Germany

****Correspondence:***
Pablo Valdes-Donoso
pablovd@umn.edu

Specialty section:
This article was submitted to Veterinary Epidemiology and Economics, a section of the journal Frontiers in Veterinary Science

Received: *12 September 2016*
Accepted: *04 January 2017*
Published: *19 January 2017*

Citation:
Valdes-Donoso P, VanderWaal K, Jarvis LS, Wayne SR and Perez AM (2017) Using Machine Learning to Predict Swine Movements within a Regional Program to Improve Control of Infectious Diseases in the US. Front. Vet. Sci. 4:2. doi: 10.3389/fvets.2017.00002

INTRODUCTION

Between-farm direct or indirect contact *via* movement of animals or biological materials (e.g., semen), or cross-contamination through inputs such as machinery or human workers, is among the most important factors contributing to disease spread in food animals (1). Farm-to-farm contacts spread diseases that affect the US swine industry, including porcine reproductive and respiratory syndrome (PRRS) and porcine epidemic diarrhea (PED). For both PRRS and PED, animal movements (e.g., gilts, boars, weaned pigs, feeder pigs, and cull animals) represent one of the most important disease transmission routes between farms (2–6).

Understanding the network structure of food animal industries is critical for efficient production and disease control. For example, the sharing of information among agents (e.g., farmers, suppliers, and brokers) within a network may result in an increase of economic efficiency due to the selection of strategies that can decrease production and/or transaction costs (7). Indeed, social network analysis (SNA) is an analytical tool that has been widely used in the field of veterinary medicine to design disease control plans (8). SNA has been used to quantify the nature of connections (referred to as *edges* or *contacts*) among elements (*nodes* or *vertices*) in a population (9). Nodes may be farms or other facilities (e.g., slaughter houses, truck wash disinfection stations, or feed plants) from, to, or through which, animal populations are connected, and contacts among nodes may be categorized as direct or indirect (8, 10). SNA enables researchers to better understand animal movement patterns and, consequently, provide insights on how diseases diffuse in a given industry (11–13). For example, in many livestock industries, a minority of farms typically account for the majority of animal movements (14–16). Identification of those few farms, often referred to as "hotspots" or "super-spreaders" for disease transmission, may help formulating contingency plans to control high impact diseases, as timely intervention to targeted farms may enhance the probability of such plans being successful (13–15, 17). Similarly, efforts to improve animal management and biosecurity in super-spreaders may also contribute to reducing disease risk and prevalence (17, 18).

The US swine industry is characterized by large numbers of documented pig movements within and between states and regions. From 1970 to 2001, the number of pigs moved from one state to another (or from Canada to the US) increased from 30 to 50 million (19). Increases in the number and distance of movements reflect growth in the number of farms specializing in specific phases of the production cycle. Indeed, a growing proportion of feeder and finishing swine farms are located in the Midwest in close proximity to the grain used to feed pigs (20). In contrast, breeding populations tend to be located in areas distant from the major growing pig regions, such as the southeastern US, where grain inputs are not as critical (20–22). Whereas the regional specialization of different industry components has undoubtedly improved efficiency, the necessary movement of animals between the two regions alters the risk of long-distance disease spread (1, 22–24).

Animal movements are only partially regulated in the US, and no source provides complete information on such movements. For example, the United State Department of Agriculture, through the animal disease traceability program, collects information on movements of cattle, bison, equines, sheep and goats, swine, and poultry, only when movements cross state boundaries, except when livestock are moved to slaughter facilities or chicks moved from hatcheries (25). The lack of movement data creates a particular problem for the control of diseases, such as PRRS. In that context, regional control programs (RCPs), voluntarily organized and coordinated by producers, serve as means to share sanitary status information among farmers located in a given area. Sharing information within an RCP in Minnesota (RCP-N212) has been correlated with a decrease in PRRS incidence (26), and thus, one may hypothesize that sharing additional information about pig movements would further improve control program effectiveness. Unfortunately, lack of information about between-farm movements hinders attempts to describe network structure, hence impairing ability to prevent and control disease.

To elucidate the role of network structure in the spread of swine diseases, the relation between PRRS manifestation and animal movements between farms (and other related sites, such as buyer stations or market sites) was assessed in two counties in Minnesota (27). A positive association between positive PRRS status and the number of direct and indirect suppliers (in-reach degree) was observed in one county, but no additional network measures were significantly correlated with positive PRRS status (27). Although that early study provided valuable insights about pig movements between sites and their potential contribution to disease spread, a more complete assessment of the structure of contacts is required to understand disease spread. We use data from Wayne (27) and more recent data collected by the RCP-N212 to build a predictive movement model between sites, which is then used to estimate a complete movement network for the RCP-N212 in Minnesota. The results may be incorporated into a disease-spread model to help explain disease dynamics and support disease prevention and control activities within the RCP-N212.

MATERIALS AND METHODS

Data Sources

We used two complementary sets of data to construct our model. The first dataset, referred to as the network building data set, included information on pig movements related to two counties being used to fit the model, whereas the second dataset included information on sites located within the broader RCP-N212 area, which was used for prediction purposes. The first dataset included information collected in two counties that were geographically located within the boundaries of the second dataset; however, the two datasets were collected separately. The first data set was based on surveys conducted with owners, managers, and veterinarians on farms and at market sites located in Stevens and Rice counties, Minnesota in 2006 (27). Animal movement data included origin and destination of sites in and out of Stevens and Rice, geographic locations, and the production type of sites and owner. Production types included boar stud (BS), farrowing

(Fa), nursery (N), finishing (Fi) farms, and market sites (M). This last type encompasses buying stations or slaughter plants. Two networks were described in the building dataset, one for each county, i.e., a Stevens network (SN) and a Rice network (RN). Each network contained data on directional animal movements between any given site located within the county and a number of sites located either inside or outside the county.

The second data set, referred to as RCP-N212, contained information on geographical location, owner, and type of site for premises enrolled in the RCP-N212. This data set contains roughly 38% of total swine premises with 100 or more animals located in Minnesota (28). Data were collected between 2012 and 2015. The RCP-N212 comprised 34 counties in Minnesota, including Stevens and Rice counties (26). The University of Minnesota manages the RCP-N212 data under the terms of an agreement with swine producers that protects the confidentiality of the data.

Network Description

The structures of SN and RN were described using SNA representing directional flows of animal movements between sites. The site-level connectivity of each network was described using *in-* and *out-degree*, calculated as the number of pig movements received or sent by a specific site to or from other sites. *Betweenness*, defined as the number of directed paths that pass through a given site, when the shortest paths between other pairs of sites are traced (9), was also estimated. Metrics were stratified by site type (e.g., BS, Fa, N, Fi, and M) for SN and RN, and differences in centrality measures between types were analyzed using Kruskal–Wallis tests. To assess the correlation between types of sites in each network, the assortativity coefficient (r) for a mixing matrix was used, as defined by elsewhere (29), so that

$$r = \frac{\sum_i e_{ij} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

where e_{ij} is the fraction of animal movements in the network that connects sites of type i to type j, and a_i and b_i are fractions of destination-or-origin, respectively, of a movement that is attached to site type i. A value of $r = 0$ indicates no assortative mixing or a random network, while a value of $r = 1$ indicates complete assortativity, i.e., that all movements are between sites of the same production type. Alternatively, if $r < 0$ links are more likely to connect two different types of nodes, which is closer to having a randomly mixed network where links often connect unlike nodes (e.g., different type of sites) (29).

Four metrics at the network level were estimated, namely, (1) *network density*, calculated as the fraction of movements that are present in the network relative to the total number possible; (2) *clustering coefficient*, used as a measure of cohesiveness and defined as the probability that two sites that are linked to a common site are also linked to each other; (3) *diameter*, calculated as the largest distance between two sites in the network, where distance is the shortest path between two sites; and (4) the *mean path length*, calculated as the mean length of the shortest paths connecting two sites (9, 30).

Figures and statistical computations were preformed using R V.3.1.1 (31), including the packages *ggplot2* (32), *maps* (33), *MASS* (34), and *igraph* (35).

Network Prediction

Due to the inherent attributes of nodes, their dimensional distribution, connection features, etc. predicting networks can be challenging (36, 37). Unsupervised and supervised methods have been used to try to elucidate network structures. Unsupervised approaches seek to assign scores to possible links between nodes mainly based on the neighborhood characteristics of each node and path-distances between nodes. While the first estimates the likelihood of a link between two nodes based on the degree of overlap of their neighbors, the second searches for the shortest path-distance among all possible combinations between nodes (36). For example, the preferential attachment prediction has been used to estimate potential connections of a node given the proportional number of neighbors that it has (38), or the Katz coefficient scores the possible links between two nodes subject to a given length paths (39). While unsupervised methods have been popular in network prediction, they fail to handle network dynamics, the mutual dependence of components, and other features inherent of the network structure (e.g., an unbalanced number of links), thus often leading to unstable performance (37). Among supervised methods, random forest (RF) has shown high levels of classification accuracy compared to others techniques such as bagging (37, 40), so it is the approach used here. Here, information provided by SN and RN was used in a RF model to predict animal movements between sites. After predictions were obtained, parameters were extrapolated to predict movements for the entire number of farms within the RCP-N212.

RF Model

Models based on classification trees are built using a single rule or a set of rules for a number of variables that split data to predict possible outcomes. A RF is an ensemble of independent classification trees created from bootstrap samples chosen with replacement from a training data set, in which aggregated estimates from each ensemble generates a final prediction of the probability that a given outcome occurs (40, 41), e.g., a link between two sites. The samples that are not selected as bootstrap samples are called "out-of-bag" (OOB) samples and are used to estimate the error rate. The OOB error rate is reduced by ranking predictors and subsequently removing those considered less important. Calculating the difference in accuracy between models in which predictors are present or removed is used to assess predictor importance. Differences are normalized across all trees generated and then ranked based on accuracy of prediction (40, 41).

Using all sites from SN and RN, we created a new dataset (referred to as RF-data) that contained all possible origin-destination pairs of sites within each network. Per each possible pair of sites, we assigned a dichotomous outcome (yes, no) variable (also referred to as a class variable) indicating whether or not the animal movement has occurred between that pair. We used the geographical location of each site to estimate the pairwise Euclidean distance (kilometers) between farms, and generated a

dichotomous variable (yes, no) indicating whether or not each pair had a common owner. Additionally, we generated 25 dummy variables, each denoting a possible pair of site types (e.g., Fa–Fi, Fi–M, BS–Fa, etc.), being 1 if the pair site type combination was true and 0 otherwise.

The effectiveness of model prediction is determined using a portion of the data that has not been used to build and tune the model (40). Thus, we split the RF-data randomly, using 75% of observations to build and tune the model (referred to as the training dataset), and the remaining 25% to test or validate our model (referred to as the testing dataset or validation set). In other words, we used the training dataset to create (i.e., train and tune) the RF model and then used the testing dataset to qualify its performance through a confusion matrix: a two by two table displaying the number of observed and predicted movements reported from the model (**Table 1**). While there is no widely accepted rule-of-thumb for splitting the data, it is preferred to use a larger amount of information for the training set in order to reduce the variance of the parameter estimates (40). Also, we insured that the training dataset contained the same proportion of class variables (yes and no) as the original RF-data by using a data partition function executed by the *caret* package (42) in R (31).

On the other hand, because we anticipated that RF-data would be unbalanced (i.e., only a small fraction of observations were class variable "yes"), a *post hoc* down-sampling approach was implemented to balance the data, i.e., we used a sample that has roughly the same proportion of each outcome class. The down-sampling technique is an efficient way to improve predictions, particularly when using bootstrap samples, given that no information is lost during the process (40, 43). We used a wrapper provided by the *train* function in the *caret* package (42) in R (31) to improve model consistency and to determine the desired standard resampling and performance testing (40, 44). We ran and tuned the RF model using 1,500 trees for each training dataset (unbalanced and balanced), and we implemented 10-fold cross-validations to estimate and rank the most important predictors.

Subsequently, we compared performance comparing predictive (or expected) versus observed movements for both, unbalanced and balanced testing datasets by using their confusion matrixes. As result, we compared the accuracy, Kappa statistic, specificity, sensitivity, and the area under the receiver-operating characteristic (ROC) curve. The ROC curve is a graphical method to test predictive performance by contrasting true positive and negative values. The accuracy rate $\left(\mathrm{AR}=\frac{a+d}{N}\right)$ was used to measure the agreement between the predicted and the observed classes, although AR does not provide any information on the type of error the model is producing. The Kappa statistic (κ) was used to quantify the relation between observed $\left(O=\frac{a+d}{N}\right)$ and expected accuracy $\left(E=\frac{\left((d+b)*(d+c)\right)+\left((c+a)*(b+a)\right)}{N^2}\right)$, so that $\kappa=\frac{O-E}{1-E}$ serves as a proxy for model performance (40). Sensitivity $\left(\mathrm{Se}=\frac{a}{a+c}\right)$ and specificity $\left(\mathrm{Sp}=\frac{d}{b+d}\right)$ were used to measure the capability of the model to predict true movements (i.e., "yes") and non-movements (i.e., "no"), respectively, whereas the area under the ROC (AUC) was used to assess the trade-off between increasing sensitivity and decreasing specificity or vice versa. With RF, the final prediction as to whether movement occurs between a pair of sites (i.e., class variable = yes) is based on a given probability threshold (i.e., 0.5). We tested varying threshold probabilities (i.e., ≥0.5) to maximize the κ value.

TABLE 1 | Confusion matrix for the class variable (i.e., animal movements = yes or no).

Predicted	Observed		Total
	Yes	No	
Yes	a	b	a + b
No	c	d	c + d
Total	a + c	b + d	*N*

Cells indicate the number of true positives (a), false positives (b), true negatives (d), and false negatives (c).

Our RF model utilized a data set of 14,307 observations (75% of all observations) and 28 variables, thus complexity of the algorithm is given by $O(v \times n\log(n))$, where v is the number of variables and n is the number of observations. This analysis took around 45 min to complete on a standard MacBook Pro®, though other packages such as ranger and random jungle may achieve faster performances for larger data sets and down-sampling can further optimize run times (45).

Finally, using the observed and predicted animal movements, we conducted an SNA to contrast centrality measurements between the observed and predicted network using the non-parametric Kruskal–Wallis test.

Prediction of the RCP-N212 Network

We used our final model to predict animal movements among sites in the RCP-N212. Using data from RCP-N212, we generated all possible pair combinations among sites located within that RCP area. Similar to the analyses performed with the RF-data, we estimated Euclidean distances (kilometers) between each pair of sites and generated a dichotomous (yes, no) variable for ownership and 25 dummy variables for possible combinations of types of sites. Acknowledging that movements of animals must also occur to and from sites located out of the RCP-N212 and to avoid overestimations in the number of movements occurring within sites in the RCP, we restricted the number of predicted movements among sites in the RCP-N212 using the maximum values of *in-* and *out-degree* per each type of site observed in SN and RN.

We summarized the distributions of centrality measures of the predicted network for the entire RCP-N212 and for each of its 34 counties. We used the same metrics as described in Section "Network Description" at site and network levels. We performed a spatial analysis of centrality measures using a 2D-kernel

density estimation. This allowed us to evaluate the intensity of pig movements (e.g., to, through, and from other sites) in a given unit of space by approximating its probability density function (34, 46, 47).

RESULTS

Network Description

The network building data set included 237 sites (220 farms and 17 market sites), of which 33 and 19% were located within Stevens County and Rice County, respectively. The remaining 48% of sites were not located within those counties, but involved animal movements to or from Stevens and Rice (**Table 2**), some covering long distances (**Figure 1**). We identified 474 animal movements (286 movements in SN and 215 movements in RN), some of which connected the two networks (**Table 2**). Only 14% of all site types located in Stevens or Rice had movements with sites located in the other county, and all these were movements from finishing farms to market sites.

TABLE 2 | Number of sites by production type and network.

Network	BS		Fa		N		Fi		M		Total	
Rice network	0	(0)	21	(8)	18	(7)	52	(28)	6	(1)	97	(44)
Stevens network	3	(1)	43	(20)	12	(7)	43	(23)	6	(1)	107	(52)
Both	0	(0)	0	(0)	0	(0)	28	(27)	5	(2)	33	(29)
Total	3	(1)	64	(28)	30	(14)	123	(78)	17	(4)	237	(125)

Values in parentheses indicate number of sites inside the noted county.
BS, boar stud; Fa, farrowing; N, nursery; Fi, finishing; M, market sites.

Graphical representations of the networks indicate a confluence of paths toward finishing farms and then to market sites (**Figure 2**). Thus, markets and finishing farms served as hubs in the network. As expected, the most likely movements occurred between sites of different types ($r = -0.13$ and $r = -0.16$ for SN and RN, respectively) that followed downstream flows, i.e., a vertical structure (**Table 3**). For example, movements from farrowing (Fa) or nursery (N) farms to finishing farms (Fi) were more frequent compared to other possible types of destinations, i.e., market sites (M), boar studs (BS), farrowing (Fa), or nursery (N) farms (**Table 3**). Markets were the most likely destinations for finishing farms ($e_{\mathrm{FiM}} = 0.40$ and $e_{\mathrm{FiM}} = 0.41$ for SN and RN, respectively), although finishers also sent pigs into upstream destinations, including nurseries and farrowing farms (e.g., N, Fa, etc.), probably to provide replacement animals (**Table 3**). The most likely destination for farrowing farms was finishers, followed by nurseries, consistent with the industry trend to eliminate nurseries as midpoint stations (20) (**Figure 2**; **Table 3**).

Whereas *in-degree* and *betweenness* were slightly higher in SN than RN ($P = 0.05$ and $P = 0.04$, respectively), there was no statistical difference between the two networks in out-degree ($P = 0.31$). In contrast, *in-degree* varied across production types for both networks ($P < 0.01$ for both), with markets having a higher *in-degree* (mean = 11.7, SD = 16.8, min = 0 and max = 57) (**Figure 3**). Nurseries exhibited significantly higher out-degree than other production types ($P = 0.01$ for SN and $P < 0.01$ for RN), each shipping animals to three different sites on average, with a maximum of 12. *Betweenness* did not significantly differ across production types within RN ($P = 0.17$) but was statistically different among different types of sites in SN ($P = 0.02$) (**Table 4**).

Both networks exhibited similar densities, 0.014 for SN and 0.013 for RN. However, RN was relatively more cohesive than SN,

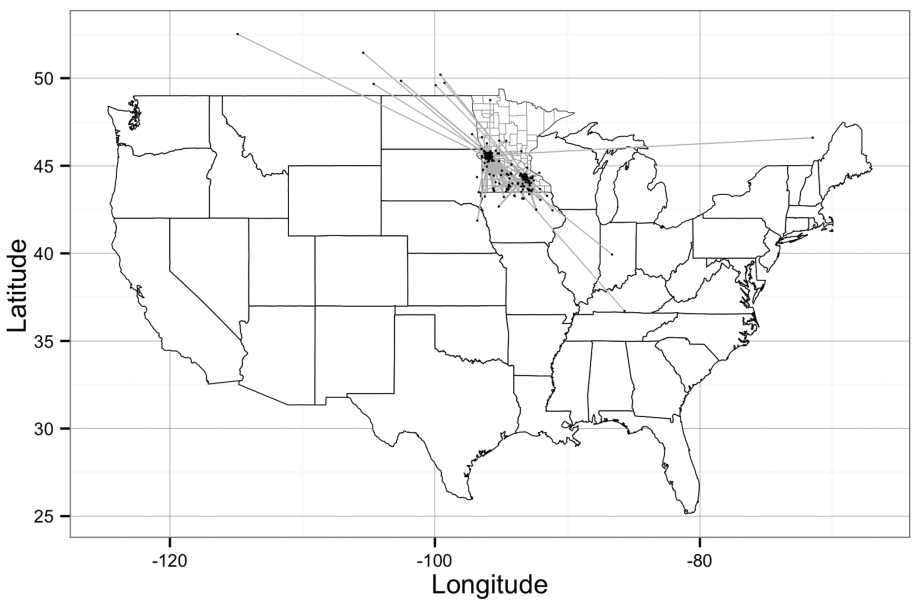


FIGURE 1 | Geographical representation of the Stevens network and Rice network of animal movements between swine farms and/or market sites. Dots represent geographical location of sites and straight gray lines represent animal movements. [Source: Wayne (27)].

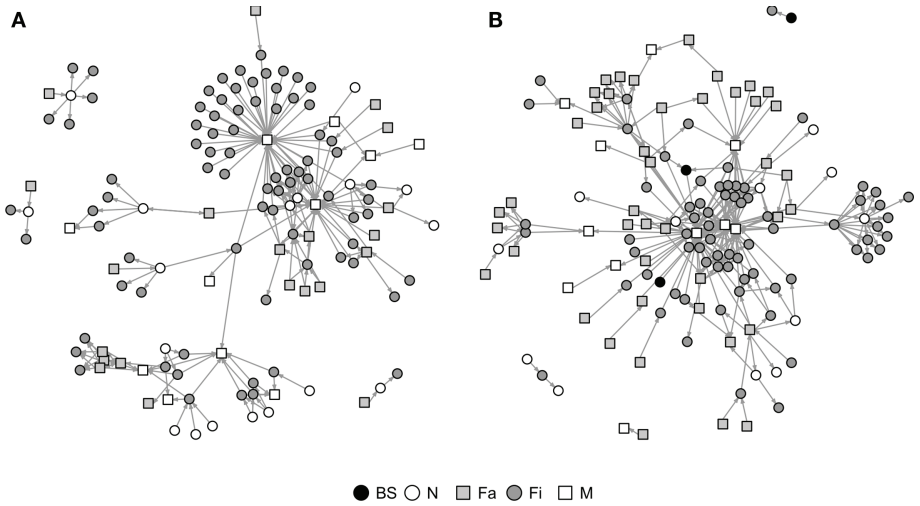


FIGURE 2 | Graphical representation of Rice (A) and Stevens networks (B). Circles and squares represent either swine farms (BS, boar stud; Fa, farrowing; N, nursery; and Fi, finishing) or market sites (M).

TABLE 3 | Mixing matrix e_{ij} for type of farm in Rice network (RN) and Stevens network (SN).

Network		SN						RN					
Destination		**BS**	**Fa**	**Fi**	**M**	**N**	a_i	**BS**	**Fa**	**Fi**	**M**	**N**	a_i
Origin	BS	0.00	0.00	0.00	0.01	0.00	0.01	–	–	–	–	–	–
	Fa	0.00	0.03	0.09	0.10	0.04	0.26	–	0.06	0.12	0.05	0.05	0.27
	Fi	0.01	0.10	0.07	0.40	0.00	0.58	–	0.03	0.00	0.41	0.00	0.45
	M	0.00	0.00	0.00	0.02	0.00	0.02	–	0.00	0.00	0.02	0.00	0.02
	N	0.00	0.00	0.13	0.00	0.00	0.13	–	0.00	0.24	0.01	0.00	0.26
b_i		0.01	0.14	0.29	0.53	0.04	1.00	–	–	0.36	0.50	0.05	1.00

BS, boar stud; Fa, farrowing; N, nursery; Fi, finishing; M, market sites.
$r_{SN} = -0.13$ *and* $r_{RN} = -0.16$.

as shown by a higher clustering coefficient, revealing a 0.007 and a 0.056 probability, respectively, that two sites moving animals to a common site were also connected to each other. As a result, RN also had a smaller diameter (4) than SN (5), though mean path lengths were relatively similar (1.82 for RN and 1.85 for SN). In turn, the distances between sites varied considerably, from less than 1 to more than 1,000 km. The overall mean distance between sites was 111 km, with nurseries and farrowing farms receiving animals from longer distances and boar studs shipping animals to sites located more than 500 km away (**Table 5**).

Network Prediction

Random Forest

There were 19,075 possible pairs for RN and SN. Among them, a minority (2.6%) corresponded to true movements (i.e., class = yes). RF models for both balanced (similar proportion of class variable "yes" and "no") and unbalanced datasets used 1,500 trees, and the optimal number of predictors (m_{try}) estimated was 27 and 20, respectively. We observed a higher κ for the unbalanced dataset, indicating a higher accuracy (**Table 6**). However, use of the unbalanced datasets resulted in predictions that were strongly biased toward the majority class, with the class variable "no" accounting for 97.4% of total pairs.

The balanced dataset optimized sensitivity with a low penalty to specificity. Moreover, inspection of the AUC indicated that false positives and negatives were minimized with the balanced dataset (**Figure 4**). However, the 0.5 default probability threshold used by the RF to predict animal movement between a pair of sites (i.e., class variable = yes) resulted in low agreement (i.e., when $\kappa < 0.3$) between observed versus predicted movements (40) (**Table 6**). Increasing the threshold from 0.5 to 0.85 resulted in an increase in agreement ($\kappa = 0.5$, **Figure 5**) between observed and predicted movements. The most important variables predicting movements were farm type (downstream combinations from finishers and farrowing farms to market sites), sharing the same owner, and distance (**Figure 6**).

Comparing the observed (*O*) and predicted (or expected, *E*) networks based on observed and predicted animal movements from use of the *testing dataset*, model predictions provided a reasonable approximation of real movements (**Figure 7**). Overall, there were no statistical differences between both networks in *betweenness* ($P = 0.38$) and *in-degree* ($P = 0.97$), whereas values

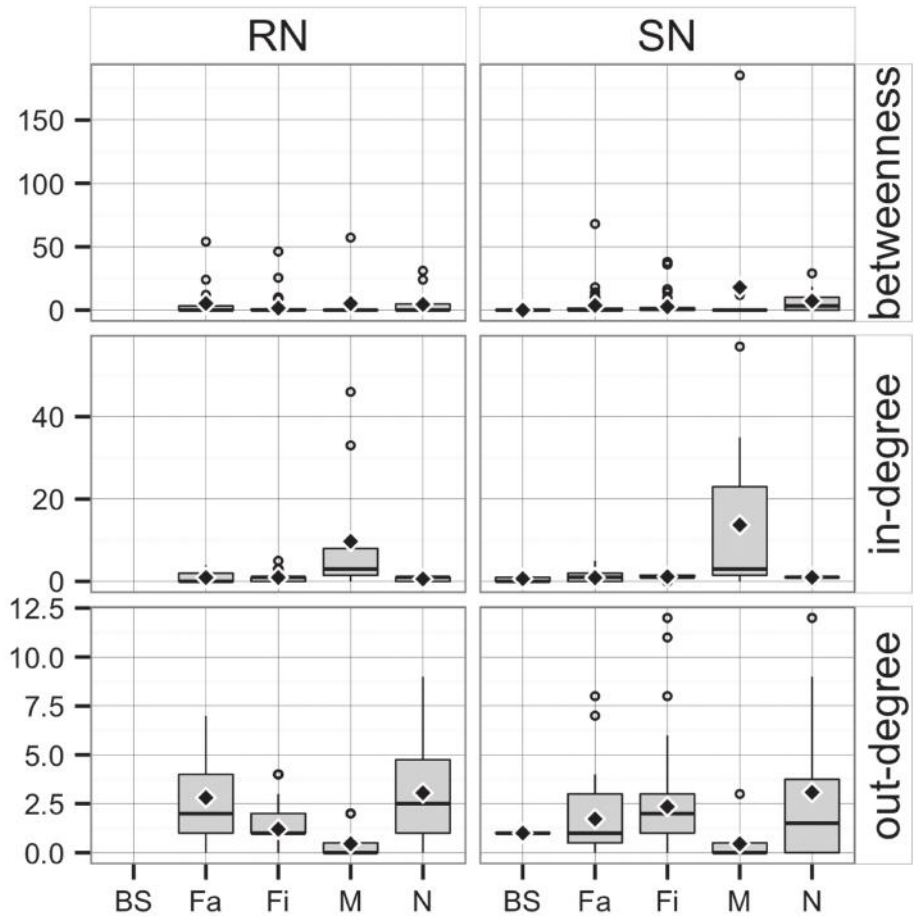


FIGURE 3 | Boxplot of betweenness, in- and out-degree in Rice network and Stevens network by production type (BS, boar stud; Fa, farrowing; N, nursery; Fi, finishing; M, market sites). Boxes indicate the first and third percentile; middle bars represent the median and diamonds represent the mean.

TABLE 4 | Summary of centrality measures at site-level using both Stevens network and Rice network.

Centrality measure	BS	Fa	Fi	M	N
Betweenness	0.00	4.20	2.07	11.56	5.61
	(0)	(1.43)	(0.51)	(8.67)	(1.66)
	0[a]	68[a]	46.21[a]	185.14[a]	31[a]
In-degree	0.67	0.92	1.05	11.73	0.77
	(0.67)	(0.14)	(0.07)	(3.59)	(0.1)
	2[a]	5[a]	5[a]	57[a]	2[a]
Out-degree	1.00	2.08	1.74	0.46	3.07
	(0)	(0.26)	(0.15)	(0.18)	(0.62)
	1[a]	8[a]	12[a]	3[a]	12[a]

Values denote means, SEs in parenthesis, and maximums with superscript "a."
BS, boar stud; Fa, farrowing; N, nursery; Fi, finishing; M, market sites.

TABLE 5 | Summary of distances (km) between origin and destination by type of site.

Type destination		BS	Fa	Fi	M	N	Mean
Type origin	BS			1,894.81	20.36		645.18
				(–)	(13.68)		
				1,894.81[a]	34.04[a]		
	Fa		113.71	122.50	32.95	170.83	102.98
			(46.72)	(38.56)	(7.78)	(61.61)	
			700.64[a]	1,582.98[a]	214.84[a]	1,066.67[a]	
	Fi	9.94	181.90	98.16	122.97	21.51	128.49
		7.65	(43.15)	(22.28)	(8.75)	(–)	
		17.58[a]	1285.83[a]	271.21[a]	354.65[a]	21.51[a]	
	M				267.49		267.49
					(49.57)		
					432.79[a]		
	N		126.73	48.98	42.12	22.08	50.21
			(101.77)	(7.73)	(30.68)	(–)	
			228.50[a]	270.83[a]	72.81[a]	22.08[a]	
Mean		9.94	155.76	90.30	109.95	158.91	111.14

Values denote means, SEs in parenthesis, and maximums with superscript "a."
SEs could not be calculated in all cases due to small sample size.
BS, boar stud; Fa, farrowing; N, nursery; Fi, finishing; M, market sites.

for the *out-degree* were significantly different ($P = 0.02$). For the latter, we predicted that, on average, a site would deliver animals to 1.4 sites (SD = 1.3), compared to 1 site (SD = 1.0) observed in the real network.

Furthermore, the patterns of connectivity across farm types were qualitatively similar (**Figure 7**). Whereas comparisons

TABLE 6 | Results of the random forest (RF) analyses for balanced and unbalanced datasets.

Model	Accuracy	Kappa	Sensitivity	Specificity	Area under receiver-operating characteristic
Balanced RF model	0.88	0.23	0.808	0.883	0.93
Unbalanced RF	0.98	0.34	0.232	0.997	0.85

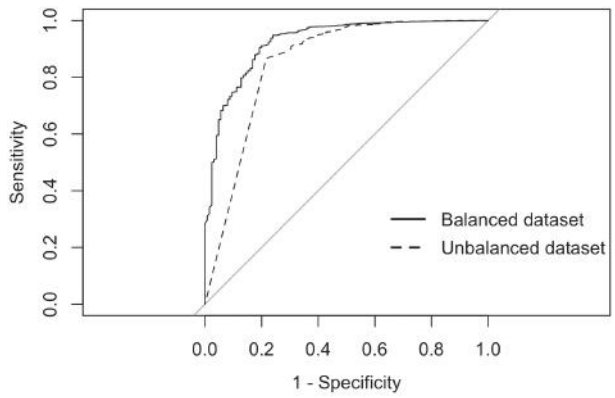


FIGURE 4 | Receiver-operating characteristic curves for the RF analyses using balanced and imbalanced datasets.

across production types within observed and predicted networks did not show significant differences in *betweenness* ($P = 0.32$ and $P = 0.35$, respectively), *in-* and *out-degree* were statistically different across production types in both observed ($P < 0.01$ for both) and predicted networks ($P < 0.01$, and $P = 0.01$, respectively). For example, market sites in the observed data received animals from 7.4 different sites, whereas the model predicted receptions from 8.4 different sites. Similarly, whereas the model predicted that a nursery would ship animals into 2.1 farms, observed values indicated 1.7 different farms (**Figure 7**). On the other hand, there were no significant differences when comparing the observed to predicted centrality metrics by production type ($P > 0.05$) for all, except *betweenness* of market ($P = 0.02$) and *out-degree* of finishing farms ($P = 0.002$).

Additionally, average distances between observed and predicted movements did not significantly vary across types of sites (N, Fa, Fi, and M with *P*-values of 0.70, 0.05, 0.29, 0.94, respectively) (**Figure 8**). Among farms, we noticed that finishing farms shipped animals the longest distances (observed and predicted averages 128.4 and 130.7 km, respectively), whereas markets on average received animals from 107.3 km away versus a prediction of 115.7 km.

Prediction of the RCP-N212 Network

The RC-N212 dataset contained 830 sites, 65.1% of which specialized in the last stage of production (e.g., growing, finisher, wean-to-finish), and 32% characterized as farrowing farms or nurseries. Only 1.2% of total sites recorded in RCP-N212 were market sites, which were located in only 4 out of the 34 counties in which RCP-N212 sites were located. We generated 688,070 possible origin-destination pairs, and our model predicted that 0.9% of those pairs were likely to move animals between them, using a probability threshold >0.85. However, if the number of likely links for a given farm exceeded the maximum observed *in-* or *out-degree* for its production type (**Table 4**), the number of contacts was restricted to the maximum degree by randomly selecting from the highly probable links. This process resulted in a network where 0.4% of the total pairs were likely to move pigs between them.

Unsurprisingly, market sites reached the maximum allowable *in-degree*, receiving pigs from 57 sites, whereas farrowing farms, nurseries, and finishers were expected to receive animals (perhaps replacements), on average, from 4, 1, and 2 sites, respectively. On the other hand, the model predicted that nurseries and farrowing farms would ship pigs (i.e., *out-degree*) to 12 and 10 farms, respectively (**Figure 9A**). *Betweenness* was highest in farrowing and nursery farms, followed by finishing farms (**Figure 9A**).

The model using RCP-N212 data predicted animal movement distances that were slightly different from those predicted using the testing dataset presented in the previous section. Finishing farms were expected to ship (i.e., *out-degree*) animals through, on average, 141.4 km (SD = 75.5 km), whereas nurseries, on average, 113.6 km away (SD = 67.7 km) to sites within the RCP-N212. In turn, farrowing farms and market sites were expected to receive animals from longer distances (mean = 168.4 km, SD = 51.0 km, and mean = 104.3 km, SD = 98.5 km, respectively) (**Figure 9**). The density of the predicted network in RCP-N212 was 0.004, with a clustering coefficient of 2.8%, and a mean path length of 7.26.

Predicted pig movements in the RCP-N212 covered large spatial areas, and only 14% were within the same county. In general, most predicted pig movements passed through several counties, with a maximum of 11 counties. Finally, the predicted network for the RCP-N212 suggested a major aggregation of movements to and from sites located in areas toward the southern part of the regional program (**Figure 10**).

DISCUSSION

The aim of this research was to predict animal movements among sites located within a given RCP. Unfortunately, data of movement networks are often incomplete or unavailable for food animal industries characterized by a large number of animal movements between sites, such as the US swine industry (19, 25). Therefore, we employed machine-learning techniques to illustrate how models may be fitted by using a subset of the data to increase their completeness and accuracy. Specifically, using information available in only two counties, we studied the likelihood of possible movements among sites in a larger-scale swine disease RCP in Minnesota, referred to as RCP-N212. In general, networks predicted by the RF model were consistent with the observed data

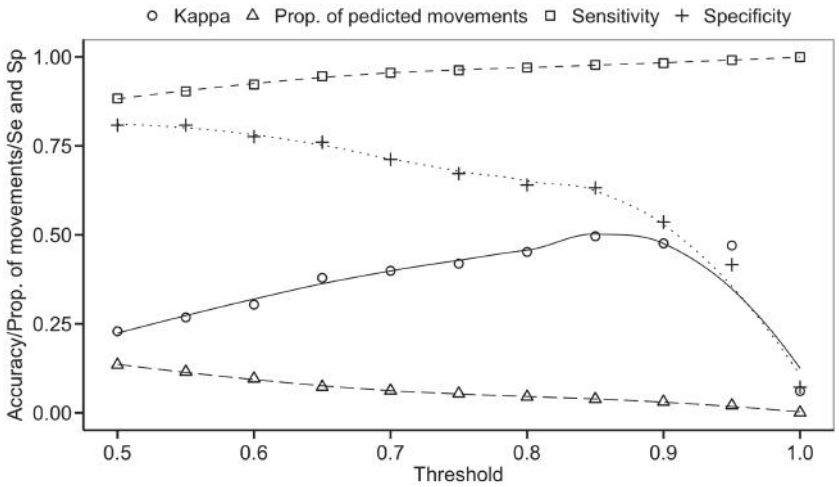


FIGURE 5 | Kappa statistic accompanying the threshold probability for the proportion of predicting movements out of the total pairs using balanced data. Sensitivity (Se) and specificity (Sp) are also reported through different thresholds.

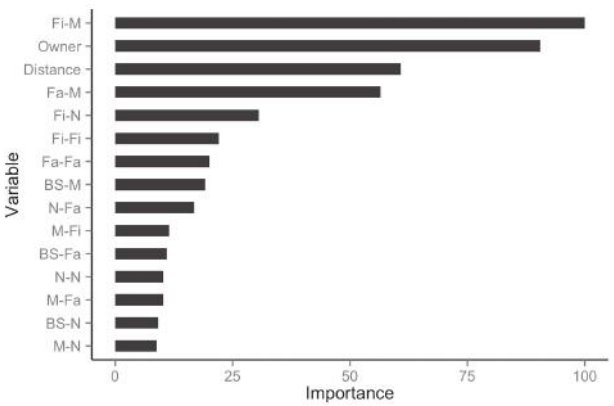


FIGURE 6 | Rank of the 15 most important variables for network prediction using the balanced dataset.

used for model training and testing in terms of both spatial and production-type connectivity patterns.

The SN and RN networks exhibited relatively similar centrality patterns and a marked flow of animal movements from upstream to downstream sites in the production chain. However, the network structure of both counties also indicated that finishing (and even market) sites might provide animal replacement (e.g., gilts and boars) to upstream sites. Because outbreaks of diseases, such as PRRS, are also common in downstream sites (26), movements from those sites to upstream sites could perpetuate disease in those areas. Indeed, previous research has shown that despite an overall decrease in the occurrence of PRRS, spatial and temporal aggregations of that disease allowed for continued hotspots throughout the period of study (26). These interactions merit further analysis to explain swine disease dynamics, especially for industry-persistent diseases such as PRRS (2, 6, 48).

While model results suggest that ownership and distance are strongly related to the probability of pig movement between sites, the production type of the origin and destination sites also influenced the probability of pig movements from one location to another. Moreover, if we consider that different types of farms might share transportation services, whereby farms may ship or receive different type of animals (e.g., feeder pigs and finishing pigs), such mixing might facilitate spread disease *via* contaminated vehicles (4, 49). Thus, we suggest that a complete evaluation of disease risks associated with transportation of pigs between facilities should take into account factors such as the commercial relationship between sites, including contractual agreements in the US swine industry (50), and site production type (6).

As mentioned previously, we used information available from two small, county-based networks to estimate parameters for predictions of animal movements that closely fit observed movements as judged by standard statistical tests. We were able to validate our predictions within SN and RN. The parameters generated by our model were used to predict animal movements between sites over a larger area, i.e., RCP-N212. This is essentially an out-of-sample prediction. As data on actual animal movements were not available for RCP-N212, we cannot directly test the accuracy of our predictions for the larger network. The results for larger network appear reasonable in that they are consistent with the topology of the SN and RN networks, and these results may be useful in helping understand actual (but unobservable) animal movements in Minnesota. We believe that such out-of-sample prediction is warranted for the scale of RCP-N212, given that farms within this program are similar to the farms in SN and RN in terms of geography, demography, and management. However, predictions that would rely on more extensive extrapolation (such as at the scale of multiple states) would not be appropriate given the scope of our sampling. In addition, the value of being able to assemble a full network may not be to target individual farms, but rather to capture possible regional patterns in connectivity.

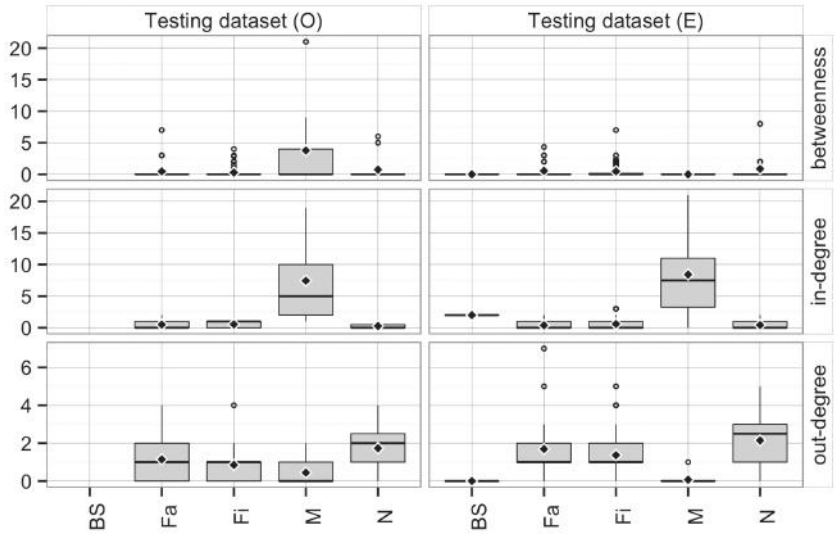


FIGURE 7 | **Boxplot of betweenness, in- and out-degrees in the testing dataset for observed (*O*) and predicted (*E*) movements by type of site**. Boxes indicate the first and third percentile; middle bar represents the median and diamonds represent the mean values.

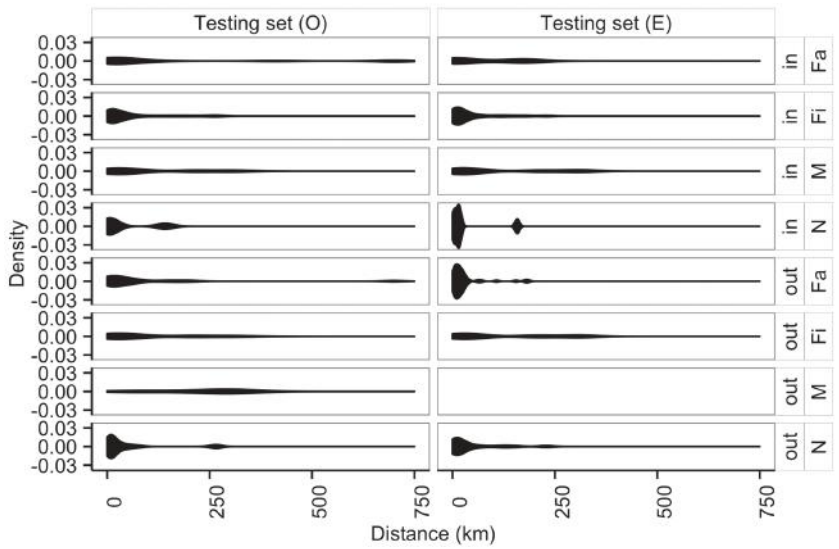


FIGURE 8 | **Distribution of distances (km) of shipments (out) and entries (in) for observed (*O*) and predicted (*E*) movements by type of location**.

Based on comparisons between observed, county-level data and regional-level model predictions, such as the distribution of movement distances and general attributes of the network, we believe that our predicted full network for RCP-N212 has structural features that are similar to the partial data used to estimate the probability of movement between farms. Thus, our findings appear reasonable and provide insight to better understand animal movement patterns within the RCP-N212. This, in turn, may help farmers design private strategies for sanitary management, as well as aid policy-makers in structure-based decisions. However, given inherent limitations to predictive modeling, we acknowledge that our results may provide only general insights about movement patterns, thus additional work must be done before strong conclusions can be made regarding the utility of the predictions achieved in this way.

The RCP-N212 covers 34 out of 87 counties in Minnesota, accounting for 38% of the total swine facilities with 100 or more heads in the state. Because farm distribution is heterogeneous within Minnesota, with a greater number of farms toward the south (28), it is reasonable to infer that the distribution and type of sites across the RCP-N212 should influence our network predictions. Furthermore, given that 65.1% of the sites are dedicated to the last stage of production, we acknowledge that a fraction of sites within the RCP-N212 must trade animals with sites

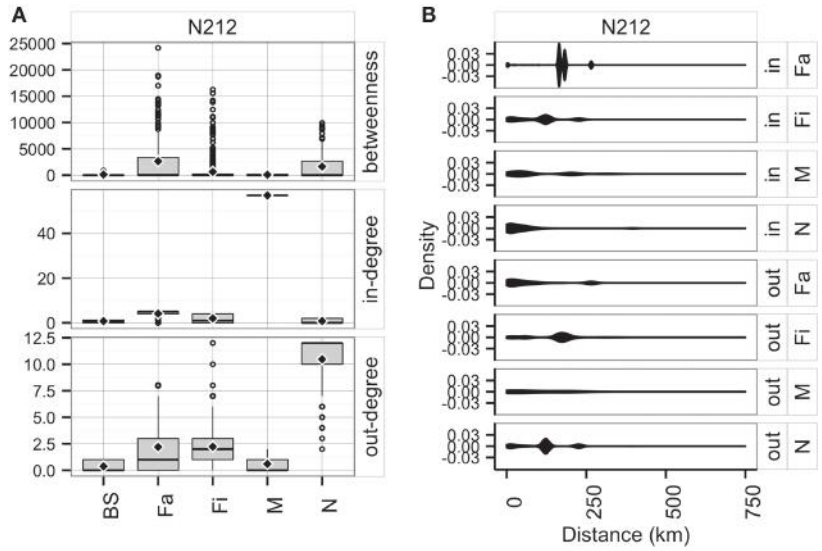


FIGURE 9 | Boxplot of betweenness, in- and out-degrees distributions (A), and distance (km) distribution of shipments (out) and entries (in) (B) for the predicted network in RCP-N212.

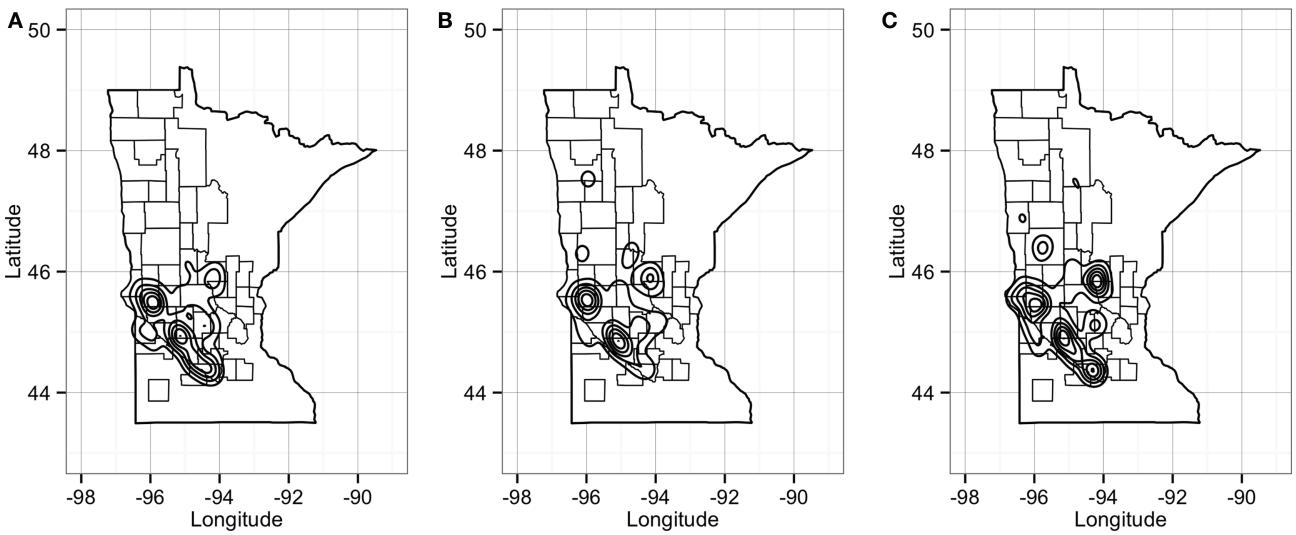


FIGURE 10 | Plot of geographical density for expected out-degree (A), in-degree (B), and betweenness (C) of sites in the RCP-N212. Counties delineated compose the area of the RCP-N212.

located in neighboring states, such as Iowa or Illinois, or even more distant states such as North Carolina, where a high number of sow farms are located (20, 51, 52). This could have led to an overestimation of movements between some sites in our model, especially for those underrepresented in the RCP. To tackle this issue, we applied two constraints to our model: (1) we restricted prediction of movements by increasing the probability threshold to 85% when assigning a possible link between two sites and (2) we constrained the maximum number of links per site using maximum values of *in*- and *out-degree*. As a result our movement predictions within the RCP-N212 were conservative, showing a lower density in the expected network (0.4%) than in the two observed networks (1.4% and 1.3% for SN and RN, respectively), although some metrics might have been overestimated, such as *out-degree* for nurseries. This is because there are very few nursery farms in this region, and many of the finishing farms are actually sourcing for pigs from other states outside of RCP-N212. However, our algorithm restricted their choice of nurseries to those within RCP-N212, perhaps leading to a false inflation of their *out-degree*. Future work should expand to larger geographic regions that encompass all stages of production, capturing movements between states. We anticipate that we could improve

model predictions by obtaining information regarding contract relationships and animal movements among Minnesota sites and suppliers located outside RCP-N212.

In the predicted RCP-N212 network, we found that sites with higher *in-* and *out-degrees* overlap with areas where spatial and temporal aggregations of PRRS have occurred (26). Therefore, it is possible that movements, in addition to farm density, might play an important role in the persistent circulation of disease in the area. The co-aggregation of animal movements and PRRS, a disease believed to be transmitted, at least in part, by animal movements (2, 6, 49), suggests that our predicted network might be capturing important features of the underlying industry structure, which indirectly supports the validity of our network predictions.

The characterization of network structures often may help for planning production and designing strategies to control animal disease (7, 8, 11, 18). The approach developed here is an early step for helping in design strategies to control swine diseases regionally. For example, the spread of swine pathogens within the full network can be simulated using computational models, which would be valuable for both predicting patterns of between-farm spread and for evaluating alternate intervention and control strategies. Among them, for instance, vaccination strategies that maximize the collective good could be quantitatively explored, including minimum levels of coverage that may prevent disease circulation in the network. Additionally, the approach developed here may reduce time and cost for data collection, as collection of movements among a partial set of sites might be sufficient to predict movements among a larger set of sites.

Among classification techniques, there are several approaches that might be used to predict possible outcomes, such as links between sites. While the focus of this paper is not to provide an exhaustive review of these techniques, here we offer some ground for further discussion and perhaps comparative studies. The RF approach has high accuracy without overfitting, it is also fairly stable to the presence of outliers and noise, and it may handle the correlation between predictors (40, 41, 53). This may be important in the context of this study, as some atypical movements between sites may occur, predictors may be correlated, and the probability of animal movement between two or more sites may often occur in a non-linear fashion. Alternatively, other supervised techniques might be used. For example, support vector machines, a vector function based technique that splits the data for classification purposes, might resolve non-linearity in the data by using a non-linear kernel function, though its performance sometimes might be compromised (40).

In conclusion, we present an approach to predict the network structure of contacts between and among farms in a region by using partial data. Our results, combined with information on the occurrence of disease in the area (i.e., outbreaks of PRRS within the RCP-N212), may be incorporated into a disease transmission model that will help to evaluate the effectiveness of prevention and control strategies in a region, with the ultimate objective of mitigating the impact of endemic disease and hypothetical epidemic incursions. The approach here may also be applied to other regions and production systems, where information on animal movements is only partially regulated, thus improving decision-makers' ability to plan and implement disease surveillance and control activities.

AUTHOR CONTRIBUTIONS

All the authors have met the four criteria described at the guidelines: PVD designed and data interpretation, revised and approved the version to be published, and agreed to be accountable for all aspects of the work. KV and LJ designed, revised and approved the version to be published, and agreed to be accountable for all aspects of the work. SW data interpretation, revised and approved the version to be published, and agreed to be accountable for all aspects of the work. AP designed, revised and approved the version to be published, and agreed to be accountable for all aspects of the work.

ACKNOWLEDGMENTS

The Becas-Chile program of the National Commission for Scientific and Technological Research (CONICYT) has supported PV-D to develop his Master and PhD in the US. Additional support has been provided by the University of Minnesota MnDrive program and by the National Pork Board. Authors thank Jaclyn R. Aliperti, PhD(c), UCD, for providing valuable editorial advice. LS acknowledges support from the National Institute of Food and Agriculture (NIFA).

FUNDING

No funding was provided for the development of this article.

REFERENCES

1. Fèvre EM, Bronsvoort BMDC, Hamilton KA, Cleaveland S. Animal movements and the spread of infectious diseases. *Trends Microbiol* (2006) 14:125–31. doi:10.1016/j.tim.2006.01.004
2. Albina E. Epidemiology of porcine reproductive and respiratory syndrome (PRRS): an overview. *Vet Microbiol* (1997) 55:309–16. doi:10.1016/S0378-1135(96)01322-3
3. Dee S, Deen J, Rossow K, Weise C, Eliason R, Otake S, et al. Mechanical transmission of porcine reproductive and respiratory syndrome virus throughout a coordinated sequence of events during warm weather. *Can J Vet Res* (2003) 67:12–9.
4. Dee SA, Deen J, Otake S, Pijoan C. An experimental model to evaluate the role of transport vehicles as a source of transmission of porcine reproductive and respiratory syndrome virus to susceptible pigs. *Can J Vet Res* (2004) 68:128–33.
5. Dee S, Clement T, Schelkopf A, Nerem J, Knudsen D, Christopher-Hennings J, et al. An evaluation of contaminated complete feed as a vehicle for porcine epidemic diarrhea virus infection of naïve pigs following consumption via natural feeding behavior: proof of concept. *BMC Vet Res* (2014) 10:1–9. doi:10.1186/s12917-014-0220-9
6. Perez AM, Davies PR, Goodell CK, Holtkamp DJ, Mondaca-Fernández E, Poljak Z, et al. Lessons learned and knowledge gaps about the epidemiology and control of porcine reproductive and respiratory syndrome virus in North America. *J Am Vet Med Assoc* (2015) 246:1304–17. doi:10.2460/javma.246.12.1304
7. Sydow J, Windeler A. Organizing and evaluating interfirm networks: a structurationist perspective on network processes and effectiveness. *Organ Sci* (1998) 9:265–84. doi:10.1287/orsc.9.3.265
8. Martínez-López B, Perez AM, Sánchez-Vizcaíno JM. Social network analysis. Review of general concepts and use in preventive veterinary medicine. *TransboundEmergDis* (2009) 56:109–20. doi:10.1111/j.1865-1682.2009.01073.x

9. Jackson MO. *Social and Economic Networks*. Princeton: Princeton University Press (2008).
10. Hagerman AD, Mccarl BA, Carpenter TE, Ward MP, O'Brien J. Emergency vaccination to control foot-and-mouth disease: implications of its inclusion as a U.S. policy option. *Appl Econ Perspect Policy* (2011) 34:119–46. doi:10.1093/aepp/ppr039
11. Nöremark M, Håkansson N, Lewerin SS, Lindberg A, Jonsson A. Network analysis of cattle and pig movements in Sweden: measures relevant for disease control and risk based surveillance. *Prev Vet Med* (2011) 99:78–90. doi:10.1016/j.prevetmed.2010.12.009
12. Rautureau S, Dufour B, Durand B. Structural vulnerability of the French swine industry trade network to the spread of infectious diseases. *Animal* (2012) 6:1152–62. doi:10.1017/s1751731111002631
13. Büttner K, Krieter J, Traulsen A, Traulsen I. Epidemic spreading in an animal trade network – comparison of distance-based and network-based control measures. *Transbound Emerg Dis* (2016) 63:e122–34. doi:10.1111/tbed.12245
14. Natale F, Giovannini A, Savini L, Palma D, Possenti L, Fiore G, et al. Network analysis of Italian cattle trade patterns and evaluation of risks for potential disease spread. *Prev Vet Med* (2009) 92:341–50. doi:10.1016/j.prevetmed.2009.08.026
15. Bajardi P, Barrat A, Savini L, Colizza V. Optimizing surveillance for livestock disease spreading through animal movements. *J R Soc Interface* (2012) 9:2814–25. doi:10.1098/rsif.2012.0289
16. VanderWaal KL, Picasso C, Enns EA, Craft ME, Alvarez J, Fernandez F, et al. Network analysis of cattle movements in Uruguay: quantifying heterogeneity for risk-based disease surveillance and control. *Prev Vet Med* (2016) 123:12–22. doi:10.1016/j.prevetmed.2015.12.003
17. Mardones FO, Martinez-Lopez B, Valdes-Donoso P, Carpenter TE, Perez AM. The role of fish movements and the spread of infectious salmon anemia virus (ISAV) in Chile, 2007–2009. *Prev Vet Med* (2014) 114:37–46. doi:10.1016/j.prevetmed.2014.01.012
18. Lentz HHK, Koher A, Hövel P, Gethmann J, Sauter-Louis C, Selhorst T, et al. Disease spread through animal movements: a static and temporal network analysis of pig trade in Germany. *PLoS One* (2016) 11:e0155196. doi:10.1371/journal.pone.0155196
19. Shields DA, Mathews K. *Interstate Livestock Movements*. Economic Research Service (2003). Available from: http://www.ers.usda.gov
20. McBride WD, Key N. *U.S. Hog Production from 1992 to 2009: Technology, Restructuring, and Productivity Growth*. Washington, DC: US Department of Agriculture, Economic Research Service (2013).
21. Hurt C. Industrialization in the pork industry. *Choices* (1994) 9:9–13.
22. Key N, McBride W. *The Changing Economics of U.S. Hog Production, ERR-52*. Washington, DC: United States Department of Agriculture, Economic Research Service (2007).
23. Tilman D, Cassman KG, Matson PA, Naylor R, Polasky S. Agricultural sustainability and intensive production practices. *Nature* (2002) 418:671–7. doi:10.1038/nature01014
24. MacDonald JM, McBride WD. *The Transformation of US Livestock Agriculture: Scale, Efficiency, and Risks*. Washington, DC: US Department of Agriculture, Economic Research Service (2009).
25. USDA. *Animal Disease Traceabililty*. Washington DC: (2016). Available: https://www.aphis.usda.gov/aphis/ourfocus/animalhealth/SA_Traceability
26. Valdes-Donoso P, Jarvis LS, Wright D, Alvarez J, Perez AM. Measuring progress on the control of porcine reproductive and respiratory syndrome (PRRS) at a regional level: the Minnesota N212 regional control project (Rcp) as a working example. *PLoS One* (2016) 11:e0149498. doi:10.1371/journal.pone.0149498
27. Wayne SR. *Assessment of Demographics and Network Structure of Swine Populations in Relation to Regional Disease Transmission and Control*. St Paul, MN: University of Minnesota (2011).
28. USDA. *Census of Agriculture 2012. Minnesota: State and County Data*. N.a.S. Service (2014). Available from: http://www.agcensus.usda.gov/Publications/2012/Full_Report/Census_by_State/Minnesota/index.asp
29. Newman MEJ. Mixing patterns in networks. *Phys Rev E* (2003) 67:1–13. doi:10.1103/PhysRevE.67.026126
30. Wasserman S, Faust K. *Social Network Analysis: Methods and applications*. Cambridge: Cambridge University Press (1994).
31. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R.F.F.S. Computing (2015).
32. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer (2009).
33. Becker RA, Wilks AR. *maps: Draw Geographical Maps*. (2014).
34. Venables WN, Ripley BD. *Modern Applied Statistics with S*. New York: Springer (2002).
35. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal Complex Syst* (2006).
36. Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. *J Am Soc Inf Sci Technol* (2007) 58:1019–31. doi:10.1002/asi.20591
37. Lichtenwalter RN, Lussier JT, Chawla NV. New perspectives and methods in link prediction. *16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA (2010).
38. Newman MEJ. Clustering and preferential attachment in growing networks. *Phys Rev E* (2001) 64:025102. doi:10.1103/PhysRevE.64.025102
39. Katz L. A new status index derived from sociometric analysis. *Psychometrika* (1953) 18:39–43. doi:10.1007/BF02289026
40. Kuhn M, Johnson K. *Applied Predictive Modeling*. New York, U S: Springer (2013).
41. Breiman L. Random forests. *Mach Learn* (2001) 45:5–32. doi:10.1023/A:1017934522171
42. Kuhn M. *Caret: Classification and Regression Training*. (2015).
43. Chen C, Liaw A, Breiman L. *Using Random Forest to Learn Imbalanced Data*. Berkeley, CA: University of California (2004).
44. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* (2008) 28:1–26. doi:10.18637/jss.v028.i05
45. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409 [Online] (2015). Available from: https://arxiv.org/abs/1508.04409 (accessed January 11, 2016).
46. Silverman BW. *Density Estimation for Statistics and Data Analysis*. London: CRC Press (1986).
47. Duong T. ks: kernel density estimation and kernel discriminant analysis for multivariate data in R. *J Stat Softw* (2007) 21:1–16. doi:10.18637/jss.v021.i07
48. Corzo CA, Mondaca E, Wayne S, Torremorell M, Dee S, Davies P, et al. Control and elimination of porcine reproductive and respiratory syndrome virus. *Virus Res* (2010) 154:185–92. doi:10.1016/j.virusres.2010.08.016
49. Dee S, Deen J, Burns D, Douthit G, Pijoan C. An evaluation of disinfectants for the sanitation of porcine reproductive and respiratory syndrome virus-contaminated transport vehicles at cold temperatures. *Can J Vet Res* (2005) 69:64–70.
50. Giamalva J. *Pork and Swine. Industry and Trade Summary*. Washington, DC: International Trade Commission (2014).
51. McBride WD, Key N. *Economic and Structural Relationships in U.S. Hog Production*. USDA-ERS Agricultural Economic Report – SSRN Electronic Journal (2003).
52. Alvarez J, Valdes-Donoso P, Tousignant S, Alkhamis M, Morrison R, Perez A. Novel analytic tools for the study of porcine reproductive and respiratory syndrome virus (PRRSv) in endemic settings: lessons learned in the U.S. *Porcine Health Manag* (2016) 2:1–9. doi:10.1186/s40813-016-0019-0
53. Liaw A, Wiener M. Classification and regression by randomForest. *R News* (2002) 2:18–22. doi:10.1057/9780230509993

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Veterinary Science

METHODS
published: 03 May 2018
doi: 10.3389/fvets.2018.00071

Effective Network Size Predicted From Simulations of Pathogen Outbreaks Through Social Networks Provides a Novel Measure of Structure-Standardized Group Size

Collin M. McCabe[1,2,3] *and Charles L. Nunn*[3,4]*

[1] Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, United States, [2] Division of Infectious Diseases and Global Health, Department of Medicine, Duke University, Durham, NC, United States, [3] Department of Evolutionary Anthropology, Duke University, Durham, NC, United States, [4] Triangle Center for Evolutionary Medicine (TriCEM), Durham, NC, United States

OPEN ACCESS

Edited by:
Kimberly VanderWaal,
University of Minnesota,
United States

Reviewed by:
Cédric Sueur,
UMR7178 Institut pluridisciplinaire Hubert Curien (IPHC), France
Paul Cross,
United States Geological Survey,
United States

***Correspondence:**
Charles L. Nunn
clnunn@duke.edu

Specialty section:
This article was submitted to Veterinary Epidemiology and Economics, a section of the journal Frontiers in Veterinary Science

Received: 18 April 2017
Accepted: 26 March 2018
Published: 03 May 2018

Citation:
McCabe CM and Nunn CL (2018) Effective Network Size Predicted From Simulations of Pathogen Outbreaks Through Social Networks Provides a Novel Measure of Structure-Standardized Group Size. Front. Vet. Sci. 5:71. doi: 10.3389/fvets.2018.00071

The transmission of infectious disease through a population is often modeled assuming that interactions occur randomly in groups, with all individuals potentially interacting with all other individuals at an equal rate. However, it is well known that pairs of individuals vary in their degree of contact. Here, we propose a measure to account for such heterogeneity: effective network size (ENS), which refers to the size of a maximally complete network (i.e., unstructured, where all individuals interact with all others equally) that corresponds to the outbreak characteristics of a given heterogeneous, structured network. We simulated susceptible-infected (SI) and susceptible-infected-recovered (SIR) models on maximally complete networks to produce idealized outbreak duration distributions for a disease on a network of a given size. We also simulated the transmission of these same diseases on random structured networks and then used the resulting outbreak duration distributions to predict the ENS for the group or population. We provide the methods to reproduce these analyses in a public R package, "enss." Outbreak durations of simulations on randomly structured networks were more variable than those on complete networks, but tended to have similar mean durations of disease spread. We then applied our novel metric to empirical primate networks taken from the literature and compared the information represented by our ENSs to that by other established social network metrics. In AICc model comparison frameworks, group size and mean distance proved to be the metrics most consistently associated with ENS for SI simulations, while group size, centralization, and modularity were most consistently associated with ENS for SIR simulations. In all cases, ENS was shown to be associated with at least two other independent metrics, supporting its use as a novel metric. Overall, our study provides a proof of concept for simulation-based approaches toward constructing metrics of ENS, while also revealing the conditions under which this approach is most promising.

Keywords: social network analysis, compartmental modeling, simulation modeling, group size, parasites, disease ecology, disease outbreaks

INTRODUCTION

Theoretical models allow us to make sense of complex phenomena by applying a set of simplifying assumptions. In many cases, however, empirical observations of the phenomena do not conform to these assumptions. Understanding how observations compare to their theoretical ideals is thus critical to the interpretation of any such model. Within biology, one of the earliest attempts to compare observations to their theoretical ideals was the work of Wright on effective population size (1). Effective population size models take an observed population with a certain amount of genetic diversity and predict the size of an idealized population under the assumptions of Fisher–Wright populations that groups are of finite and fixed sizes, individuals mate randomly, and generations do not overlap (2–4). The generalizability of effective population size allows biologists to compare populations, which is useful in many contexts, including wildlife management and conservation policies (5).

Infectious disease represents another phenomenon in which the concept of an idealized population is useful. As with effective population size, a set of simplifying assumptions exist that can be repurposed to formulate theoretically idealized populations, given an observed population. Compartmental disease models aim to predict disease transmission by using assumptions similar to those in Fisher–Wright populations. For example, they assume that individuals transmit pathogens freely throughout the population, similar to the Fisher–Wright assumption of random mating (the free association assumption); individuals do not immigrate or emigrate, maintaining a Fisher–Wright constant population size; and there is no age structure within the population, with non-overlapping generations (6). However, these assumptions are rarely met in natural populations. As shown through early critiques of compartmental disease models (6) and more recently through the resurgence of social network studies, interactions are not random, but instead structured along social ties between specific individuals based on affiliative interactions, mating, and other social behaviors (7).

Here, we investigated how changes specifically to the free association assumption, through structuring in social networks, affect the time it takes for a disease to transmit through a population. To assess the deviation of an observed population from a theoretical ideal in disease transmission through structured groups, we must define what represents an idealized population and disease outbreak. Many ecological and environmental factors can affect group size and structure, including food distribution and predation. By "ideal," we are referring to a perfect adherence to the assumption of free association. By "free association," we are referring to the fact that all individuals have equal probabilities to interact with every other individual in the population, perfectly mirroring the mass action properties of traditional compartmental disease models at infinite population sizes.

In a review of network modeling of epidemics, Keeling and Eames (8) suggest that a variety of idealized networks exist, depending on the end goal of the model. The purpose of our model is to allow free association between individuals in a social network. The earliest modeling of disease transmission through networks was conducted on lattices (9), with regularly structured connections between individuals (**Figure 1A**). However, lattices show too much deviation from the Fisher–Wright assumption of completely free and random association to be used as an idealized population. Instead, given the assumptions of basic compartmental models, the most fitting network arrangement to be used as an ideal is a maximally complete network, in which each individual has uniform ties to each other individual in the network, allowing for effectively free association among all nodes (**Figure 1B**).

As for the epidemiological model, either deterministic or stochastic models are used to model the transmission of disease. As we are aiming to simplify assumptions about the transmission of disease, deterministic models would provide more straightforward, less complicated views of disease transmission. However, deterministic models require an intimate knowledge of the dynamics of disease transmission within a population; unknown variables, such as the effect of social structure on outbreaks, make this sort of modeling impossible. Stochastic models, which are often more representative of real-world heterogeneity in disease transmission, allow for uncertainty in variables or dynamics by simulating many different, randomly selected values for important variables (11). For this reason, we employed stochastic models for our study.

Infectious diseases that are transmitted and maintained in populations can be modeled using a variety of epidemiological models. For instance, susceptible-infected (SI) models are useful for investigating the transmission of diseases caused by lifelong infections, where no recovery is possible; these models include specialized types of SI diseases, like sexually transmitted diseases, where transmission rates vary depending on which sex of individual is interacting. For following disease outbreaks through a population where recovery and resistance is possible, the simplest sufficient compartmental model would be a susceptible-infected-recovered (SIR) model, where susceptible individuals become infected from other infected individuals, but they will eventually be removed from the population of susceptible and infectious individuals, either recovering with full immunity to further infection or dying from the disease (which, for the purposes of our research, are functionally equivalent). To capture the large amount of variation among diverse types of diseases, and to be as relevant as possible to researchers studying a potentially wide variety of pathogens, we investigated SI and SIR models in this study using per contact transmission and recovery rates that were realistic but would still allow time for recovery or extinction in SIR models.

Previous work on determining the effective size of a network has focused on very specific aspects of network structure and has thus maintained a restricted conception of what constitutes an idealized network. In the only comparable epidemiological research on this topic, Caillaud et al. (12) proposed a measure of "epidemiological effective group size." This metric considered the variation in sub-group size within a meta-population and the impact of this variation on the outbreak of a disease within the meta-population. By using maximally complete networks of sub-groups connected to other maximally complete sub-groups, the researchers calculated the likelihood of an epidemic outbreak throughout the meta-population based on the size of the index sub-group. Thus, Caillaud et al.'s (12) metric is essentially a novel

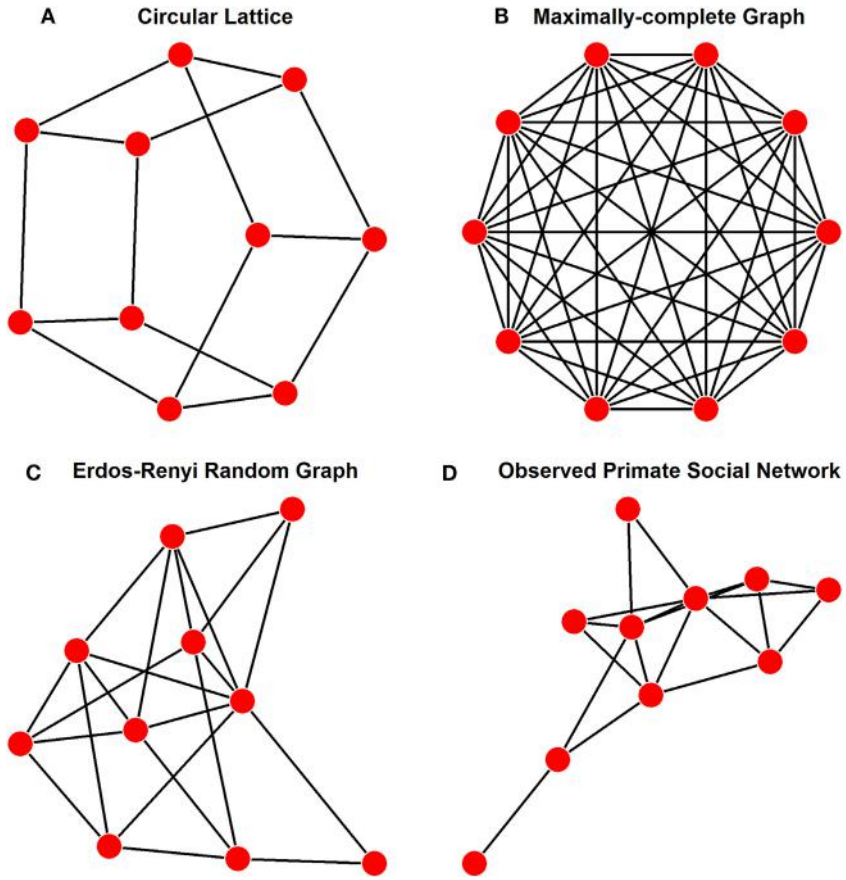


FIGURE 1 | Examples of a population of 10 individuals showing various representative network structures, as discussed in the text. These different structures and their applications are **(A)** lattice structure as has been used in other network models for disease transmission, where ties are regular, but not exhaustively complete; **(B)** maximally complete structure as was used for our idealized networks with free association, each individual is connect to all other individuals in the population; **(C)** Erdõs-Rényi generation structure with every possible tie existing with a probability of 0.25, thus the number of ties in this graph are a quarter of those present in panel **(B)**; and **(D)** an example of an empirically observed network of social interactions among primates [*Pan troglodytes* (10)].

measure of an invasion-specific critical community size needed to maintain an outbreak (13), which in addition to the previous measure, also takes variation in meta-population structure and sub-group size into account.

In addition, another notion of effective group size has been utilized in estimating the number of distinct cultural behaviors, or cultural richness, that is maintained within a human population. This approach was first theoretically developed by Henrich (14) using assumptions of even mixing for cultural transmission of multiple behaviors through a population; results of this analysis demonstrated that a decrease in the size of a population through geographic isolation could explain the loss of complex cultural behaviors among Tasmanian islanders. This method was further developed by Powell et al. (15) to incorporate spatial and temporal variability through estimates of population density and migration rates, respectively. Using this method, the researchers showed that the variability in human population density and migratory activity, resulting in "effective population sizes" for human groups, explained much of the geospatial distribution in cultural behaviors during the Late Pleistocene Epoch. These methods are closely related to those described in our study, in that each is using population structure to explain observed richness, either cultural or parasitic. However, the models for explaining observed richness of human behavior did not explicitly incorporate social network structure; this is the main contribution of our own method.

While our methods do incorporate the complexities of network structure in diseases transmission, we are omitting many other important factors of social structure and disease ecology. As just noted, social structure can have important impacts on the maintenance of cultural behaviors (14, 15), and cultural behaviors themselves have been shown to have significant impacts on disease transmission (16). In addition, several other factors can influence the structure and size of a group, including the relative despotism or tolerance of a group (17, 18), ecology (19), or resource availability (20). Our goal in omitting these factors from the following analyses is not to downplay any of their impacts on social network structure or disease transmission but rather to isolate the effect of social network structure on disease transmission using a simplified model.

The first specific aim of our study is to quantify the relationship between networks of various sizes and outbreak durations for diseases with and without immunity, and with variation in epidemiological parameters (focusing on variation in per contact probability of transmission). Here, we expect that infectious diseases transmitted through larger networks will show longer outbreak durations than disease transmitted through smaller networks (12). We investigate the relationship between group size and outbreak duration to provide a basis for calculating effective group size. The second specific aim is to generate randomly structured networks and to simulate disease transmission through those randomly structured networks to predict what sized maximally complete network would have the same outbreak duration; we call these the effective network size (ENS) of the social group. Just as we establish a relationship between outbreak duration and maximally complete network size to provide a baseline relationship between them, we use this same relationship between network size and outbreak duration to predict the ENS of randomly structured groups from the outbreak durations of their SI and SIR simulations. It is important to note that our measure of ENS will always be equal to or larger than the original group size, which differs significantly from effective population size, which is always equal to or smaller than the original group size. Among these simulations, we compare the accuracy and precision of using regression models to predict ENS from distributions of outbreak durations on the randomly structured networks. All of the methods described in this study can be easily replicated with a publicly accessible R package, enss, developed specifically for this study (https://www.github.com/collinmmccabe/enss), and the relevant functions for each step of the analysis are noted throughout the Section "Methods."

Finally, as a proof of concept, we apply our new metric for representing disease transmission to a collection of primate networks (21). We then compare the information represented by ENS to other, more established network metrics to determine the novelty of our metric, as well as its associations with other metrics. The specific metrics that we investigate here are leading eigenvector modularity, mean distance, diameter, clustering coefficient, and eigenvector centralization, as were also investigated by Nunn and colleagues (22). In Data Sheet S1 in Supplementary Material (Supplementary Analysis), we also provide an example use case of ENS from these same primate species, comparing it to raw group size as a predictor of parasite richness.

METHODS

Simulation and Regression of Disease Transmission on Maximally Complete Networks

To address the first aim of correlating idealized networks with disease transmission times, we generated maximally complete, unweighted, undirected networks for groups of size 3–200 in R, version 3.3.2 (23) with packages igraph (24), statnet (25), and functions that we developed and distribute in enss. We then simulated SI models (with a per contact transmission rate, β, of 0.10, and per capita interactions per day set at three times the group size) and SIR models (with an additional parameter, γ, or the daily recovery rate set at 0.10) to saturation or extinction (the points at which pathogens could not be transmitted further) on each of these networks 1,000 times. β and γ were both parameterized at 0.10, following previous disease simulations as described in Griffin and Nunn (26). For β, this value of 0.10 indicated that for every interaction between a susceptible and an infected individual, there was a 10% probability that the susceptible would become infected; for γ, the 0.10 indicated that per daily timestep, each infected individual had a 10% probability of recovering. These methods are available through R package enss as functions "clust_sim_SI" and "clust_sim_SIR," respectively. Although we chose to focus our efforts by using unweighted networks and testing only one value for β, we also present analyses that investigate the effects of incorporating weighted ties and varying values of β.

Per capita social interaction rates per day were chosen arbitrarily to be at a rate of three interactions per individual per day in the analyses presented here. This means that for a group of size 10, 30 random interactions were independently chosen from the set of all available interactions between individuals in the group, which was then repeated for each daily timestep of the model's simulation. Days and per capita interaction rates per day were used as familiar, but ultimately arbitrary demarcations of time in our models so that "outbreak duration" could be measured in a uniform manner.

Our algorithms for disease transmission on networks took place in multiple stages. The first stage involved generating and recording social networks as edgelists, where each social tie between two individuals is recoded as its own row of data. This method can be replicated using enss function "gen_max." We also tracked the infection status of each node, or individual in the network, as susceptible, infected, or recovered. From among these nodes, one was selected as an index case and was infected at the outset of the simulation.

Following previous disease simulations from Griffin and Nunn (26), we then selected consecutive random edges, or social ties between individuals, to determine whether the disease could be transmitted from one node to another (with a probability β of transmission for each interaction); the number of edges that were selected depended on the per capita interactions per day, or 3N, and the number of individuals in the network (ranging from 3 to 200). So, for a network of 10 individuals, we chose 30 random edges each day, allowing for the possibility of repeated sampling of social ties. For each of these edges, we checked whether transmission was possible; in our models, the only opportunity for disease transmission was the case where an edge connected an infected individual with a susceptible one, ignoring any directionality in the interaction. Each edge over which transmission was possible resulted in an actual transmission event (where the susceptible individual becomes infected) with probability $\beta = 0.10$, as described above; this would result in 10% of interactions between susceptibles and infecteds resulting in transmission. After all random edges had been considered for a day, each infected individual in SIR models randomly recovered with a probability γ. The simulation then moved to the next day, and only stopped when the criteria for simulation completion

were met. No maximum duration was set for either SI or SIR models (because these models would eventually reach either saturation or extinction).

We also considered transmission models where each tie in a graph was sampled once per day, rather than randomly in proportion to the number of nodes, because the number of edges in networks grows exponentially with the number of nodes (Equation S1 in Supplementary Material), and per capita interaction rates in large networks would be less likely to represent a given tie than in a smaller network. We call this the "alternative model," and give results in Figure S1 in Supplementary Material. These methods are available in enss as functions "clust_sim_SI_unif" and "clust_sim_SIR_unif." In addition, we also considered transmission models where ties were weighted. In such models, ties with greater weights, or intensity of interaction between two individuals, were sampled more often than lesser weighted ties. In these models, ties were still sampled randomly at the per capita interaction rate per day, but the likelihood of sampling a given tie was proportional to its weight. This model is called the "weighted model" in analyses that follow. These methods are also available in enss as functions "clust_sim_SI_w" and "clust_sim_SIR_w."

We recorded the number of days until the simulation ended as "outbreak duration." For SI models, simulations ended at saturation, defined as the point at which all individuals had transitioned from susceptible to infected. For SIR models, simulations ended at extinction, defined as the point at which no infected individuals were present in the population, either because all susceptible individuals had been infected and subsequently recovered, or because all infected individuals recovered without being able to sustain further transmission to remaining susceptible nodes. We then found a line of best fit through the results for each epidemiological model, using regression models to predict network sizes from outbreak durations. The output for these linear models can be generated in enss with functions "predict_SI_max" and "predict_SIR_max," respectively, as can a graphical representation of these models with function "plot_predict." For SIR models, only simulations where all individuals had been infected at some point in the simulation were considered sufficient. This resulted in exclusion of 26.9% of simulations in which the disease failed to infect every individual. The purpose of this screening was to ensure that a single continuous metric, outbreak duration, could be used to compare all simulations.

To determine under which conditions our method would be most useful, regression models were calculated with raw network size as the response and outbreak duration as the predictor. The association between raw network size and outbreak duration was exponential rather than linear, as would be expected from an exponential growth system like disease transmission in SI models (27). To determine the area of the graphs where we could reliably predict network size from outbreak duration, we used piecewise OLS regressions to predict two separate relationships between outbreak. We did not transform these data at this point, because by splitting the relationship into two separate regressions with piecewise regression, this approach allowed us to identify portions of the graph where prediction could be made appropriately. In the first portion, duration outbreak would show a relatively shallow relationship with network size, making prediction reasonable. But in the second, much steeper portion, relatively small increases in outbreak duration would show much larger increases in predicted network size, making prediction tenuous. We estimated piecewise regression models in R with package segmented (28) to determine where the breakpoint between the two portions of the graph would be; this method optimizes the linear fit of each portion by randomly varying the breakpoint until the best split is achieved. This can be replicated in enss with function "breakpoint_max." We also simulated the simpler SI models with varying values of β to determine if raising or lowering this parameter had any effect on the breakpoint in these piecewise regressions. Such a result would indicate that altering β would allow for better or worse predictions of large network sizes from longer outbreak durations.

In addition to considering piecewise regression models, we separately ran regression models with log-transformed network sizes to achieve a linear fit. For each set of 1,000 iterations of disease simulation on a given network, outbreak durations were quite variable. Thus, we used reduced major axis (RMA), estimates of model II regressions to control for the uncertainty in outbreak duration in addition to that in network size, calculated in R with package lmodel2 (29). RMA estimates consider the variation in both the independent and dependent variable when fitting regression models rather than, as in OLS models, only considering variation in the dependent variable. RMA provided the most suitable control for estimating how variation in outbreak duration would affect our predictions of fixed network sizes.

We then exponential transformed the output of these equations to back-transform for the log-transformation. These exponential-transformed equations formed the basis for calculating "effective" network sizes from outbreak durations of diseases simulated on observed networks. Back-transformations from log-transformed data introduce bias into predicted values because of the difference between errors in log-transformed variables and their untransformed counterparts (30, 31). We considered accounting for this bias by using the "consistent I estimator" from Hayes and Shonkwiler (30), and compared this approach to our own method of calculating network size from the uncorrected RMA models; the equation for the consistent I estimator is:

$$y = \mathbf{e}^{\ln(a)+b\ln(x)+\left(\frac{s^2}{2}\right)}$$

where a is the intercept, b is the slope, x is the independent variable, and s^2 is the mean squared error for the model. Because mean squared error is constant within each model, such a correction would create a consistent upward shift in all estimates of network size by a value of $s^2/2$; this would not have any impact on further linear models' slope coefficients, and so uncorrected RMA model back-transformations were chosen for simplicity of interpretation throughout the main text. Back-transformed predictions can be obtained in enss with function "estimate_backtrans_ens." Comparisons of observed versus effective outbreak duration distributions are given in Figures S2 and S3 in Supplementary Material.

Accuracy, Precision of Predicting ENS From Randomly Structured Graphs

To investigate the second aim, we generated large sets of Erdős-Rényi (E-R) graphs (**Figure 1C**) for predetermined group sizes and predetermined density of ties present; to reduce variability, these were used as set numbers of ties, rather than probability that ties would be present between two given nodes, as is more typical in density-determined E-R graphs in R with package igraph (24). Random graphs were used as the baseline in this case because they represented the only source from which we could obtain a large enough sample size to validate our methods. Group sizes for these were kept smaller than the maximally complete networks to allow for direct comparison of outbreak duration distributions, and they are in good agreement with the observed network sizes of primates ranging from 4 to 35 typically (32). Tie proportions were kept relatively low to increase differentiation from maximally complete networks. We sampled blocks of 111 networks for each combination of group size ($n = 10$, 30, and 50) and tie proportion (15, 25, and 35% of possible ties), generating 999 total random networks. To ensure that disease simulations could reach full saturation and (for SIR) subsequent extinction, we screened each randomly generated network to ensure that all nodes were part of a single, connected network. This method can be reproduced using function "gen_erg" in package enss.

We then simulated the same SI and SIR models (as discussed in Section "Simulation and Regression of Disease Transmission on Maximally Complete Networks") over 1,000 iterations on each of our 999 randomly generated models, recording outbreak durations of the models (again with enss functions "clust_sim_SI" and "clust_sim_SIR," respectively). Because all outbreak durations for random networks of size N are expected to be greater than those of the idealized network of size N, these simulations were conducted to determine the scale of increase in outbreak durations and consequently in ENS. The mean of outbreak durations for a given random network with a given epidemiological model were used as the predictor variable in the RMA regression equations described in Section "Simulation and Regression of Disease Transmission on Maximally Complete Networks." Only simulations which reached saturation were analyzed here, and so some runs of the SIR simulations were removed due to stochastic extinction events. This reduced the sample size of analyzed simulation runs and may have biased our results for SIR comparisons. These values were then exponential-transformed and rounded to the nearest integer to arrive at a directly comparable ENS for each random network (using enss function "estimate_backtrans_ens"). Thus, ENS were calculated twice for each random network; once for SI models and once for SIR models.

To gauge the accuracy and precision of our methods, we compared each distribution of outbreak durations on a given E-R network (hereafter, called the "observed network") to that of the original outbreak durations on the maximally complete network of the same size as the predicted ENS of the observed network (hereafter, "effective network"). We compared these distributions graphically (with enss functions "plot_compare_SI" and "plot_compare_SIR") and statistically (with enss functions "compare_SI_erg_ens" and "compare_SIR_erg_ens"). For accuracy, we compared the observed and effective network distributions in means of outbreak durations, with more similar means indicating that simulating disease spread on effective networks is more accurately capturing expected spread on the observed network. For precision, we compared the observed and effective network distributions in SDs of outbreak durations, with more similar SDs indicating that the precision of simulating disease spread on effective networks is similar to what would be obtained on the actual networks. We statistically compared the distributions of outbreak durations between observed and effective network simulations with Kolmogorov–Smirnov tests in R with package dgof (33). Significance on these tests indicates that the two distributions likely did not come from the same original distribution.

Associations Between ENS and Other Metrics

As one example application of our methods, we used our predictive models to estimate ENS of primate social networks that had been recorded in the literature (e.g., **Figure 1D**). These networks mainly consisted of the dataset of weighted sociomatrices collected by Griffin and Nunn (26), supplemented with more recent publications. Batch importing of empirical social networks was accomplished in enss using function "import_emp." A full listing of the sources for each of these networks, as well as the species and interaction type to which each corresponds, is provided in **Table 1**. ENS were again calculated by simulating SI and SIR models and then inputting the resulting outbreak duration means into the equations described in Section "Simulation and Regression of Disease Transmission on Maximally Complete Networks." For each of the empirical social networks, we then calculated weighted and unweighted versions of five common network metrics leading eigenvector modularity, which is a measure of how subdivided a network is into cliques, with higher values indicating more extreme subdivision, was calculated using function leading.eigenvector.community from package igraph (24). Mean distance, or the average of the shortest paths between each combination of two nodes, was calculated using function distance_w from package tnet (34); greater distances between nodes indicate that information will take longer to spread across the network. A related metric, diameter, measures the longest of these shortest paths across the entire network; it was calculated using function diameter from package igraph (24). Clustering coefficient, a measure of complete connectedness among triplets of nodes which have at least two connections among them, was calculated using function clustering_w from package tnet (34); higher clustering coefficients indicate that if three nodes are connected by at least two connections, they likely also include the third connection. Eigenvector centralization measures the skewness in the centrality, or connectedness of each node within the network, with higher values indicating greater skew from a uniform distribution of centralities; this was calculated using function evcent from package igraph (24). These metrics can be calculated for any set of networks using the "calculate_metrics" function in enss, which simply automates the calculations performed by functions provided in packages igraph (24) and tnet (34). Then, we compared models with all combinations of these metrics as

TABLE 1 | Raw and ENSs of primate species included in the established metric comparison models, as well as source information for each of the networks.

Species	Group size	ENS SI	ENS SIR	Weighted ENS SI	Weighted ENS SIR	Group status	Interaction class	Source
Alouatta caraya	5	7	8	9	7	Captive	Grooming	(36)
Ateles geoffroyi	15	36	22	75	23	Free-ranging	Grooming	(37)
Cebus apella	12	20	18	43	18	Wild	Grooming	(38)
Cebus capucinus	6	9	10	9	9	Wild	Grooming	(39)
Cercopithecus aethiops	8	11	13	15	12	Wild	Grooming	(40)
Cercopithecus mitis	16	43	21	57	26	Wild	Grooming	(41)
Colobus guereza	8	13	13	43	13	Wild	Grooming	(42)
Eulemur fulvus	11	16	16	20	15	Free-ranging	Proximity	(43)
Lemur catta	12	16	17	20	16	Wild	Proximity	(44)
Macaca arctoides	19	31	26	53	26	Captive	Grooming	(45)
Macaca assamensis	19	36	26	79	28	Wild	Grooming	(46)
Macaca fascicularis	10	20	15	70	17	Captive	Grooming	(47)
Macaca mulatta	28	34	35	37	30	Captive	Proximity	(48)
Macaca radiata	16	25	22	32	22	Wild	Grooming	(49)
Miopithecus talapoin	8	11	13	16	12	Captive	Grooming	(50)
Pan troglodytes	7	10	11	12	9	Wild	Grooming	(10)
Papio ursinus	14	24	21	27	18	Wild	Grooming	(51)
Saguinus fuscicollis	7	10	12	16	11	Captive	Grooming	(52)
Saguinus mystax	6	9	10	10	9	Wild	Grooming	(53)
Theropithecus gelada	7	15	12	16	10	Captive	Sociopositive	(54)

"Network size" is the count of nodes in the observed primate network. "ENS" indicates effective network size, with "SI" or "SIR" indicating the type of transmission model used for estimating ENS, and "weighted" indicating that tie weights were also included in simulations for estimating ENS. In some cases, weighted ENS measures were very different from their unweighted counterparts, indicating a strong effect of adding in tie weight information. In other cases, SI and SIR ENS estimates varied widely within species; these typically indicate an effect of removing non-total transmission simulations from SIR models, lowering ENS.
SI, susceptible-infected; SIR, susceptible-infected-recovered.

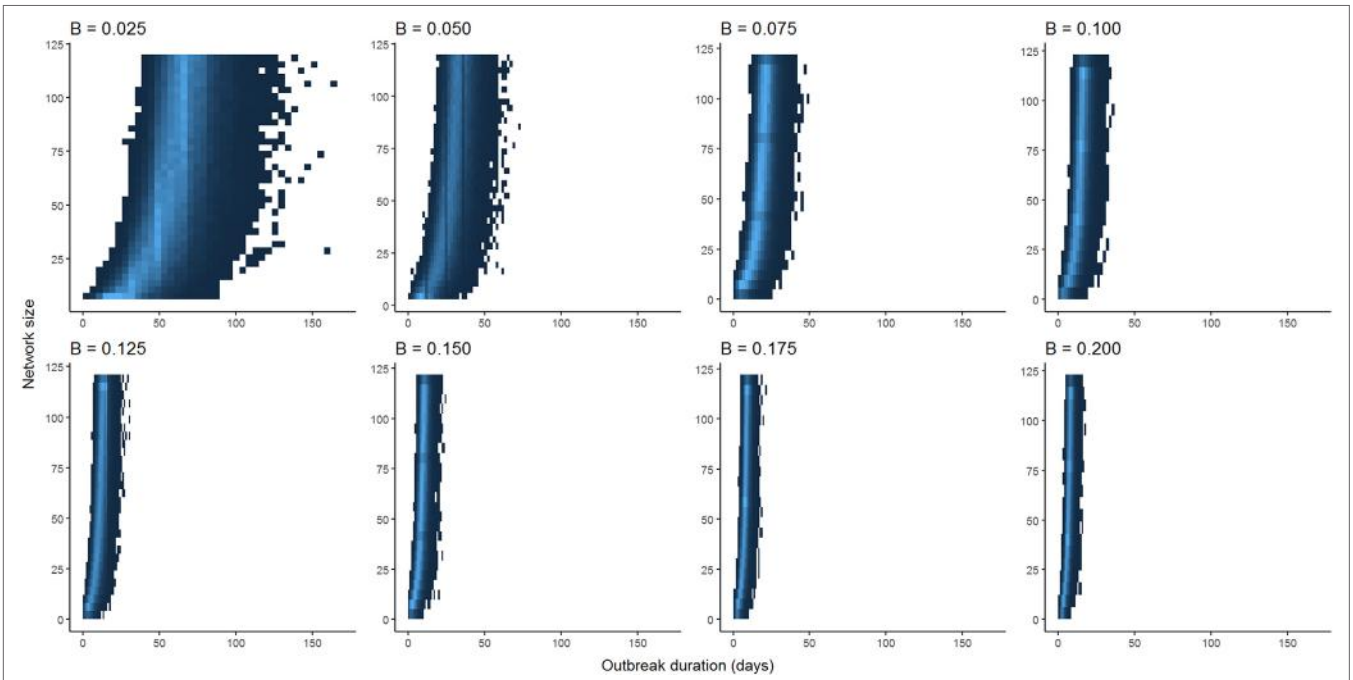


FIGURE 2 | Comparison between distributions of outbreak durations for susceptible-infected simulations with varying values for β. Lower values for β have larger ranges of outbreak durations, but the shapes of curves are qualitatively similar when scaled to the maximum outbreak duration for a given value of β.

predictors of ENS in a model comparison framework with AICc as the model selection criterion, using a cutoff of two AICc units for preferring a model over other models. AICc values were calculated in R with package MuMIn (35) and can be calculated in batch form with the enss function "AICc_ens_metrics."

RESULTS

Optimization of piecewise regression models estimated a break at a network size of 80 nodes, indicating that predictions of ENS above 80 individuals would be considerably less reliable than

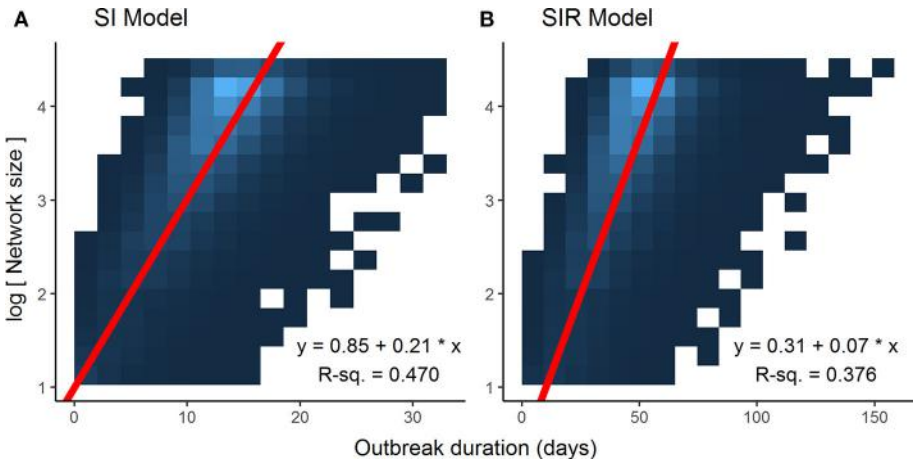


FIGURE 3 | Associations between log-transformed network size and outbreak duration for different disease models. Data points for each graph, limited to networks of 80 nodes or less (n = 78,000), were too dense to make scatterplot representations intelligible, thus heatmaps were used to illustrate the results, with lighter colors of blue representing a higher density of data points. Log-transforming network size makes for a linear relationship, and reduced major axis model 2 regression lines, represented in red, account best for the joint variation in the x and y axes. **(A)** Susceptible-infected (SI) model. **(B)** Susceptible-infected-recovered (SIR) model.

those of 80 or below. Furthermore, altering the values for β had no effect on the breakpoints, although as would be expected, the ranges of outbreak durations were inversely related to the value for β (**Figure 2**). All piecewise regressions revealed breaks at between 79.45 and 81.75 nodes. RMA model II regressions of log-transformed maximally complete network size versus outbreak duration for SI and SIR models fit relatively well, with R^2 of 0.470 and 0.376, respectively (**Figure 3**). The regression equations, listed in **Figures 3A,B**, were then used to calculate ENS. Alternative model results, with ties sampled regularly rather than randomly, showed similar results for SIR models, but tended to oversample ties in large networks for SI models, leading to unreasonably short outbreak durations in these networks (Figure S1 in Supplementary Material).

We then compared the distributions of E-R graph (observed) outbreak durations to those of their equivalent maximally complete (effective) network's outbreak durations to assess accuracy and precision. This was done to determine whether disease outbreaks on observed networks were accurate, or similar to those on maximally complete networks, in terms of the distributions of the outbreak durations from simulations on effective and observed networks. **Figure 4** shows the results of the SI model comparisons. Accuracy of our RMA predictive model was high, with means similar between observed and effective network outbreak durations (**Figure 4B**), but outbreak durations from observed network simulations showed higher SDs than those from effective networks (**Figure 4C**). Kolmogorov–Smirnov tests show that these two sets of distributions were often significantly different, with a critical value for the D-statistic at 0.60 (**Figure 4D**). However, this method is extremely sensitive to small changes in distributions and may not be best suited for determining similarity between the observed and effective network outbreak duration distributions.

Figure 5 shows the results of the SIR model comparisons between effective and observed network simulations. Again, similarity between mean values of outbreak durations for simulations on effective and observed networks (i.e., accuracy) was high (**Figure 5B**), but outbreak durations from observed network simulations actually showed lower SDs than those from effective networks (**Figure 5C**); this was likely due to the exclusion of simulations where the disease went extinct, which would have drastically reduced the variance of results. Kolmogorov–Smirnov tests show that these two sets of distributions were often significantly different, again with a critical value for the D-statistic at 0.60 (**Figure 4D**).

In our model selection framework comparing unweighted ENS to other established unweighted network metrics, the best fitting model for SI ENS included positive associations with raw group size ($b = 1.13$), mean distance ($b = 116.93$), and clustering coefficient ($b = 66.71$), as well as a negative association with eigenvector centralization ($b = -107.31$); the model had an adjusted R^2 of 0.971. There were four best fitting models for SIR ENS within two units of the minimum AICc value, and thus each of the following models were tied for best fit: SIR best fit #1 included positive associations with raw group size ($b = 1.15$), clustering coefficient ($b = 3.44$), and eigenvector centralization ($b = 17.61$); the model had an adjusted R^2 of 0.993. SIR best fit #2 included positive associations with raw group size ($b = 1.16$) and eigenvector centrality ($b = 20.27$), as well as a negative association with mean distance ($b = -4.25$); the model had an adjusted R^2 of 0.993. SIR best fit #3 included positive associations with raw group size ($b = 1.17$) and eigenvector centralization ($b = 17.23$), as well as a negative association with leading eigenvector modularity ($b = -13.35$); the model had an adjusted R^2 of 0.993. SIR best fit #4 included positive associations with raw group size ($b = 1.15$) and eigenvector centralization ($b = 5.69$); the model had an adjusted R^2 of 0.992.

Meanwhile, for the weighted models, the best fit for weighted SI ENS included positive associations with raw group size ($b = 2.40$) and mean weighted distance ($b = 51.26$), as well as a negative association with weighted diameter ($b = -15.15$); the model had an adjusted R^2 of 0.706. Again, there were four best fitting

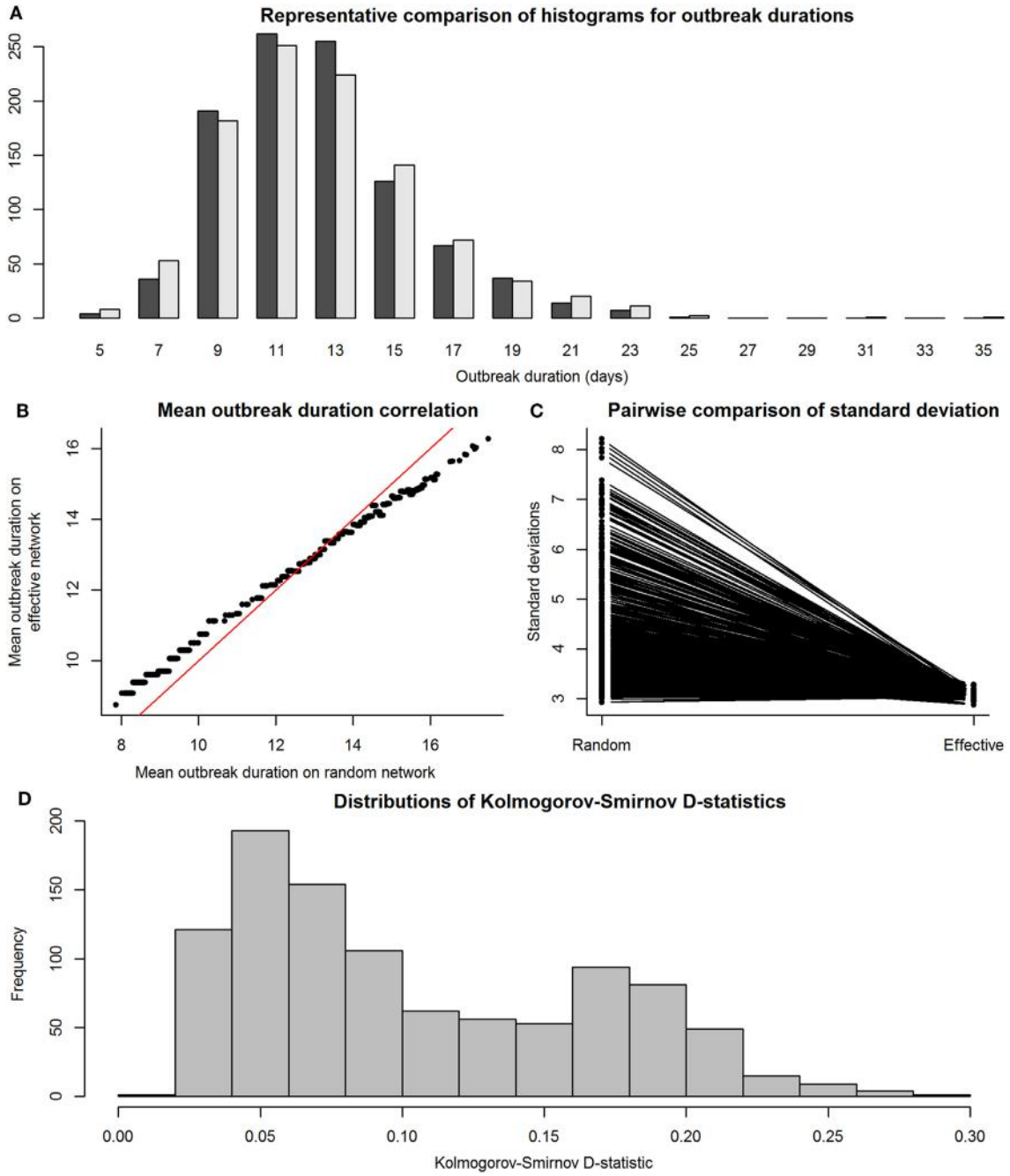


FIGURE 4 | Comparison between distributions of outbreak durations for susceptible-infected simulations on observed and effective network. Throughout the figure, the term "observed" refers to results from simulations on Erdős-Rényi graphs, and "effective" refers to results from simulations on reduced major axis-predicted equivalent maximally complete networks. Network sizes are limited to a maximum of 80 individuals, as this was the condition under which we were reasonably confident in our results. Panel **(A)**, a histogram with a representative pair of observed (dark gray) and effective (light gray) distributions of outbreak durations plotted together for viewing overlaps, shows that the distributions, compared on a pairwise scale had a considerable amount of overlap. Panel **(B)** shows means of outbreak durations from observed networks plotted against those from their predicted effective networks; red line indicates 1:1 equivalence, at which effective means match observed means. Panel **(C)** shows a paired line plot of SDs in outbreak durations for simulations on observed and effective networks; observed networks showed higher SDs than their paired effective networks. Panel **(D)** shows a histogram of Kolmogorov–Smirnov D-statistics for pairwise statistical comparisons between observed and effective network outbreak durations, with values above 0.60 indicating significantly different distributions.

models for weighted SIR ENS within two units of the minimum AICc value, and thus each of the following models were tied for best fit: weighted SIR best fit #1 included positive associations with raw group size ($b = 1.14$), leading eigenvector modularity ($b = 15.88$), and eigenvector centralization ($b = 4.29$); the model had an adjusted R^2 of 0.949. Weighted SIR best fit #2 included

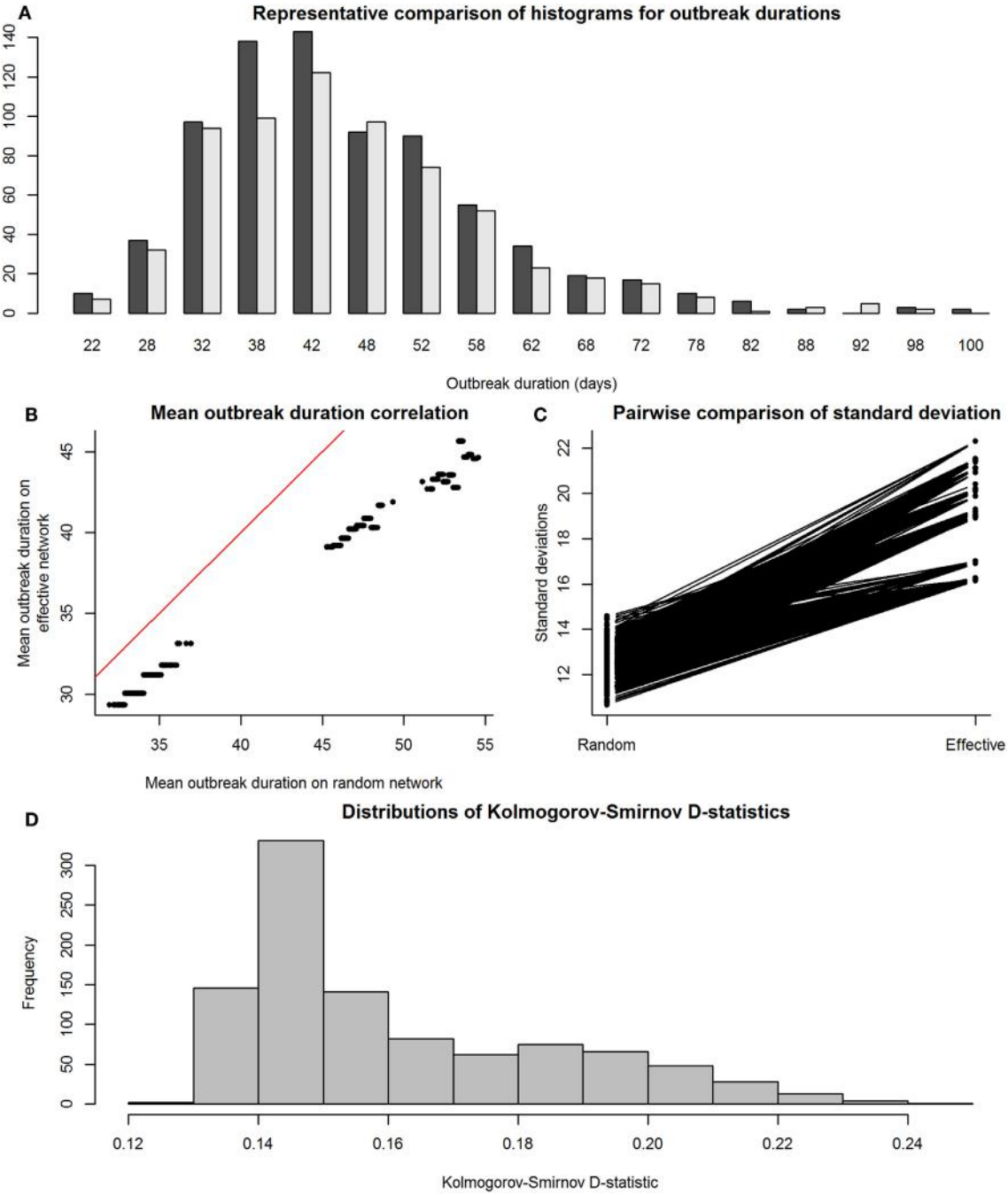


FIGURE 5 | Comparison between distributions of outbreak durations for susceptible-infected-recovered simulations on observed and effective network. Again, the term "observed" refers to results from simulations on Erdős-Rényi graphs, and "effective" refers to results from simulations on reduced major axis-predicted equivalent maximally complete networks. Network sizes are also limited to a maximum of 80 individuals, as this was the condition under which we were reasonably confident in our results. Panel **(A)**, a histogram with a representative pair of observed (dark gray) and effective (light gray) distributions of outbreak durations plotted together for viewing overlaps, shows that the distributions, compared on a pairwise scale had a considerable amount of overlap. Panel **(B)** shows means of outbreak durations from observed networks plotted against those from their predicted effective networks; red line indicates 1:1 equivalence, at which effective means match observed means. Panel **(C)** shows a paired line plot of SDs in outbreak durations for simulations on observed and effective networks; observed networks showed higher SDs than their paired effective networks. Panel **(D)** shows a histogram of Kolmogorov–Smirnov D-statistics for pairwise statistical comparisons between observed and effective network outbreak durations, with values above 0.60 indicating significantly different distributions.

positive associations with raw group size ($b = 1.16$) and mean weighted distance ($b = 5.30$), as well as a negative association with weighted diameter ($b = -1.84$); the model had an adjusted R^2 of 0.949. Weighted SIR best fit #3 included positive associations with raw group size ($b = 1.13$), mean weighted distance ($b = 1.83$), and leading eigenvector modularity ($b = 14.44$); the model had an

adjusted R^2 of 0.950. Weighted SIR best fit #4 included positive associations with raw group size ($b = 1.10$) and leading eigenvector modularity ($b = 18.15$); the model had an adjusted R^2 of 0.948.

DISCUSSION

These results demonstrate the potential for using ENS to compare infectious disease risk across groups of different sizes, including potentially for understanding disease transmission across a mosaic of many loosely connected groups within a larger meta-population structure, as well as for simplifying entire meta-populations to a single ENS. Previous studies have applied similar network-level metrics, like centrality and modularity, to the study of disease transmission through contact, grooming, and sociopositive networks in both wild and captive populations (32, 55–59). But nearly all of these measures capture only one aspect of networks, and they require this aspect to be considered in isolation from other important information about the network, specifically, its size. This issue is especially problematic for some metrics like modularity, whose value is mathematically positively associated with network size (22, 60). When compared to established network metrics, our single metric of ENS was best predicted by a combination of group size plus at least two other metrics. Thus, our measure of ENS provides a metric for disease transmissibility among individuals in a group that also accounts for the size of the group from which it was estimated. This differs from the previously mentioned approach by Caillaud et al. (12), which focused on understanding sub-group heterogeneity of meta-populations in light of epidemic thresholds. Specifically, our approach uses network structure and group size to predict how quickly a disease can be transmitted and maintained by individuals in a population.

Many more established network metrics covaried consistently with our measures of ENS, although there were differences most noticeably between transmission modes. For both SI and SIR ENS, the raw, original group size (number of nodes in the observed network) covaried strongly and positively with ENS, supporting the claim that ENS presents a novel, "size standardized" network metric. In addition, for SI models, both weighted and unweighted, mean distance was positively associated with ENS, perhaps indicating that SI models function through a simple diffusion process, where distance traveled is the best indicator of disease spread time. On the other hand, for SIR models, again both weighted and unweighted, network metrics like centralization and modularity, which generally indicate the skewness of tie distributions, showed generally positive associations with ENS. These relationships may point more toward the importance of skewness of connections in impeding or bottlenecking the spread of diseases specifically for SIR transmission models.

Of course, social networks can be represented in many ways, and our approach still simplifies networks considerably from their real-world manifestations. First, nearly all social ties in the real world vary in intensity (i.e., the networks are weighted), yet we conducted most of our tests using unweighted networks. The unweighted networks were used as a less "noisy" test of our methods. We did, however, also test for associations between ENS and other network metrics using weighted primate networks, which generally showed weaker effects compared with using unweighted ENS, likely due to the increased variation introduced by tie weights. Additional sources of variability are also worth considering. For example, individuals may vary in traits that make them more or less susceptible to a disease or to transmitting it, including trade-offs between reproductive status, dominance, and immune system, as well as age-related effects on immune function (59). Networks may also vary in their structure across time, adding yet another variable that complicates analyses (58, 61–63). However, the majority of research focuses on the importance of structural aspects of static networks for predicting and mitigating disease transmission, as this allows for more straightforward interpretation and comparison among different populations (64–66).

Additional applications of the method may open a variety of new routes for wildlife management and infection control. ENS could be used in disease outbreak risk assessments for wild or captive populations with known social networks. In addition, meta-populations of groups with known social networks could be simplified to their respective ENS to make prediction of future outbreaks easier in the future. Groups of sufficient size or structure could be targeted for vaccination campaigns in the wild or in captivity. In addition to comparing groups in a meta-population to one another, ENS could be used as a rough heuristic at a larger scale, reducing entire meta-populations to a single ENS. Finally, if further work is conducted to develop our method into a mathematical one rather than a simulation-based one, this approach could be applied to policy and management applications where simulation modeling is prohibitively time-consuming.

Although this study has only focused on simulation-based solutions for determining ENS, mathematical solutions for determining ENS should be investigated to obtain more succinct and resource-efficient calculations. One such approach for these mathematical solutions was shown by Caillaud et al. (12), but mathematicians and theoreticians interested in the effects of group size on disease transmission could still significantly further such research. In addition to this, the number of studies that have published social network structures is still small. For this reason, we encourage scientists researching social interaction to publish network information on species for which they already have data and to begin more studies of social network analysis in primate groups.

AUTHOR CONTRIBUTIONS

CM and CN designed the research and wrote the manuscript. CM developed R code and analyzed data.

ACKNOWLEDGMENTS

We thank Duke University for providing facilities for research, and Richard Wrangham, Joe Henrich, Hillary Young, Sean Riley, Kelsey Sumner, the editors of this research topic and our two reviewers for feedback. Kelsey Sumner also helped develop R code and shared primate social networks from her literature search.

FUNDING

CM was supported by Harvard University, the NSF Graduate Research Fellowship Program (DGE-1144152) and a Cora du Bois Charitable Trust Dissertation Writing Fellowship.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at https://www.frontiersin.org/articles/10.3389/fvets.2018.00071/full#supplementary-material.

REFERENCES

1. Wright S. Evolution in Mendelian populations. *Genetics* (1931) 16:97–159.
2. Felsenstein J. Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics* (1971) 68:581–97.
3. Crow JF. Wright and fisher on inbreeding and random drift. *Genetics* (2010) 184:609–11. doi:10.1534/genetics.109.110023
4. Weissman DB, Barton NH. Limits to the rate of adaptive substitution in sexual populations. *PLoS Genet* (2012) 8:e1002740. doi:10.1371/journal.pgen.1002740
5. Gomez-Uchida D, Palstra FP, Knight TW, Ruzzante DE. Contemporary effective population and metapopulation size (N e and meta-N e): comparison among three salmonids inhabiting a fragmented system and differing in gene flow and its asymmetries. *Ecol Evol* (2013) 3:569–80. doi:10.1002/ece3.485
6. Anderson RM, May RM. Population biology of infectious diseases: part I. *Nature* (1979) 280:361–7. doi:10.1038/280361a0
7. White LA, Forester JD, Craft ME. Using contact networks to explore mechanisms of parasite transmission in wildlife. *Biol Rev* (2017) 92:389–409. doi:10.1111/brv.12236
8. Keeling MJ, Eames KTD. Networks and epidemic models. *J R Soc Interface* (2005) 2:295–307. doi:10.1098/rsif.2005.0051
9. Cardy JL, Grassberger P. Epidemic models and percolation. *J Phys A Math Gen* (1985) 18:L267–71. doi:10.1088/0305-4470/18/6/001
10. Sugiyama Y. Grooming interactions among adult chimpanzees at Bossou, Guinea, with special reference to social structure. *Int J Primatol* (1988) 9:393–407. doi:10.1007/BF02736216
11. Kimura M, Saito K, Nakano R. Extracting influential nodes for information diffusion on a social network. In: Cohn A, editor. *AAAI'07 Proceedings of the 22nd National Conference on Artificial Intelligence*. Vol. 2. Vancouver: AAAI Press (2007). p. 1371–6.
12. Caillaud D, Craft ME, Meyers LA. Epidemiological effects of group size variation in social species. *J R Soc Interface* (2013) 10:20130206. doi:10.1098/rsif.2013.0206
13. Bartlett MS. The critical community size for measles in the United States. *J R Stat Soc Ser A* (1960) 123:37–44. doi:10.2307/2343186
14. Henrich J. Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses—the Tasmanian case. *Am Antiq* (2004) 69:197. doi:10.2307/4128416
15. Powell A, Shennan S, Thomas MG. Late Pleistocene demography and the appearance of modern human behavior. *Science* (2009) 324:1298–301. doi:10.1126/science.1170165
16. McCabe CM, Reader SM, Nunn CL. Infectious disease, behavioural flexibility and the evolution of culture in primates. *Proc Biol Sci* (2014) 282:20140862. doi:10.1098/rspb.2014.0862
17. Matsumura S. The evolution of "egalitarian" and "despotic" social systems among macacques. *Primates* (1999) 40:23–31. doi:10.1007/BF02557699
18. Chapman CA, Rothman JM. Within-species differences in primate social structure: evolution of plasticity and phylogenetic contraints. *Primates* (2009) 50:12–22. doi:10.1007/s10329-008-0123-0
19. Eisenberg JF, Muckenhirn NA, Rudram R. The relationship between ecology and social structure in primates. *Science* (1972) 176:863–74. doi:10.1126/science.176.4037.863
20. Wrangham RW. An ecological model of female-bonded primate groups. *Behaviour* (1980) 75:262–300. doi:10.1163/156853980X00447
21. Nunn CL, Altizer SM. *Infectious Diseases in Primates: Behavior, Ecology, and Evolution*. New York: Oxford University Press (2006).
22. Nunn CL, Jordán F, McCabe CM, Verdolin JL, Fewell JH. Infectious disease and group size: more than just a numbers game. *Philos Trans R Soc Lond B Biol Sci* (2015) 370:20140111. doi:10.1098/rstb.2014.0111
23. R Core Team. *R: A Language and Environment for Statistical Computing*. (2016). Available from: https://www.R-project.org/ (Accessed: April 18, 2017).
24. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal Complex Systems 1695* (2006) 1695.
25. Handcock MS, Hunter DR, Butts CT, Goodreau SM, Morris M. statnet: Software tools for the representation, visualization, analysis and simulation of network data. *J Stat Softw* (2008) 24:1–11. doi:10.18637/jss.v024.i01
26. Griffin RH, Nunn CL. Community structure and the spread of infectious disease in primate social networks. *Evol Ecol* (2012) 26:779–800. doi:10.1007/s10682-011-9526-2
27. Fraser C, Riley S, Anderson RM, Ferguson NM. Factors that make an infectious disease outbreak controllable. *Proc Natl Acad Sci U S A* (2004) 101:6146–51. doi:10.1073/pnas.0307506101
28. Muggeo VMR. Estimating regression models with unknown break-points. *Stat Med* (2003) 22:3055–71. doi:10.1002/sim.1545
29. Legendre P. *lmodel2: Model II Regression*. (2014). Available from: https://cran.r-project.org/package=lmodel2 (Accessed: April 18, 2017).
30. Hayes JP, Shonkwiler JS. Allometry, antilog transformations, and the perils of prediction on the original scale. *Physiol Biochem Zool* (2006) 79:665–74. doi:10.1086/502814
31. Smith RJ. Logarithmic transformation bias in allometry. *Am J Phys Anthropol* (1993) 90:215. doi:10.1002/ajpa.1330900208
32. Kasper C, Voelkl B. A social network analysis of primate groups. *Primates* (2009) 50:343–56. doi:10.1007/s10329-009-0153-2
33. Arnold TB, Emerson JW. Nonparametric goodness-of-fit tests for discrete null distributions. *R J* (2011) 3:34–9.
34. Opsahl T. *Structure and Evolution of Weighted Networks*. London, UK: University of London (Queen Mary College) (2009). p. 104–22. Available from: http://toreopsahl.com/publications/thesis/; http://toreopsahl.com/tnet/ (Accessed: April 18, 2017).
35. Barton K. *MuMIn: Multi-Model Inference*. (2016). Available from: https://cran.r-project.org/package=MuMIn (Accessed: April 18, 2017).
36. Jones CB. Social organization of captive black howler monkeys (*Alouatta caraya*): social competition and the use of non-damaging behavior. *Primates* (1983) 24:25–39. doi:10.1007/BF02381451
37. Ahumada JA. Grooming behavior of spider monkeys (*Ateles geoffroyi*) on Barro Colorado Island, Panama. *Int J Primatol* (1992) 13:33–49. doi:10.1007/BF02547726
38. Izawa K. Social behavior of the wild black-capped Capuchin (*Cebus apella*). *Primates* (1980) 21:443–67. doi:10.1007/BF02373834
39. Perry S. Female-female social relationships in wild white-faced capuchin monkeys, *Cebus capucinus*. *Am J Primatol* (1996) 40:167–82. doi:10.1002/(SICI)1098-2345(1996)40:2<167:AID-AJP4>3.0.CO;2-W
40. Seyfarth RM. The distribution of grooming and related behaviours among adult female vervet monkeys. *Anim Behav* (1980) 28:798–813. doi:10.1016/S0003-3472(80)80140-0
41. Cords MA. Agonistic and affiliative relationships in a blue monkey group. In: Whitehead PF, Jolly CJ, editors. *Old World Monkeys*. Cambridge: Cambridge University Press (2000). p. 453–79.
42. Dunbar RIM, Dunbar EP. Contrasts in social structure among black-and-white colobus monkey groups. *Anim Behav* (1976) 24:84–92. doi:10.1016/S0003-3472(76)80102-9
43. Jacobs A, Sueur C, Deneubourg JL, Petit O. Social network influences decision making during collective movements in brown lemurs (*Eulemur fulvus fulvus*). *Int J Primatol* (2011) 32:721–36. doi:10.1007/s10764-011-9497-8
44. Kendal RL, Custance DM, Kendal JR, Vale G, Stoinski TS, Rakotomalala NL, et al. Evidence for social learning in wild lemurs (*Lemur catta*). *Learn Behav* (2010) 38:220–34. doi:10.3758/lb.38.3.220

45. Butovskaya ML, Kozintsev AG, Kozintsev BA. The structure of affiliative relations in a primate community: allogrooming in stumptailed macaques (*Macaca arctoides*). *Hum Evol* (1994) 9:11–23. doi:10.1007/BF02438136
46. Cooper MA, Berntein IS, Hemelrijk CK. Reconciliation and relationship quality in Assamese macaques (*Macaca assamensis*). *Am J Primatol* (2005) 65:269–82. doi:10.1002/ajp.20114
47. Butovskaya M, Kozintsev A, Welker C. Conflict and reconciliation in two groups of crab-eating monkeys differing in social status by birth. *Primates* (1996) 37:261–70. doi:10.1007/BF02381858
48. Massen JJM, Sterck EHM. Stability and durability of intra- and intersex social bonds of captive rhesus macaques (*Macaca mulatta*). *Int J Primatol* (2013) 34:770–91. doi:10.1007/s10764-013-9695-7
49. Sugiyama Y. Characteristics of the social life of bonnet macaques. *Primates* (1971) 12:247–66. doi:10.1007/BF01730414
50. Wolfheim JH. A quantitative analysis of the organization of a group of captive talapoin monkeys (*Miopithecus talapoin*). *Folia Primatol* (1977) 27:1–27. doi:10.1159/000155773
51. King AJ, Clark FE, Cowlishaw G. The dining etiquette of desert baboons: the roles of social bonds, kinship, and dominance in co-feeding networks. *Am J Primatol* (2011) 73:768–74. doi:10.1002/ajp.20918
52. Vogt JL. The social behavior of a marmoset (*Saguinus fuscicollis*) group II: behavior patterns and social interaction. *Primates* (1978) 19:287–300. doi:10.1007/BF02382798
53. Löttker P, Huck M, Zinner DP, Heymann EW. Grooming relationships between breeding females and adult group members in cooperatively breeding moustached tamarins (*Saguinus mystax*). *Am J Primatol* (2007) 69:1159–72. doi:10.1002/ajp.20411
54. Dunbar RIM. Structure of social relationships in a captive gelada group: a test of some hypotheses derived from studies of a wild population. *Primates* (1982) 23:89–94. doi:10.1007/BF02381440
55. Borgatti SP. Centrality and network flow. *Soc Networks* (2005) 27:55–71. doi:10.1016/j.socnet.2004.11.008
56. Potterat JJ, Rothenberg RB, Muth SQ. Network structural dynamics and infectious disease propagation. *Int J STD AIDS* (1999) 10:182–5. doi:10.1258/0956462991913853
57. Romano V, Duboscq J, Sarabian C, Thomas E, Sueur C, MacIntosh AJJ. Modeling infection transmission in primate networks to predict centrality-based risk. *Am J Primatol* (2016) 78:767–79. doi:10.1002/ajp.22542
58. Rushmore J, Caillaud D, Matamba L, Stumpf RM, Borgatti SP, Altizer SM. Social network analysis of wild chimpanzees provides insights for predicting infectious disease risk. *J Anim Ecol* (2013) 82:976–86. doi:10.1111/1365-2656.12088
59. Cohen S, Doyle WJ, Skoner DP, Rabin BS, Gwaltney JM. Social ties and susceptibility to the common cold. *JAMA* (1997) 277:1940–4. doi:10.1001/jama.277.24.1940
60. Sah P, Leu ST, Cross PC, Hudson PJ, Bansal S. Unraveling the disease consequences and mechanisms of modular structure in animal social networks. *Proc Natl Acad Sci U S A* (2017) 114:4165–70. doi:10.1073/pnas.1613616114
61. Springer A, Kappeler PM, Nunn CL. Dynamic vs. static social networks in models of parasite transmission: predicting *Cryptosporidium* spread in wild lemurs. *J Anim Ecol* (2017) 86:419–33. doi:10.1111/1365-2656.12617
62. Read JM, Eames KTD, Edmunds WJ. Dynamic social networks and the implications for the spread of infectious disease. *J R Soc Interface* (2008) 5:1001–7. doi:10.1098/rsif.2008.0013
63. Hamede RK, Bashford J, McCallum H, Jones M. Contact networks in a wild Tasmanian devil (*Sarcophilus harrisii*) population: using social network analysis to reveal seasonal variability in social behaviour and its implications for transmission of devil facial tumour disease. *Ecol Lett* (2009) 12:1147–57. doi:10.1111/j.1461-0248.2009.01370.x
64. Craft ME. Infectious disease transmission and contact networks in wildlife and livestock. *Philos Trans R Soc Lond B Biol Sci* (2015) 370:20140107. doi:10.1098/rstb.2014.0107
65. Glass RJ, Glass LM, Beyeler WE, Min HJ. Targeted social distancing design for pandemic influenza. *Emerg Infect Dis* (2006) 12:1671–81. doi:10.3201/eid1211.060255
66. Andre M, Ijaz K, Tillinghast JD, Krebs VE, Diem LA, Metchock B, et al. Transmission network analysis to complement routine tuberculosis contact investigations. *Am J Public Health* (2007) 97:470–7. doi:10.2105/AJPH.2005.071936

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Veterinary Science

ORIGINAL RESEARCH
published: 05 April 2017
doi: 10.3389/fvets.2017.00046

Estimation of Time-Dependent Reproduction Numbers for Porcine Reproductive and Respiratory Syndrome across Different Regions and Production Systems of the US

Andréia G. Arruda[1], Moh A. Alkhamis[2], Kimberly VanderWaal[1], Robert B. Morrison[1] and Andres M. Perez[1]*

[1] *Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota, St Paul, MN, USA,*
[2] *Environment and Life Sciences Research Center, Kuwait Institute for Scientific Research, Kuwait City, Kuwait*

Porcine reproductive and respiratory syndrome (PRRS) is, arguably, the most impactful disease for the North American swine industry, due to its known considerable economic losses. The Swine Health Monitoring Project (SHMP) monitors and reports weekly new PRRS cases in 766 sow herds across the US. The time-dependent reproduction number (TD-R) is a measure of a pathogen's transmissibility. It may serve to capture and report PRRS virus (PRRSV) spread at the regional and system levels. The primary objective of the study here was to estimate the TD-R values for PRRSV using regional and system-level PRRS data, and to contrast it with commonly used metrics of disease, such as incidence estimates and space–time clusters. The second objective was to test whether the estimated TD-Rs were homogenous across four US regions. Retrospective monthly incidence data (2009–2016) were available from the SHMP. The dataset was divided into four regions based on location of participants, and demographic and environmental features, namely, South East (North Carolina), Upper Midwest East (UME, Minnesota/Iowa), Upper Midwest West (Nebraska/South Dakota), and South (Oklahoma panhandle). Generation time distributions were fit to incidence data for each region, and used to calculate the TD-Rs. The Kruskal–Wallis test was used to determine whether the median TD-Rs differed across the four areas. Furthermore, we used a space–time permutation model to assess spatial–temporal patterns for the four regions. Results showed TD-Rs were right skewed with median values close to "1" across all regions, confirming that PRRS has an overall endemic nature. Variation in the TD-R patterns was noted across regions and production systems. Statistically significant periods of PRRSV spread (TD-R > 1) were identified for all regions except UME. A minimum of three space–time clusters were detected for all regions considering the time period examined herein; and their overlap with "spreader events" identified by the TD-R method varied according to region. TD-Rs may help to measure PRRS spread to understand, in quantitative terms, disease spread, and, ultimately, support the design, implementation, and monitoring of interventions aimed at mitigating the impact of PRRSV spread in the US.

Keywords: time-dependent reproductive number, surveillance, porcine reproductive and respiratory syndrome, space–time clusters, porcine reproductive and respiratory syndrome incidence

OPEN ACCESS

Edited by:
Saraya Tavornpanich, Norwegian Veterinary Institution, Norway

Reviewed by:
Sara Amirpour Haredasht, University of California Davis, USA
Mathieu Andraud, Anses, France

****Correspondence:***
Andréia G. Arruda
arrud002@umn.edu

Specialty section:
This article was submitted to Veterinary Epidemiology and Economics, a section of the journal Frontiers in Veterinary Science

Received: *23 November 2016*
Accepted: *21 March 2017*
Published: *05 April 2017*

Citation:
Arruda AG, Alkhamis MA, VanderWaal K, Morrison RB and Perez AM (2017) Estimation of Time-Dependent Reproduction Numbers for Porcine Reproductive and Respiratory Syndrome across Different Regions and Production Systems of the US. Front. Vet. Sci. 4:46. doi: 10.3389/fvets.2017.00046

INTRODUCTION

Although porcine reproductive and respiratory syndrome (PRRS) is, arguably, one of the most important diseases of swine affecting the North American industry, aspects of its transmission within production systems and within regions are not completely understood (1). Even though PRRS is endemic in North America, recurrent emergence of new PRRS virus (PRRSV) strains results in an epidemiological dynamic that resembles an epidemic condition for the disease (2, 3). PRRSV epidemics impact the swine industry and commonly require prompt mobilization of resources for diagnostics (i.e., sequencing of the virus), thorough investigations to understand the origin of the emerging PRRSVs, and implementation of effective control measures.

Surveillance is an integral part of strategies for control and elimination of PRRSV. There are a number of surveillance activities currently in place in the US; however, because PRRS is not reportable, surveillance strategies vary dramatically according to factors such as region and production system. A few examples of such surveillance activities are ongoing monitoring in breeding herds and gilt development units, and passive surveillance triggered by clinical symptoms.

The concept of near real-time disease surveillance is important in the context of emerging PRRSV strains given that rapid identification of an epidemic (i.e., emergence of novel PRRSV strains) will likely result in a reduction of outbreak duration due to timely implementation of prevention and control measures to decrease virus spread within and across geographical regions.

In the absence of a regulatory framework, initiatives aimed at monitoring PRRS in North American swine farms are voluntary in nature. One example of an effort intended to coordinate surveillance efforts in the US at the national level is the Swine Health Monitoring Project (SHMP). The SHMP is a voluntary project that aims to monitor the incidence of PRRS; it currently enrolls approximately 42% of the US sow population distributed in 19 states in the country. Interpretation of collected data to participants and the swine industry currently focuses on incidence. Additionally, the number of new cases and spatial–temporal clustering have been previously investigated and reported to describe PRRS trends and to identify PRRS epidemics (4–6). However, the rate of new cases over time, referred to as incidence, serves as a proxy for risk but does not contribute as a metric for the epidemic progression or prediction of its evolution.

There are other methods, however, that could serve as proxy for disease progression and that have not been sufficiently explored in measuring PRRS transmissibility. The basic reproductive number (R_0) refers to the average number of secondary infections caused by a primary case and is commonly used to characterize the transmissibility potential of a disease in a completely susceptible population (7). In contrast, the effective reproductive number (R_e) can be used to characterize transmissibility once a certain proportion of the population has been infected and is resistant (immune) (8), which would be an example for the case of PRRS in the US. The time-dependent reproduction number (TD-R) is a measure of disease transmissibility that can be estimated over the course of disease progression (9). The TD-R has been particularly useful for monitoring epidemic trends, identifying "super-spreader events," measuring progress of interventions over time and for providing parameters for mathematical models (e.g., models to test interventions) (10).

The overall hypothesis of this study was that PRRS transmissibility, as measured by the TD-R, would not differ between regions and swine production systems within the US. This result would indicate that epidemiological dynamics are somewhat synchronized across regions, either because of seasonal weather changes or high connectivity among regions due to animal movements, as opposed to each region experiencing distinct temporal dynamics. Thus, our primary objective was to estimate the TD-R values for PRRSV using regional and system-level PRRS data from across the US, and to contrast it to incidence estimates and commonly investigated space–time clusters. We hypothesized that the peaks on the TD-R, incidence, and the space–time clusters would overlap. Furthermore, the secondary objective was to test whether the estimated TD-R were homogenous across four US regions. For this objective, the hypothesis was that the TD-R would be homogenous across all regions. Ultimately, results presented here will contribute to support the design and implementation of strategies for PRRS surveillance and control in the US.

MATERIALS AND METHODS

Data Source

Source of data for the study here was the SHMP, which includes a cohort of farms that voluntarily agree to share PRRS data weekly. The PRRS status captured in this dataset followed slightly modified guidelines described elsewhere (11). Briefly, status included categories 1, 2, 2fvi, 2vx, 3, and 4. Status 1 designates actively infected herds (in which pigs were shedding the virus), status 2 indicates stable herds (no shedding detected in weaned pigs after following certain sample size requirements); 2fvi and 2vx refer to herds that were using live-virus exposure or modified live-virus vaccination as control strategies, respectively; status 3 defines herds that were provisionally negative (negative gilts introduced into the herd and remained negative); and status 4 denotes herds that were seronegative.

Four areas across the US were chosen, including farms located in the states of North Carolina [South East (SE)], Oklahoma [South (S)], Minnesota/Iowa [Upper Midwest East (UME)], and Nebraska/South Dakota [Upper Midwest West (UMW)], and some neighboring locations (**Figure 1**). Those regions represented areas within the US characterized by high (SE and UME) and low (S and UMW) swine density, as reflected by the FAO's GeoNetwork data repository for global livestock densities (**Figure 1**).

Estimation of TD-R

The TD-R was estimated over time for farms participating in the SHMP project. Values of TD-R $>$ 1 were interpreted as an indication that the number of cases would increase over time (propagating phase of an epidemic), whereas values of TD-R $<$ 1

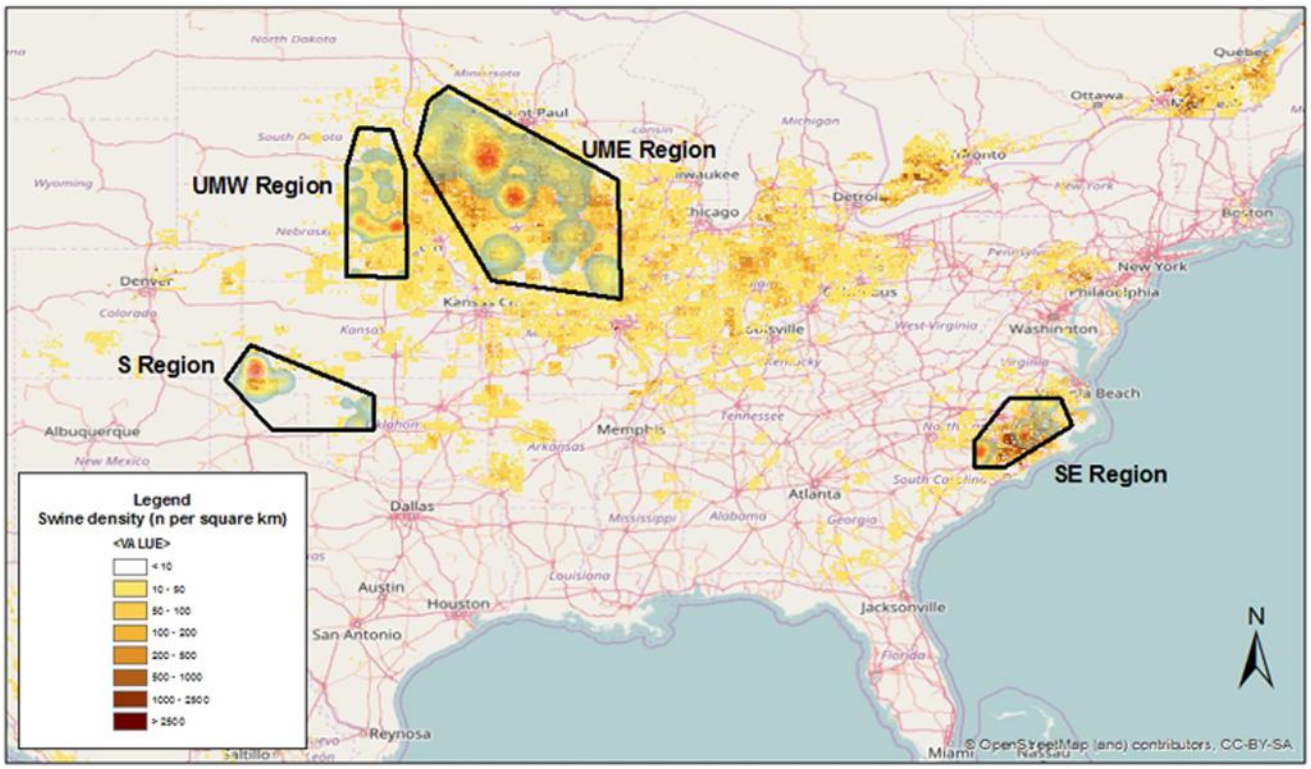


FIGURE 1 | **Kernel smoothed density of swine sites used for this project overlaid on swine animal density as modeled by the FAO's GeoNetwork data repository**.

served as an indication that the epidemic was fading out (8). The estimation of reproductive numbers is commonly considered an indirect process, because the parameters needed (e.g., contact rate and probability of transmission given contact) are usually difficult, if not impossible, to estimate. Here, data available to compute the TD-R included weekly number of cases reported from July 2009 through March 2016.

Effective time-dependent reproductive numbers were estimated from observed incidence data using a likelihood-based procedure described elsewhere (8) and implemented through the R package "R0" in R v.3.2.3 (9). In summary, the TD-R was calculated based on averaging over all transmission networks compatible with the observed cases (9).

Firstly, incidence data were aggregated at the monthly level to reduce the prevalence of time intervals with 0 values in the time series. For months in which no cases were reported, the count of new cases was set to 1 with the assumption that at least one outbreak was missed, which is a reasonable assumption for PRRS, because sow herds have different levels of immunity due to variable management strategies, and these can impact detection of disease. Secondly, the generation time distribution that best fit the observed occurrence of cases was estimated. This refers to the time between detection of a primary case and detection of a secondary case (8), and in our case, we considered the time lag between consecutive reported outbreaks and estimated its mean and SD from the observed epidemic curve using a function in R (9). Thirdly, the number of secondary cases for each case was estimated by averaging over all transmission chains compatible with the epidemic curves during the course of epidemics. This was done in two steps (8):

First, the probability that a certain reported outbreak i (that occurred at a certain time) was infected by another reported outbreak j (occurring at a previous time) was calculated by $p_{ij} = w(t_i - t_j)/\sum_{i \neq k} w(t_i - t_k)$; where w corresponds to the generation time distribution, and $t_i - t_k$ corresponds to the difference in time of recording of outbreaks i and j.

Second, the TD-R for reported outbreak j was calculated by the sum over all outbreaks i weighted by the likelihood that outbreak i was infected by outbreak j: $R_j = \sum_i p_{ij}$; and this was finally averaged considering all reported outbreaks with the same date of recording (9): $1/N_t \sum_{\{tj=t\}} R_j$.

Confidence intervals (CIs) were obtained by simulation; and statistically significant periods of PRRS spreading were defined as periods [month(s)] for which the TD-R's 95% CI did not include 1.

Time-dependent reproductive numbers were described separately for the four investigated geographical areas, as well as for each participating production system (20 systems represented by letters A–T). A production system was defined as two or more swine sites with a common owner or management structure. The Kruskal–Wallis test was used to determine whether the median TD-Rs differed across the four areas. Furthermore, the

Dunn's test of multiple comparisons (12) was applied, adjusting for multiple comparisons using the Bonferroni correction method. All statistical analyses were performed using STATA/IC version 14.1.

Space–Time Permutation Model

Clustering of cases in space and time was explored using the permutation model of the scan statistic (13) implemented using the SaTScan™v.9.4.2 software (14). Briefly, the permutation model of the scan statistic compares the number of observed cases in any candidate cluster to the number of cases that would had been expected if the spatial and temporal location of all outbreaks were evenly distributed so that no space–time dependency occurred. The scan statistic has been proposed (15) as a surveillance tool to track clusters of disease, and it is especially useful because it does not require information on the background population at risk (16). Statistically significant clusters were declared when $P < 0.05$.

RESULTS

The number of outbreaks varied according to region, with SE and UME (North Carolina and Minnesota/Iowa), the most swine densely populated regions of the country, reporting the highest number of outbreaks over the 2009–2016 period (**Table 1**). A given swine site may have had more than one outbreak through the study period; the number of outbreak per site reporting an outbreak was higher for the S and UMW regions (1.76 and 2.21 outbreaks per site, respectively) when compared to SE and UME (1.44 and 1.48 outbreaks per site, respectively). Those two last areas, however, did contribute with a larger number of months of data (**Table 1**).

The generation time distribution followed a lognormal distribution for the regions of SE (mean 1.30 months, SD: 1.26), S (mean: 1.09 months, SD: 1.01), UME (mean: 7.93 months, SD: 7.76) and UMW (mean months: 1.30, SD: 1.23).

The median and mean values for TD-R were similar across all regions and oscillated around 1.0, which is expected for endemic diseases. Interestingly, even though the mean and median values were close to 1 for all regions, incidence peaks and temporal variation in TD-R appeared remarkably different (**Figure 2**). A difference was observed in regards to the maximum number of TD-R values observed across regions; specifically, the TD-R was highest for SE, followed by S, UMW, and UME (**Table 1**). There were also remarkable differences in PRRS immune status classification for sites reporting outbreaks across the four regions (**Table 1**): of note; for the S region, the vast majority of sites reporting outbreaks were vaccinating the herd prior to the outbreak, which was not observed in such proportion for other regions. The SE region had a higher proportion of sites breaking that were classified as status 1 (active infection) when compared to sites that reported outbreaks from other regions (**Table 1**).

Between-region difference on median TD-R values was evident on the Kruskal–Wallis test, which showed that at least one of the regions had a different median ($P = 0.03$). Further *post hoc* pairwise comparison showed that the UME region was only statistically significant from the SE and UMW regions (**Table 2**).

It has been previously reported that PRRS has an evident seasonal pattern, showing high incidence during fall and winter (October through January), and low during spring and summer [February through September (5)]. Surprisingly, after stratifying the data by geographical region, there was no obvious visual indication of predictable yearly patterns for any of the regions besides Minnesota/Iowa (**Figure 2**).

The TD-R description showed variation according to geographical region, a phenomenon similar to the one previously described for the incidence estimate (**Figure 2**). The TD-R values showed statistically significant peaks before the incidence peaked for SE, S, and UMW (**Figure 2**). Interestingly, when comparing raw number of new cases or incidence with the TD-R estimates, all statistically significant peaks of TD-R ($P < 0.05$) preceded a meaningful increase in the number of cases (>2) for the regions of SE and S, showing the potential the tool has for early signaling outbreaks (**Figure 2**). The TD-R and the incidence peaks

TABLE 1 | Basic regional descriptors and description of time-dependent reproduction number (TD-R) values calculated in the study for porcine reproductive and respiratory syndrome (PRRS) transmissibility between swine sites located across four different regions of the US.

Region	*N* sites[a]	Period (months)[b]	*N* cases[c]	Median[d]	Mean (SD)[d]	Max [95% confidence interval (CI)][d]	PRRS status before outbreak[e] (% of sites reporting an outbreak)					
							1	2	2fvi	2vx	3	4
SE	72	81	104	0.99	1.14 (0.73)	5.42 (2.00, 9.00)	25.0	17.3	2.9	18.3	4.8	31.7
S	42	76	74	1.0	1.14 (0.54)	3.22 (1.00, 6.00)	0	4.0	0	85.1	0	10.8
UME	218	81	324	1.12	1.30 (0.68)	2.22 (0.45, 4.47)	5.4	8.8	39.3	20.5	7.7	18.3
UMW	38	76	84	1.002	1.10 (0.52)	2.80 (1.00, 5.00)	8.3	7.1	26.2	14.3	17.9	26.2

SE, South East (North Carolina); S, South (Oklahoma); UME, Upper Midwest East (Minnesota/Iowa); UMW, Upper Midwest West (Nebraska/South Dakota).

[a]*Number of swine sites.*

[b]*Number of months the region contributed with data.*

[c]*Number of incident cases from 2009 to 2016.*

[d]*Median, mean (SD), and maximum (95% CI) for TD-R values calculated in this study.*

[e]*Status is according to AASV guidelines: status 1 designates actively infected herds (in which pigs were shedding the virus); status 2 indicates stable herds (no shedding detected in weaned pigs after following certain sample size requirements); 2fvi and 2vx refer to herds that were using live-virus exposure or modified live-virus vaccination as control strategies, respectively; status 3 defines herds that were provisionally negative (negative gilts introduced into the herd and remained negative); and status 4 denotes herds that were seronegative.*

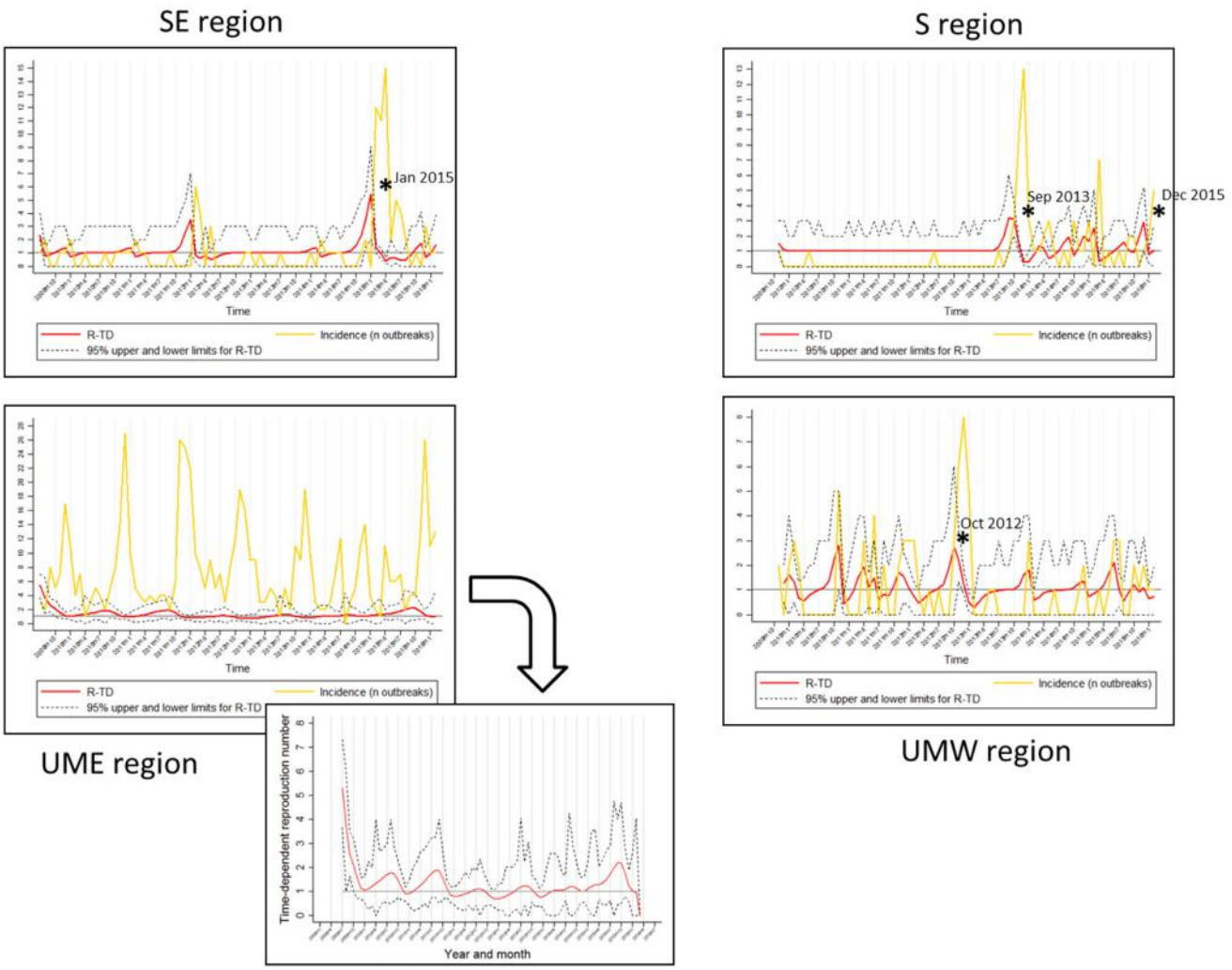


FIGURE 2 | Regional graphs showing time-dependent reproduction number (TD-R) and incidence for porcine reproductive and respiratory syndrome. The *Y* axis corresponds to TD-R with 95% upper and lower confidence intervals (CIs) (red) and number of new cases (yellow; represented after replacing 0 counts with "1's" as described in Section "Materials and Methods"). Stars (*) represent peaks on TD-R in which the 95% CI did not include 1; region Upper Midwest East (UME) has a zoom-out on the TD-R values to improve visualization.

TABLE 2 | Multiple pairwise comparisons for Kruskal–Wallis test using Dunn's test of multiple comparisons, showing the estimate (*P*-value), with applied Bonferroni correction.

Region	South East	South (S)	Upper Midwest East (UME)
S	−0.99 (0.96)	–	–
UME	−3.38 (0.002)*	−2.35 (0.0565)	–
Upper Midwest West	−0.42 (1.0)	0.56 (1.0)	2.91 (0.0107)

**Statistically significant difference.*

occurred at approximately the same time for the UMW region, and there were no statistically significant TD-R peaks for the region of MN/IA (**Figure 2**). For instances in which the TD-R peaks preceded the incidence peak, the lag time between peaks varied between 1 and 2 months. Likewise, indications that the epidemic was waning (TD-R < 1) occurred 2 months earlier than declines in incidence for SE and S (March 2015 versus May 2015 and November 2013 versus January 2014, respectively), and 1 month earlier for UMW.

Finally, the 20 systems examined contributed with a population at risk of on average 35 farms (min: 7, max: 83) per system. The average number of outbreaks per farm reporting at least one outbreak was 48.15 (min: 12, max: 189). Separate system-specific TD-R appeared to vary (**Table 3**; **Figure 3**), even for systems located within the same geographical region. Four systems were selected to illustrate differences between TD-R and incidence curves (systems A–D, **Figure 3**). Systems C and D, for example, were located within the same geographical region and showed one peak within the same time period (February 2015), even though the TD-R peak was not significant for system D. System

TABLE 3 | Time-dependent reproduction number summary estimates for each system enrolled in the SHMP project.

System	*N* months[a]	Mean	Median	SD	Min	Max
A	80	1.21	1.09	0.53	0.34	2.60
B	34	1.36	1.28	0.78	0.29	3.55
C	80	1.12	1.00	0.61	0.20	4.31
D	22	1.19	1.18	0.58	0.45	2.77
E	80	1.09	1.00	0.41	0.32	2.10
F	80	1.04	1.00	0.31	0.33	3.00
G	80	1.19	1.00	0.63	0.32	3.62
H	80	1.04	1.00	0.31	0.25	2.42
I	80	1.04	1.00	0.29	0.25	2.42
J	80	1.15	1.00	0.62	0.28	3.97
K	80	1.02	1.00	0.23	0.33	2.93
L	80	1.03	1.00	0.32	0.33	2.99
M	53	1.07	1.00	0.58	0.24	5.00
N	80	1.01	1.00	0.22	0.5	2.00
O	80	1.07	1.00	0.47	0.33	3.52
P	80	1.05	1.00	0.37	0.38	2.80
Q	45	1.07	1.00	0.44	0.41	2.72
R	57	1.11	1.00	0.53	0.27	3.98
S	36	1.14	1.00	0.84	0.23	5.81
T	45	1.06	1.00	0.42	0.41	2.72

[a]*Number of months the systems are participating in the SHMP.*

A showed no significant peaks on TD-R, but it showed frequent increases in incidence; and system B showed a peak in TD-R not observed when farms are aggregated at the region level.

The spatial–temporal model showed statistically significant clusters for all examined regions (**Table 4**). There were 3 clusters in space and time for the SE region, 3 clusters for the S region, 10 clusters for UME, and 4 clusters for UMW.

DISCUSSION

The study here is the first to investigate and report the use of the time-dependent reproductive number for PRRS reporting purposes, and to contrast it with commonly used methods for describing PRRS epidemics (i.e., number of cases and spatial–temporal cluster detection). Strengths of this study include the availability of monthly PRRS incidence data from a large number of US swine herds spread across different geographical regions, as well as the inclusion of a large number of swine production systems.

Results support the observation that region-level insights cannot be provided by using data that are aggregated from large national projects. Furthermore, regional-level control and prevention strategies should not be made based on the assumption that PRRS transmission dynamics are the same across geographical regions of the same country. Stratification of data would be able to provide a better estimate on which control and prevention measures, if any, would work best and provide the best benefit for specific regions.

Comparison of PRRS transmissibility across regions and production systems has not been previously reported for PRRS, and the statistical differences among TD-R estimates between regions were somewhat surprising, given that it is commonly believed that all regions have similar PRRS transmissibility patterns. Some reasons that might explain the observed differences include climatic factors (e.g., temperature variation), demographic and biosecurity factors (e.g., presence of filtered farms), swine density, the presence of different production systems in the areas, and the potential introduction of PRRS strains in differing instances.

The commonly expected predictable yearly increase pattern for PRRS was not visually evident for all geographical regions across years, nor was the time periods in which PRRS was spreading (defined as the TD-R 95% CI did not include 1; **Figure 2**). Interestingly, even though both the SE and UME regions are known to have high swine density, the patterns of PRRS transmission between them were different (**Figure 2**). However, PRRS management strategies within these two regions are known to differ, which may partially explain the findings: first, the immunity status of swine sites is anecdotally observed to be different among areas. For example, among high swine dense areas (SE and UME), it is believed that a certain amount of herd immunity exists in the SE region compared to the UME because producers in the latter area are more willing to attempt PRRS elimination from herds. In contrast, producers from the SE region are commonly using vaccination or live-virus inoculation strategies to mitigate PRRS impact (SHMP data not shown). However, it was observed that, even when certain amount of underlying immunity existed for the SE region, spreading events still occurred. This is also anecdotally observed from field veterinarians and producers. Another difference between the areas might be the use of farm filtration as a preventive measure for PRRS outbreaks, with the SE area being characterized by lower frequency of filtered farms compared to the UME area. Data gather on these and other management and biosecurity factors for future projects might help elucidating regional differences described herein.

Porcine reproductive and respiratory syndrome epidemic events were recognized by the TD-R method for all regions except for the UME. These events possibly reflected the introduction of new or previously undetected PRRSV strains in the SE, S, and UMW areas (10, 17). Overall, the TD-R appeared to be particularly useful for areas where the occurrence of outbreaks is sporadic, perhaps resembling an epidemic nature. In such cases, the TD-R appeared to flag outbreaks of new strains earlier when compared to the crude increase in the number of cases (**Figure 2**), which could be valuable for near real-time disease surveillance in the context of commonly emerging PRRSV strains. The use of TD-R could aid in the rapid detection of these episodes, which, combined with communication and mobilization with key industry stakeholders, could result in faster control of disease in a region. For areas where PRRS can be characterized mostly as having endemic nature, the use of the TD-R might still be useful for signaling epidemic progression, characterizing transmissibility over time, and identifying the occurrence of "super-spreader" events.

The relatively large number of spatial–temporal clusters was not surprising. Analysis of data from a regional control project in Minnesota reported that, despite an overall decrease in PRRS incidence from 2012 to 2015, significant spatial–temporal clusters of disease incidence over 3-week periods and 3-km radii were found (5). The occurrence of spatiotemporal cluster did not

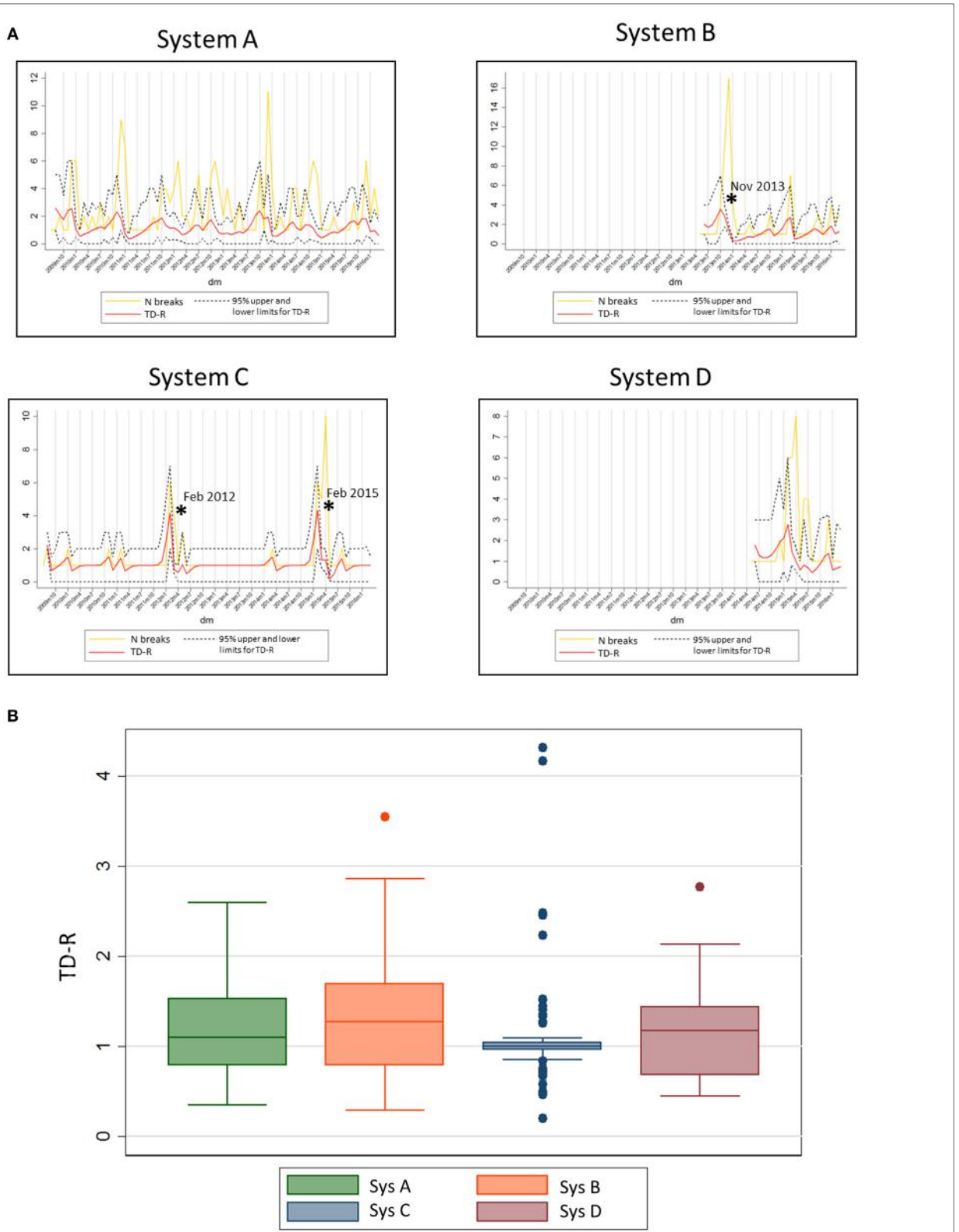


FIGURE 3 | **(A)** System-specific graphs showing time-dependent reproduction number (TD-R) and incidence for porcine reproductive and respiratory syndrome. TD-R with 95% upper and lower confidence intervals (CIs) are shown in red, and incidence curves (after replacing 0 counts with "1's" as described in Section "Materials and Methods") are shown in yellow. Stars (*) represent peaks on TD-R in which the 95% CI did not include 1. **(B)** Box plot showing TD-R distribution for four different systems of the US.

TABLE 4 | Super-spreader events and clusters found for the four regions by three different methods [time-dependent reproduction number (TD-R) estimation, purely temporal cluster detection, and spatial–temporal cluster detection].

Area[a]	TD-R[b]	Spatiotemporal cluster detection[c]	Radius of cluster (km)	O[d]	E[d]	*P*-value
SE	12/2014–01/2015	02/2015–07/2015	89	47	17.2	<0.001
		02/2012–07/2013	15	26	9.5	<0.001
		11/2010–12/2010	3.6	5	0.3	0.007
S	12/2015	08/2014–11/2014	42	10	1.9	<0.001
	10/2013	02/2014–05/2014	96	12	2.9	0.002
		11/2009–06/2012	7	4	0.1	0.007
UME		09/2015	4.3	22	1.7	<0.001
		07/2015–11/2015	44	35	9.7	<0.001
		01/2015–03/2015	22	12	0.6	<0.001
		04/2013–05/2013	44	15	0.7	<0.001
		09/2012	23	10	0.6	<0.001
		03/2012–05/2012	132	26	4.39	<0.001
		10/2011	3	15	1.52	<0.001
		07/2011–08/2011	65	13	0.7	<0.001
		05/2010–11/2010	94	47	13	<0.001
		01/2010–02/2010	7	31	6.5	<0.001
UMW	11/2012	06/2015–11/2015	42	15	1.9	<0.001
		05/2012–10/2012	31	10	1.6	<0.001
		06/2011	2	8	0.9	<0.001
		02/2010	28	14	3.9	0.01

[a]Area 1 corresponds to North Carolina, area 2 corresponds to Oklahoma panhandle, area 3 corresponds to Minnesota/Iowa, and area 4 corresponds to Nebraska/South Dakota.

[b]Epidemic events as defined by TD-R (8); 95% confidence interval does not include 1.

[c]Spatial–temporal cluster detection using the spatial–temporal permutation model (14).

[d]Observed (O) and expected (E) number of cases.

overlap with the detection of peaks in TD-R as expected (**Table 4**) but usually was recognized later than the first. At times, these clusters were quite frequent and lasted for a long period of time, which raises the point to whether the alarms they may trigger would be of concern or not.

Finally, system-specific estimates of TD-R showed recognizable peaks for systems C and D (**Figure 3**). These peaks corresponded to a known incursion of an emerging strain for the area. In addition, for system C, there is empirical evidence that intense breaks occur every 3 years, which was evidenced by our analysis. System A was characterized by multiple outbreaks over time, even though statistically significant PRRS spread periods were not detected. Predictable yearly increase patterns were visually suggestive for this area, except for the most recent years. Finally, system B appeared to have had a large outbreak in the end of November of 2013, which once more is anecdotally thought to be due to the incursion of an emerging strain in the region. The authors hypothesize that the reason why no further considerable outbreaks were observed after these is a combination between control measures being taken after the epidemic event, and the existence of a certain level of immunity in the herd after infection. For future studies, collection of such information is important to allow for testing of these and other hypotheses. We also recognize that, at time of writing of this manuscript, peer-reviewed publications on this matter are largely lacking; therefore, it is challenging to compare our study results with previous work done in PRRS or any other swine infectious disease.

This study has some limitations. First, it is important to highlight that our source population corresponded to sow sites only and did not include growing pig sites. Even though growing pig sites are responsible for adding "infection pressure" at a regional level, one could argue that this population is somewhat distinct from the sow farm population in terms of disease management. Infection of sow herds results in more dramatic consequences due to the fact that pigs produced in such facilities are commonly transported to other sites, and therefore decisions in regards to disease prevention and control are markedly different between these distinct animal populations. On a similar note, our analysis included data from voluntary participants only. Therefore, results do not necessarily apply to the overall population of swine sites in the US. The impact of this issue is hard to predict and assess, given that the representativeness of participating producers is not well documented; thus, the authors recommend results to be taken with caution.

Second, underreporting could have affected results, especially for systems and regions that have underlying immunity for PRRS. To the knowledge of the authors, there are no published methods for TD-R calculations that can account for such issue, but the authors believe the underreporting to be constant in time, thus not dramatically affecting results. Another methodology-related issue is the fact that in our case the epidemic was not observed from the first case onward; with results in overestimation of the initial reproductive numbers (9). For this reason, the authors decided not to consider initial estimated TD-R numbers (first 4 months) when summarizing these data. Finally, although the prevalence of intervals with 0 values was not high, padding the time series, by replacing the 0 values with 1's (a common practice in the time series analysis) resulted in a well-fitted distribution for the generation time, which subsequently increased the computation efficiency of the TD-R values. That said, this practice might have resulted in over or under estimation of the TD-R values in small regions with underreported outbreaks. However, as described above, the disease is endemic, and the assumption of at least one outbreak occurred in the small production region is biologically plausible in the context of epidemiology of PRRSV in the US. The authors also recognize that the CIs estimated using the time-dependent method could have been wide because few cases were observed at times. However, the authors are unaware of a method available (at time of publication) that would allow for better estimation of a transmission parameter in endemic settings.

In conclusion, this study showed the utility that TD-R estimates may have in monitoring and early signaling epidemics for PRRS, and its benefits will likely vary according to geographical region and production system. The TD-R is a promising complementary measure for incidence, because the latter is limited to measuring the amount of cases per unit of time but does not provide insights on the epidemic progression or effectiveness of control measures, which can be accomplished with the calculation of the former. The use of the TD-R may be complemented by other tools, such as, for example, the use of sequential Bayesian

R_0 for prediction of increases in the incidence as well as signaling the end of epidemics (18).

AUTHOR CONTRIBUTIONS

AA and MA formulated the main hypothesis of this study; AA was responsible for report and manuscript preparation and MA substantially helped with analysis. RM was responsible for acquisition of data. KV, AP, and RM helped with interpretation of results. All the authors contributed in critically revising the manuscript and approving its final version.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Swine Health Monitoring Project participants for sharing information.

FUNDING

Funding for the SHMP and this study was provided by the Swine Health Information Center, the National Pork Board, and the MnDrive program.

REFERENCES

1. Alvarez J, Valdes-Donoso P, Tousignant S, Alkhamis M, Morrison R, Perez A. Novel analytic tools for the study of porcine reproductive and respiratory syndrome virus (PRRSv) in endemic settings: lessons learned in the US. *Porc Health Manag* (2016) 2:3. doi:10.1186/s40813-016-0019-0
2. Ropp SL, Wees CE, Fang Y, Nelson EA, Rossow KD, Bien M, et al. Characterization of emerging European-like porcine reproductive and respiratory syndrome virus isolates in the United States. *J Virol* (2004) 78:3684–703. doi:10.1128/JVI.78.7.3684-3703.2004
3. Li C, Zhuang J, Wang J, Han L, Sun Z, Xiao Y, et al. Outbreak investigation of NADC30-like PRRSV in South-East China. *Transbound Emerg Dis* (2016) 63(5):474–9. doi:10.1111/tbed.12530
4. Tousignant SJ, Perez A, Morrison R. Comparison between the 2013-2014 and 2009-2012 annual porcine reproductive and respiratory syndrome virus epidemics in a cohort of sow herds in the United States. *Can Vet J* (2015) 56:1087–9.
5. Tousignant SJ, Perez AM, Lowe JF, Yeske PE, Morrison RB. Temporal and spatial dynamics of porcine reproductive and respiratory syndrome virus infection in the United States. *Am J Vet Res* (2015) 76:70–6. doi:10.2460/ajvr.76.1.70
6. Valdes-Donoso P, Jarvis LS, Wright D, Alvarez J, Perez AM. Measuring progress on the control of porcine reproductive and respiratory syndrome (PRRS) at a regional level: the Minnesota N212 regional control project (RCP) as a working example. *PLoS One* (2016) 11:e0149498. doi:10.1371/journal.pone.0149498
7. Dietz K. The estimation of the basic reproductive number for infectious diseases. *Stat Methods Med Res* (1993) 2:23–41. doi:10.1177/096228029300200103
8. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am J Epidemiol* (2004) 160:509–16. doi:10.1093/aje/kwh255
9. Obadia T, Haneef R, Boelle PY. The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Med Inform Decis Mak* (2012) 12:147. doi:10.1186/1472-6947-12-147
10. Alkhamis MA, VanderWaal K. Spatial and temporal epidemiology of lumpy skin disease in the Middle East, 2012-2015. *Front Vet Sci* (2016) 3:19. doi:10.3389/fvets.2016.00019
11. Holtkamp DJ, Polson DD, Torremorell M, Morrison B, Classen DM, Henry S, et al. Terminology for classifying swine herds by porcine reproductive and respiratory syndrome virus status. *J Swine Health Prod* (2011) 19(1):44–56.
12. Dinno A. *Dunntest: Dunn's Test of Multiple Comparisons Using Rank Sums*. Stata Software Package (2014). Available from: http://www.alexisdinno.com/stata/dunntest.html
13. Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F. A space-time permutation scan statistic for disease outbreak detection. *PLoS Med* (2005) 2(3):e59. doi:10.1371/journal.pmed.0020059
14. Kulldorff M. *SaTScanTM v9.4: Software for the Spatial and Space-Time Scan Statistics*. (2009). Available from: http://www.satscan.org/
15. Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. *J R Statist Soc A* (2001) 164(1):61–72. doi:10.1111/1467-985X.00186
16. Pfeiffer D, Robinson T, Stevenson M, Stevens K, Rogers D, Clements A. *Spatial Analysis in Epidemiology*. New York, NY: Oxford University Press (2008).
17. Bumgardner EA, Lawrence PK. Complete genome sequencing of recently isolated Porcine Reproductive and Respiratory Syndrome Virus RFLP 1-7-4 strains reveals a familiar series of deletions in ORF1a that are similar to those seen in previously studied strains of high virulence. *Abstract Retrieved from 2015 North American PRRS Symposium*. (2015). 34 p. Available from: https://www.vet.k-state.edu/na-prrs/docs/2015-proceedings.pdf
18. Betencourt LMA, Ribeiro R. Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS One* (2008) 3(5):e2185. doi:10.1371/journal.pone.0002185

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Veterinary Science

ORIGINAL RESEARCH
published: 28 February 2017
doi: 10.3389/fvets.2017.00016

Data-Driven Risk Assessment from Small Scale Epidemics: Estimation and Model Choice for Spatio-Temporal Data with Application to a Classical Swine Fever Outbreak

Kokouvi Gamado[1], Glenn Marion[1] and Thibaud Porphyre[2,3]*

[1] Biomathematics and Statistics Scotland, Edinburgh, UK, [2] Epidemiology Research Group, Center for Immunity, Infection and Evolution, University of Edinburgh, Edinburgh, UK, [3] The Roslin Institute, University of Edinburgh, Easter Bush Campus, Edinburgh, UK

OPEN ACCESS

Edited by:
Kimberly VanderWaal, University of Minnesota, USA

Reviewed by:
Francisco Ruiz-Fons, Spanish Research Council, Spain
Gert Jan Boender, Wageningen University and Research Centre, Netherlands

***Correspondence:**
Glenn Marion
glenn.marion@bioss.ac.uk

Specialty section:
This article was submitted to Veterinary Epidemiology and Economics, a section of the journal Frontiers in Veterinary Science

Received: *01 November 2016*
Accepted: *30 January 2017*
Published: *28 February 2017*

Citation:
Gamado K, Marion G and Porphyre T (2017) Data-Driven Risk Assessment from Small Scale Epidemics: Estimation and Model Choice for Spatio-Temporal Data with Application to a Classical Swine Fever Outbreak. Front. Vet. Sci. 4:16. doi: 10.3389/fvets.2017.00016

Livestock epidemics have the potential to give rise to significant economic, welfare, and social costs. Incursions of emerging and re-emerging pathogens may lead to small and repeated outbreaks. Analysis of the resulting data is statistically challenging but can inform disease preparedness reducing potential future losses. We present a framework for spatial risk assessment of disease incursions based on data from small localized historic outbreaks. We focus on between-farm spread of livestock pathogens and illustrate our methods by application to data on the small outbreak of Classical Swine Fever (CSF) that occurred in 2000 in East Anglia, UK. We apply models based on continuous time semi-Markov processes, using data-augmentation Markov Chain Monte Carlo techniques within a Bayesian framework to infer disease dynamics and detection from incompletely observed outbreaks. The spatial transmission kernel describing pathogen spread between farms, and the distribution of times between infection and detection, is estimated alongside unobserved exposure times. Our results demonstrate inference is reliable even for relatively small outbreaks when the data-generating model is known. However, associated risk assessments depend strongly on the form of the fitted transmission kernel. Therefore, for real applications, methods are needed to select the most appropriate model in light of the data. We assess standard Deviance Information Criteria (DIC) model selection tools and recently introduced latent residual methods of model assessment, in selecting the functional form of the spatial transmission kernel. These methods are applied to the CSF data, and tested in simulated scenarios which represent field data, but assume the data generation mechanism is known. Analysis of simulated scenarios shows that latent residual methods enable reliable selection of the transmission kernel even for small outbreaks whereas the DIC is less reliable. Moreover, compared with DIC, model choice based on latent residual assessment correlated better with predicted risk.

Keywords: spatial epidemics, kernel transmission functions, Markov Chain Monte Carlo, risk assessment, latent residuals, deviance information criterion

1. INTRODUCTION

The livestock epidemics of foot-and-mouth disease (FMD) in the United Kingdom (UK) in 2001 (1) and of classical swine fever (CSF) in the Netherlands in 1997 (2, 3) were characterized by their widespread spatial extent and significant impact on the agricultural sector. The 2001 FMD outbreak affected all UK livestock farms and is estimated to have a cost total of £8 billion (4) whereas the CSF epidemic in Netherlands totalized a short-term economic impact over £1.1 billion (5).

The desire to mitigate such impacts for future livestock-disease incursions has increased focus on preparedness for emerging and re-emerging pathogens (6, 7) and highlighted the need for quantitative tools to support such efforts (8, 9).

A key step in controlling livestock-disease incursions is to quantitatively assess the risk of localized disease spread from infected to susceptible farms. Control strategies can then make use of such risk assessments, for example they can be used to decide on which farms surrounding confirmed infected premises to impose control measures (10, 11). Quantitative study of historical epidemics has the potential to make such risk assessment faster and more robust, enabling more rapid and reliable response than possible if waiting for sufficient data to accrue during an ongoing epidemic (12). Tools to enable analysis of historic outbreaks thus may provide information critical to control operations prompted by future incursions of emerging or re-emerging diseases. Data on large outbreaks such as those described above have been shown to enable quantification of various epidemiological (13), economic (14, 15), and logistical (16) aspects of disease spread. In turn, these studies have contributed to the design of novel, more cost-efficient, control strategies against future epidemics (17–20). Fortunately large outbreaks, such as those described earlier, are rare and incursions of emerging or re-emerging disease typically lead to small and often repeated localized outbreaks (21). In this paper, we will investigate the quality of inference possible from data on small outbreaks.

Model-based inference can be conceptualized as a two-stage process; estimate the parameters for each of a set of given models and then rank or choose between the different models. Subsequent risk assessment can then be based on the model that best fits the data, or suitably weighted outputs from a set of models.

Over recent years, a number of authors have developed Bayesian inferential tools to enable statistically rigorous parameterization of discrete state continuous time Markov and semi-Markov processes (DCTMPs) from noisy and incomplete observations typical of field data, e.g., disease detections from real world epidemics (22–27). These inference tools are flexible in that they can be applied to a wide range of model structures and epidemic scenarios. Discrete state continuous time Markov and semi-Markov processes are well suited to modeling a diverse range of epidemiological systems. Their continuous time nature enables more accurate representation of the transmission process than discrete time models, and their discrete state space is ideally suited to represent epidemic spread between individuals, e.g., farms classified into distinct disease classes, susceptible, infected, etc. The inference tools for such models make use of data-augmentation approaches to account for missing information (such as the times individuals become infected) which are treated as additional parameters to be inferred. Information (including uncertainty) on key quantities obtained from Bayesian inference is encoded in the so-called posterior, which is the joint distribution of model parameters (describing processes such as transmission between hosts and disease progression within an infected host) and missing data, e.g., the infection and transition times conditional on the observed data and modeling assumptions. Markov Chain Monte Carlo (MCMC) methods are typically used to draw samples from the posterior, and from these samples it is straightforward to calculate quantities of interest such as the probability, under some defined scenario, that a given farm will become infected at some future date. It is important to note that such predictions reflect the uncertainty inherent in the inference and therefore indirectly the quality of the data. It can be shown that reliability of inference for a given process (e.g., transmission) increases with the number of associated events (e.g., infections) occurring during the observation period. In general, this supports the intuition that uncertainty in inference will depend on the quantity of data available, and that uncertainty in inferences based on small outbreaks will be greater, and the reliability of the estimates more limited, than for large outbreaks.

Although theory underlying model choice in a Bayesian framework is well established (28, 29), its implementation is often impractical, especially for missing data problems where, as of interest here, latent variables are used for data augmentation. Reversible jump MCMC (30) in principle allows calculation of the so-called Bayes factor comparing two models but suffers implementation issues (31). For example, to compare a newly proposed model with earlier models requires the rerunning (and often recoding) of the inference procedure for at least one earlier model. In practice, the only model selection tool in widespread use for DCTMPs applied to epidemiological modeling is the DIC, or Deviance Information Criteria (32–34). DIC is a Bayesian model selection method which tries to balance model complexity with fit to data (35). However, there are increasing concerns with regards to discriminatory performance (36) particularly in the presence of latent variables, such as infection times, where there is no unique definition of DIC (37). A complimentary approach to model choice is that of model assessment where statistics are constructed to detect inconsistencies between model assumptions and the real world processes being modeled (38, 39). Novel diagnostics tools for assessing the fit of DCTMP model, known as latent residuals, have recently been developed by Lau et al. (40) and have proven to be efficient in identifying misspecification of model components in the context of spatio-temporal models for disease spread. This approach allows indirect model comparison, but also has the important benefit of indicating in what manner a given model may be inadequate, thereby providing insights into ways for improvements. The application of latent residuals for DCTMPs to small outbreaks has yet to be tested (40), and model comparison based on data from small outbreaks is likely to be challenging given the potential difficulties described above associated with inference for individual models from such data.

In this paper, we focus on the spatio-temporal modeling of disease spread between farms and subsequent detection of infected premises, conducting inference from data on detection

times only. We focus on the model choice problem of selecting the appropriate functional form describing between-farm disease transmission. Local between-farm transmission is generally mediated by multiple processes influenced by a wide range of factors including human behavior and characteristics of the livestock-disease system in question. However, typically these factors are not quantified and overall local spatial spread is often summarized to be a function of the straight-line distance between farms, modeled by the so-called "kernel transmission function." In the context of CSF, Mintiens et al. (41) examined risk factors associated with the occurrence of neighborhood infections during the 1994 Belgium outbreak and found that intensity of neighboring herds was the most significant factor. Staubach et al. (42) established a distance-dependent risk function based on data obtained from real outbreaks. Building on this earlier work, Backer et al. (20) modeled between-farm spread of CSF using the kernel transmission function (described later as K_2, see Section 2.1) but modulated the kernel by the infectiousness of the infected farms (infectiousness defined by the number of infectious individuals on farm). The same transmission kernel was considered by Boender et al. (43), but they also included the influence of farm size in order to determine both the distance dependence and the farm-size dependence of the between-farm transmission risk of CSF during the Dutch 1997/1998 epidemic. As we subsequently show, identifying the correct kernel is critical in effective support of policy and decision-making for disease control since it can affect the risk profile of farms that are candidates for control.

In summary, this manuscript is organized as follows:

1. We first focus on the methodology describing the set of models that are mainly different in the kernel transmission functions used. These models' parameters are inferred in the Bayesian framework and two model selection tools are applied, namely DIC and latent residuals.
2. Having inferred parameters from the various models, we quantify the posterior risks associated with each premise by simulating repeated epidemics using posterior estimates as initial conditions.
3. The reliability of the methods are shown using simulated data and applied to field data on the small CSF outbreak that occurred in 2000 in East Anglia, UK (44).

2. MATERIALS AND METHODS

2.1. Model Structure

We use a stochastic spatio-temporal Susceptible-Infectious-Removed (SIR) epidemic model to assess local between-farm spread and farm-level detection and control of a disease. The underlying model framework is a continuous time discrete state-space semi-Markov process. In our model, individuals represent spatially distinct locations, e.g., farms, characterized by a discrete set of disease states (S, I, and R). The population is assumed initially fully susceptible prior to an initial incursion of disease into a single farm (the index case). Susceptible farms (in state S) subsequently become infected via local contacts with infected farms (in state I). Once infected, each farm remains in state I until its infectious status is detected and controls imposed. It is assumed that controls are introduced immediately upon detection and are completely effective. Therefore, once detected farms enter a restricted or controlled state (state R) in which they are removed from (i.e., play no further role in) the epidemic.

2.1.1. Detection and Control

The description of the R state assumes that effective control is rapid following detection which is a reasonable assumption for a range of disease systems (9, 45). In the context of FMD or CSF outbreaks in the UK, total depopulation of the infected premises is carried out within 24 h of the reporting of on-farm detection (1, 44). The infectious period for an infected farm can therefore be reasonably approximate as the period between the date of infection and the date at which the farm in question is reported as infected. The detection of disease is considered non-Markovian since the probability of detection is not constant but varies as a function of time since infection, e.g., as the number of on-farm cases changes. Here, we assume that the infectious period of the disease follows a gamma distribution with shape α and rate γ.

2.1.2. Secondary Infection

In the situation where farm i is infected and infectious (i.e., in state I) and farm j is susceptible (i.e., in state S), the rate at which j enters state I (i.e., becomes infected/infective) due to transmission from farm i, is $\beta_{ij} = \beta_0 h_{ij}$. Here, β_0 is the maximum rate of contact whereas h_{ij} is a kernel function that can involve individual-specific characteristics. In the case of a distance kernel function, the contact rate varies as a function of the Euclidean (straight line) distance between farms. The shape of the kernel can significantly impact the spatial spread and development of epidemic processes (46, 47). In order to evaluate the effect of the kernel selection on outbreak risk assessment, we considered four different forms for the kernel function h_{ij}:

$$K_1 = \exp\{-\tau\rho(i,j)\},$$
$$K_2 = \frac{1}{1+\left(\frac{\rho(i,j)}{d}\right)^{\tau}},$$
$$K_3 = \frac{1}{1+\frac{\rho(i,j)}{d}},$$
$$K_4 = 1-\exp\left(-\left(\frac{\rho(i,j)}{d}\right)^{-\tau}\right),$$

where $\rho(i,j)$ denotes the Euclidean distance between individuals or sites i and j ($i,j \in \{1,2,\ldots N\}$). These functions represent a broad range of kernel functions found in the literature to characterize local spread in the analysis of spatial epidemics. For example, the exponential kernel K_1 was used for modeling the spread of the *citrus tristera* virus (48) and the spread of citrus greening (49). The kernel K_2 was used to model the between-farm spread of the CSF in the Netherlands by Backer et al. (20) and Boender et al. (43). The kernels K_2 and K_4 were used to model the spread of FMD in Japan (50) and in Netherlands (51) and are parameterized in such a way they capture short-range spread. The Cauchy kernel K_3 is a special case of K_2 for $\tau = 1$ and was considered by Kypraios (52), and as a description of the dispersal of an invasive

non-native vascular plant by Lau et al. (40). K_3 allows more for long-range infections compared to the others as it presents the heaviest tail among all four. Full model specification as given in Section S1.1 in Supplementary Material.

2.2. Bayesian Inference

Here, we assume that available data consist of the time at which individual premises were detected as infected, and the geolocations not only of these premises, but all susceptible farms. In this analysis, we assume the times at which individuals are infected are not observed, since this information is typically not available in outbreak situations. However, such information could be incorporated in our framework if it were available.

Missing infection times make the likelihood for the SIR model intractable and render the use of frequentist approaches for inference less than straightforward. However, developments in data-augmentation MCMC allow Bayesian inference of both the parameters and missing infection times. The approach adopted is similar to that of Jewell et al. (25) which builds on the seminal work of Gibson and Renshaw (22) and O'Neill and Roberts (23). Full details on the model fitting framework and procedures are provided in Sections S1.2 and S1.3 in Supplementary Material.

Briefly, we considered the joint posterior distribution of the model parameters $\boldsymbol{\theta}$ and infection times $\boldsymbol{I}$ given the data, i.e., the detection times $\boldsymbol{R}$. Using Bayes' theorem, this can be expressed as

$$\pi(\boldsymbol{\theta}, \boldsymbol{I} \mid \boldsymbol{R}) \propto \pi(\boldsymbol{R}, \boldsymbol{I} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{1}$$

where $\pi(\boldsymbol{R}, \boldsymbol{I}|\boldsymbol{\theta})$ is nothing more than the probability of obtaining a given joint realization of detection and infection times $(\boldsymbol{I}, \boldsymbol{R})$ from a simulation of the model given parameter values $\boldsymbol{\theta}$. Sometimes referred to as the complete data likelihood, this can be calculated based only on the definition of the model. $\pi(\boldsymbol{\theta})$ is the joint prior distribution of the parameters and is specified in light of any information available before the data $\boldsymbol{R}$ is observed. If samples $\{(\boldsymbol{\theta}_k, \boldsymbol{I}_k)\colon k = 1 \ldots L\}$ can be drawn from the posterior distribution expressed in equation (1), then it is possible to calculate almost any statistic of interest. For example, marginalizing over $\boldsymbol{I}$ would result in estimating the distribution $\boldsymbol{\theta}|\boldsymbol{R}$, whereas marginalizing over $\boldsymbol{\theta}$ would result in estimating $\boldsymbol{I}|\boldsymbol{R}$. We note that the latent space, here corresponding to the set of infection times only, can be readily expanded to consider information such as the transmission tree, i.e., "who-infects-who." Moreover, the framework can also be expanded to consider other disease categories such as an exposed state.

Here, we employ Markov Chain Monte Carlo (MCMC) methods to draw samples from the posterior distribution specified in equation (1) only up to constant of proportionality. More specifically, we make use of a Metropolis–Hastings (53) within Gibbs algorithm (54) together with simple non-centering techniques (27). Full details of the sampling algorithms are provided in Section S1.3 in Supplementary Material.

2.3. Model Selection

Model selection is a challenging aspect of inference, especially for models with latent variables and where data are limited. In this study, we employ two contrasting model selection tools from the literature to select among our four spatial kernels, $K_1, \ldots, K_4$, which model fits best our data. In particular, we use the deviance information criterion (DIC (35)) and Bayesian latent residuals (40).

Although widely used and measuring the trade-off between the model fit and complexity, DIC is recognized to have issues such as non-invariance to reparameterization, lack of consistency, no basis on a proper predictive criterion (36), and multiple definitions in the presence of latent variables (37). In this study, we used two formulations of DIC for latent variable models (37): DIC_1 is computed accounting for the full likelihood, i.e., the data augmented likelihood, whereas DIC_2 is computed on a partial likelihood of the observation process conditional on the latent variables. We note that DIC_1 and DIC_2 correspond respectively to DIC_4 and DIC_8 in Celeux et al. (37). In either case, interpretation of DIC is straightforward with the smallest value taken to indicate which model performs the best. In practice, however, only differences of magnitude 10 or more between models are usually considered as indicative of significant differences in model fit (55).

By contrast, Bayesian latent residuals based on non-centered reparameterisations of discrete state-space continuous time semi-Markov process have recently been proposed as an approach to assess the fit of different components of the model, thereby focusing attention on aspects of the model that need improvement (40). For example, the infectious link residuals (ILRs) considered here are specifically designed to detect misspecification of the spatial transmission kernel. These are constructed by expanding the set of latent variables sampled from the posterior distribution to include both infection times and also the donors of infection, i.e., the transmission tree. This information then allows inference of a random variable, an ILR, that, under the assumption the model is correct has a uniform $U(0, 1)$ distribution (40). Evidence of non-uniform distribution of the ILRs indicates a misspecification of the kernel function. Recall that the data consist only of detection times and the infection times are unknown. Each set of infection times estimated within the MCMC algorithm (i.e., drawn from the posterior distribution) corresponds to a set of residuals. Each of these sets of residuals is subjected to an Anderson–Darling test to assess whether it conforms to a uniform $U(0, 1)$ distribution or not. Therefore, there are exactly the same numbers of p-values as there are samples from the posterior distribution. We record the proportion of p-values, e.g., $\Pr(p < 5\%)$, that are less than a confidence level, e.g., 5%. In the case that the proportion is similar to the confidence level, we take that as indicative that the model is not inconsistent with the data; more formally there is not sufficient evidence to reject the model. A high proportion of p-values greater than the 5% level (i.e., $\Pr(p < 5\%)$ large) is interpreted as evidence against the chosen form of the kernel.

To evaluate the performance of these model selection and assessment tools, we apply them within a simulation study framework where the true data-generating model is known. See Section 2.5 for full details of the simulated scenarios considered. For each iteration of the simulation study, we simulate an epidemic and generate a data set (detection times only) using the Gillespie algorithm (56) under a specified kernel. Each dataset is fitted to the models defined in Section 2.1 using MCMC, not only with

the kernel used to simulate the epidemic, but also with alternative kernels. Model selection and assessment tools based on DIC and latent residuals are used to assess each of the fitted models as described above. Full details of the model selection tools used are provided in Section S1.4 in Supplementary Material.

2.4. Risk Quantification

In order to make the results of inference more directly relevant to the development of disease control, we construct posterior predictive distributions that quantify farm-level risk. To do so, we start by drawing a large number of independent samples from the joint posterior distribution for each model variant (i.e., each kernel). For each kernel, and a fixed set of initial conditions, these sampled parameter values are used to simulate multiple realizations of an epidemic using the Gillespie algorithm from which we record the infected sites and event times. From the large number of simulated epidemic realizations obtained, distributions are calculated for each kernel by evaluating the proportion of realizations in which each premise becomes infected at any given point in time. Such posterior predictive distributions can be used to produce heat maps showing the spatial distribution of farm-level infection risk (see further in section: Results) at a given point in time. The heat maps are produced at a length of times sufficient to capture the early phase and small scale epidemics' behavior.

An alternative approach to visualizing this information is to summarize the simulated epidemics by looking at the average proportion of infected farms (farm-level prevalence) as function of time. However, interpreting such average between-farm prevalences is complicated by the fundamental characteristics of disease transmission in SIR epidemics. In particular, it is well known that when the basic reproduction rate $R_0 > 1$ disease spreads on average, but that the final size distribution for a stochastic SIR epidemic model is bimodal, showing that some epidemics die-out before becoming large, while others grow to affect a large proportion of the population (8, 57) (see Section S2.1.3 in Supplementary Material). Given that the mean of a bimodal distribution is not a meaningful summary statistic, we use the mean of the final size distribution as a boundary defining small and large epidemics. This enables us to calculate the probability of obtaining small and large outbreaks and plot average between-farm prevalence as a function of time for both small and large outbreaks, as predicted under each of the four kernels.

2.5. Simulation Studies

To assess the reliability of parameter inference and model selection and to evaluate the effect of final epidemic size, we generate data sets based on the Gillespie algorithm (56), introducing the disease via a single randomly selected primary (index) case located in a closed population of $N = 201$ farms in a square of sides [0, 2,000] km (corresponding to an average density of 5×10^{-5} farms per 10^2 km^2) and using a fixed set of parameters for each scenario considered. The simulation studies described below are divided into simulation study 1 which focuses on single data sets generated using two contrasting kernels, and study 2 which explores coverage properties of our inference procedure using data sets generated from multiple realizations of each of these scenarios. Bayesian inference is applied to every simulated data set to fit models based on all four kernels (K_1, K_2, K_3, K_4), and the model selection procedures described above are applied to every data set to discriminate between the kernels.

2.5.1. Simulation Study 1a

We first simulate a single realization of the epidemic assuming that the mechanism of disease spread is described by kernel K_1. More specifically the characteristics of underlying infection process are $\beta_0 = 0.35$, $K_1 = \exp\{-0.008\rho(i,j)\}$ with an initial condition in which all farms are susceptible except for a single, randomly selected primary case. The infectious period follows a Ga(5, 5) distribution. A total of $n_R = 43$ removed individuals were recorded together with their removal times.

2.5.2. Simulation Study 1b

A single data set of detection times is generated from a simulation using kernel K_2 instead of K_1. The detection time distribution and initial conditions are as in study 1a but here, $\beta_0 = 400$, $K_2 = \frac{1}{1+\left(\frac{\rho(i,j)}{1.5}\right)^2}$, and $n_R = 44$ removal times were obtained.

2.5.3. Simulation Study 2a

Different outbreak sizes are considered in order to assess goodness-of-fit and evaluate the performance of model selection tools as epidemic size increases. Simulations are performed to obtain $n = 30$ realizations for each epidemic size category of [6, 10], [11, 15], [16, 20], [21, 25], [26, 30], [31, 35], [36, 40], and [41, 45]. For each realization, a data set of detection times was recorded. For each randomly selected incursion event, the spread of the disease was simulated using kernel transmission function K_1 with the same parameterization used in simulation study 1a. Considering only small ($\leq$45 infected farms) completed epidemics, we subsequently inferred all parameters of interests (as well as posterior distributions of infection times for infected premises), and computed all three measures of goodness-of-fit (DIC_1, DIC_2, ILR) under the hypothesis that the spread of the disease follows a kernel transmission function with a shape either K_1, K_2, K_3, or K_4. For each scenario, coverage properties (i.e., the number of times the true parameter values fall within their respective 95% credible intervals) are recorded for each size category described above.

2.5.4. Simulation Study 2b

To evaluate the resilience of our conclusions given to the shape of the kernel transmission function of the data generation process, an analogous procedure to that described for study 2a was carried out considering data generated using K_2 (i.e., instead of the kernel transmission function K_1) with the same parameterization as used in simulation study 1b.

2.6. Field Data

In 2000, the UK experienced an outbreak of CSF across the region of East Anglia (44). Unlike in the Netherlands where the CSF outbreak has been detected in various areas of the country (58), the UK outbreak was only detected in the region of East Anglia, with n = 16 farms found infected in the 3 month-long outbreak. Records of the pig population at the time show that N = 1,703 pig farms were present in the affected area and were considered at risk

(44). This represents an approximate average population density of 6.08 per 10^2 km^2. Although all pig farms in the UK could be considered at risk, inferred transmission distances are small relative to the area occupied by the above subpopulation, and all infected premises are contained comfortably in the defined region and the risks for farms outside the subpopulation considered are considered negligible given their distances to the infected farms and the inferred transmission kernels (see Results section), unless sources of infection exist other than those considered in the modeling framework used here. The inferences presented below would therefore be essentially unchanged by accounting for a larger at risk population.

All parameters of interest including infection times were inferred using data describing farm locations and the time at which each premises was detected and reported infected. Control interventions consisted of depopulation of infected premises within 24 h of reporting of disease detection. Inference was conducted, and all model selection criteria and measures of goodness-of-fit computed, under the hypothesis that spread of the disease was described by the spatial SIR model described above with each of the kernel transmission functions K_1, K_2, K_3, or K_4. Although CSF is highly contagious and may result in the death of young animals, clinical sign are non-specific and can result in failed diagnosis and a long period of undiagnosed spread (59). Consistent with the literature (20), we therefore assumed that a minimum of 8 days was required for infected premises to be detected. Therefore, during the fitting procedure, the infectious period was assumed to follow a left-truncated gamma distribution. Full details are available in Section S1.1 in Supplementary Material.

3. RESULTS

3.1. Testing Inference Methods with Simulated Data

We first make use of simulated data to assess the performance of inference and model selection procedures and risk assessments based on them. Initially the focus is on inference from individual epidemics, but latterly we explore coverage properties, as a function of outbreak size, by considering performance on data from a representative ensemble of epidemics.

3.1.1. Inference from Single Epidemics

Posterior means for model K_1 parameters, under simulation study 1a, were 0.396 (95% credible interval 0.169, 0.761), 6.251 (2.638, 11.631), 0.00771 (0.00540, 0.01031), and 4.907 (1.986, 9.190), respectively, for β_0, γ, τ, and α. All 95% credible intervals overlap with the underlying parameters values used to generate the data, suggesting our inference framework is able to recover appropriate parameterizations when there is no mismatch between the data-generating mechanism and fitted model. This is also the case for simulation study 1b which replaces kernel K_1 with K_2 and inference for β_0, γ, τ, α, and d yields estimates 714.576 (59.319, 2,809.770), 7.931 (3.298, 13.429), 2.016 (1.639, 2.411), 5.567 (2.228, 9.782), and 2.237 (0.699, 4.717), respectively. Assessments of convergence of the outputs of the MCMC sampler reveal no evidence of lack of convergence as shown by the auto-correlation functions in Section S2.1.1 in Supplementary Material.

Table 1 shows the values obtained for each of the model selection tools implemented when assessing the fit of model variants based on each of the four transmission kernels. The correct kernel was identified when considering the ILRs making use of the measure Pr($p < 5\%$), the proportion of p-values less than 5% based on the latent residuals, and using DIC_1. It is also worth noting that analysis of the residuals suggests an equally good fit for K_2 and K_4 in simulation study 1b. By contrast, DIC_2 led to selection of the incorrect kernel in both situations, preferring K_3 rather than either K_1 or K_2. This finding is consistent with previous studies (40) and highlights that using DIC_2 may generate selection bias.

To evaluate the implication of such miss-selection on predicted risk associated with different premises, we construct posterior predictive probabilities of infection under all kernels (corresponding risk maps are shown in Section S2.1.4 in Supplementary Material). Comparing across kernels there are clear differences in risk levels predicted for many locations. K_3 predicts highest risk of infection for all farms while predicted risk under K_2 and K_4 decreases for extreme locations, whereas the predicted risk under K_1 falls off much faster with distance from the index case.

Figure 1 shows separately for both small and large outbreaks, how the average proportion of infections, i.e., the total number of cases divided the total number of farms within the area, evolves through time under different kernels. Small (large) outbreaks are defined as those less (greater) than the posterior mean epidemic size (see methods). The between kernel differences in the posterior predicted average proportion of infected premises through time, are very noticeable for small epidemics, and still visible, but to a lesser extent, for large epidemics.

We also evaluated the posterior predicted risk of a small as opposed to a large outbreak. Under simulation study 1a small epidemics occur 61.87% of times for K_1, 58.59% for K_2, and 52.88% and 59.12% for K_3 and K_4, respectively. Similar proportions are obtained under simulation study 1b except for small difference in K_3 with 57.18%, while other proportions are 60.39%,

TABLE 1 | Computed DIC_1 and DIC_2 values and the proportion of p-values less than 5% obtained from testing the distribution of ILRs when fitting model variants K_1, …, K_4 to simulated data from scenarios 1a (*a*) and 1b (*b*).

(*a*) Simulation study 1a: true kernel, K_1

	DIC_1	DIC_2	Pr($p < 5\%$)
K_1	**266**	761	**6.42%**
K_2	274	23,741	36.88%
K_3	285	**630**	90.99%
K_4	273	39,486	39.29%

(*b*) Simulation study 1b: true kernel, K_2

	DIC_1	DIC_2	Pr($p < 5\%$)
K_1	239	371	34.79%
K_2	**227**	494	**4.43%**
K_3	252	**362**	89.61%
K_4	394	489	4.64%

Bold font indicates the smallest values for the DICs and Pr(p < 5%) bringing out the selected model by each tool.

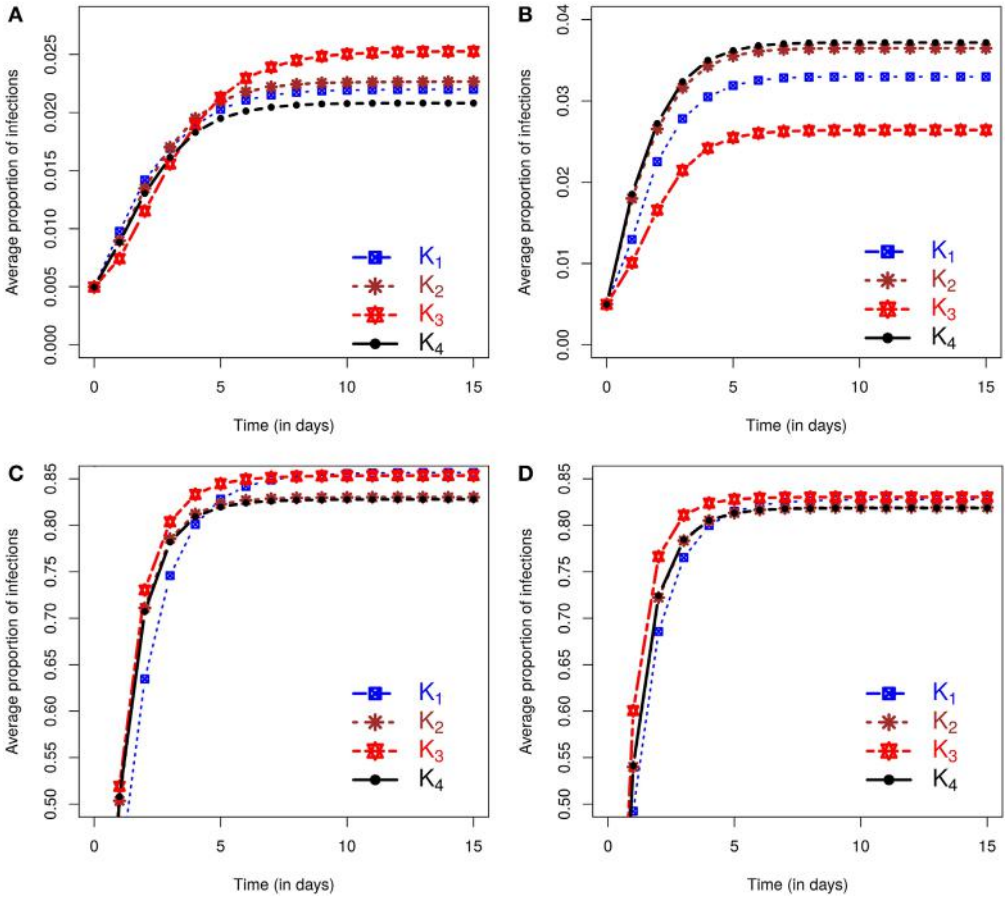


FIGURE 1 | **Posterior predicted average proportion of premises infected as the epidemics evolve in time (days)**. On each graph, the lines correspond to the results obtained when kernels $K_1 - K_4$ are fitted to data. The column on the left shows results using data from simulation study 1a and is divided into predicted outbreaks that are **(A)** small or **(C)** large. The column on the right shows results from simulation study 1b, stratified for **(B)** small or **(D)** large predicted outbreaks. The size of outbreaks was classified as either small or large based on final outbreak sizes being smaller or larger than the mean of the final size distribution (see text for details).

59.03%, and 58.83% for K_1, K_2, and K_4, respectively. It is also worth noting that in both simulation studies 1a and 1b, the observed epidemics on which inference is based are classed as small, but these lie at the upper end of the final size distribution for small epidemics according to the outbreak size classification used here. Inference from these single outbreaks therefore captures the true underlying bimodal nature of the outbreak risk associated with the SIR dynamic (see Section S2.1.3 in Supplementary Material). In terms of assessment of risks associated with future incursions this means that the probability of obtaining an outbreak similar in size to the observed outbreak is relatively low.

The posterior predicted estimates of future outbreaks produce rather different risk assessments depending on the kernel used to fit the data, suggesting that reliable model choice is of key importance.

In assessing the model selection criteria studied here several key points are noteworthy. First, the ability of DIC_2 to identify the correct model is poor, whereas DIC_1 and the ILRs do so in both simulation study 1a and 1b. Second, different kernels lead to different predicted risk. For example, the model variant with kernel K_3 produced the most extreme predictions of risk. However, this was also the model least favored when testing model assumptions using the ILRs, but was not consistently ranked by DIC_1. In simulation study 1b, analysis of ILRs leads to selection of two models, i.e., with kernels K_2 and K_4, and these models produced almost identical posterior predicted risk profiles in terms of expected proportion of infected farms. This similarity is not evident in their DIC_1 scores. These results suggest that model selection based on the ILRs may be better able to identify differences in posterior predicted risk profiles than DIC.

3.1.2. Coverage Properties and Effect of Outbreak Size

To generalize our findings, we considered epidemics generated with either a K_1 or K_2 kernel function and evaluated how inferences (description in section *Bayesian Inference*) and model selection tools (details given in section *Model selection*) considered may perform as a function of the epidemic size and across multiple realizations of the epidemic process (see sections *Simulation study 2a* and *Simulation study 2b*). Coverage properties of our

inference procedure from multiple realizations are explored and available in Section S2.2 in Supplementary Material. The coverage rates obtained show that the true parameters are contained approximately 95% of the time in their corresponding credible intervals. However, the rates are higher for the parameters of the infectious period distributions where informative priors are used on the shape parameter as in Kypraios (52) and Streftaris and Gibson (60). The uncertainty of the estimates reduces as the epidemic size increases.

For each epidemic size category, **Figure 2** plots the proportion of simulated data sets from which each of the model selection criteria successfully identifies the correct kernel. **Figure 2** shows that, although inferences of model structure are reasonably accurate for small epidemics, increasing epidemic size increases the accuracy with which the underlying infection process is identified. However, this depends on which measures of goodness-of-fit are used. While the DIC measures contradict each other, the latent residuals typically distinguish between the kernels and select the correct model used to simulate the outbreak (e.g., **Figure 2A**). However, for the smallest outbreaks, DIC_1 has the best record (e.g., **Figure 2B**), but as epidemic size increases, the reliability of the latent residuals quickly improves and outperforms both DIC_1 and DIC_2. For epidemic sizes greater than 20, ILRs identify the correct model in at least 90% of cases. In all situations considered, DIC_2 provided contradictory results to DIC_1, maintained a low success rate in choosing the true model, which for scenario 2b actually declined with outbreak size.

Figure 2B illustrates the resulting patterns from simulation study 2b which uses models based on K_2 to generate the data. This plot shows results for fitting only K_1–K_3, and K_4 is not considered since as we saw above K_2 and K_4 are difficult to discriminate, but lead to similar risk assessments. Including both here, therefore gives a biased view of the model selection process (see Section S3 in Supplementary Material). To better understand why these kernels are not efficiently discriminated, we explored the effect of the density of the population at risk. Results (details shown in Section S3 in Supplementary Material) show that increasing the density of farms provides more information about short-range transmission (i.e., more short-range transmissions occur) allowing differences between K_2 and K_4 to be distinguished. For realistic range of parameters, K_2 and K_4 can have similar long-range behavior or similar short-range behavior, but not both and therefore can agree when fitted to data which largely excludes short-range transmission, as is the case when the population density is low. These results highlight the critical importance of population density in the local spread of disease. They also explain the effect seen here where we are unable to discriminate between kernels K_2 and K_4 using simulated data from a low density population (5×10^{-5} per 10^2 km^2), but are able to detect differences between these kernels when fitting to field data below (see section *Classical Swine Fever epidemic*) where the density of farms is considerably higher (6.08 per 10^2 km^2).

3.2. Classical Swine Fever Epidemic

We now apply the inference and model selection methodologies described above to field data from a small scale outbreak of CSF in East Anglia, UK in 2000.

As described in section *Materials and Methods*, stochastic spatio-temporal models were fitted to the CSF outbreak field data (see section *Field data* for details) using each of the kernels K_1–K_4. This process yielded estimates of a number of quantities that are difficult to measure directly including the unobserved infection times, the infectious period distribution and the transmission kernel. In general, the inferred infection times agree well with the independent estimates obtained through contact tracing procedures during control activities as shown in Section S2.4 in Supplementary Material. The average infection times suggest that infections from first to last infected farm happen in an interval of approximately 65 days and around 9 farms were infected before the first detection.

As expected, the inferred transmission kernel functions all decay as a function of distance but vary according to the fitted

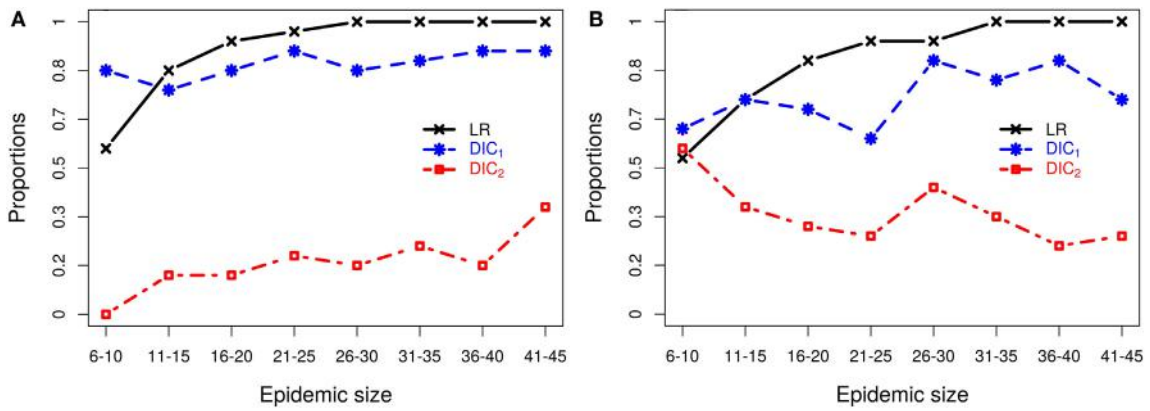


FIGURE 2 | Probabilities of correctly selecting the right model using latent residuals (LR), DIC_1, and DIC_2 in the case of (A) simulation study 2a (using K_1 to simulate the data) and (B) simulation study 2b (using K_2 to simulate the data). Both graphs show that the LR perform better (higher probabilities) in selecting the right kernels than the DICs as the epidemic size increases.

form. Posterior medians and their corresponding 95% credible intervals of the kernel transmission function are plotted on a log-scale under the four different kernels, with details on model goodness-of-fit and convergence in Section S2.4 in Supplementary Material. K_1 seems to decay fastest with the widest 95% credible interval. As before K_2 and K_4 look very similar in terms of median shape and credible interval and seem to allow for close range transmission, while K_3 decays very slowly and presents the smallest 95% credible interval. By contrast, choice of the transmission kernel has little impact on the inferred infectious period (Section S2.4 in Supplementary Material).

We now evaluate the impacts of each kernel on posterior predicted risks. **Figure 3** displays the mean posterior predicted proportion of infected farms as a function of time. Predictions of this risk measure under K_3 anticipate a larger number of infected premises, both in the case of small and large outbreaks, when compared with predictions based on the other kernels. The predictions under the remaining kernels are similar, and once again the differences between K_2 and K_4 are particularly small. As before, small (large) outbreaks are defined here as those where the final epidemic size is smaller (larger) than the mean of the final size distribution. The similarity between K_2 and K_4 and the relative difference of predictions under K_3 also hold for the inferred probabilities of a small outbreak, which are 65.45%, 66.17%, 63.75%, and 66.17%, respectively, under kernels K_1, K_2, K_3, and K_4. As with the simulation studies, the observed epidemic was classified as small but with the proportion of infections observed in the real epidemic greater than the average size predicted for small outbreaks (see plot in Section S2.5 in Supplementary Material).

Figure 4 further illustrates the importance of kernel choice by plotting risk maps of the probabilities of infection of each farm for a disease incursion. While farms close to the index case seem to present the highest risks of infection across all maps, the risk profiles of farms look different as we get further from the index case under each kernel. **Figure 4** shows that predicted risks under K_3 are the highest for farms situated at the largest distances from the index case because of its long right tail. At intermediate distances, the risks of farms under kernel K_1 are actually the greatest, but predicted risk is the lowest and decreases quickly for long distances compared to the other kernels. Kernels K_2 and K_4 predict visually similar risk with a consistent pattern of observed cases in terms of spatial extent and intensity of infection within the high risk area.

It is worth noting that most observed cases during the real CSF outbreak have a relatively high risk profile with an average probability of infection by $t = 90$ days of {0.52, 0.46, 0.46, 0.45} under K_1, K_2, K_3, and K_4, respectively, while the average risk across all farms is {0.41, 0.36, 0.40, 0.35}, i.e., a difference of {0.11, 0.10, 0.06, 0.10}. The different risk predictions under the four kernels are further quantified in **Table 2** which shows the number of farms with expected posterior probability of infection by $t = 90$ days less than 0.15, between 0.35 and 0.45, and greater than 0.45. Most farms showed probability of infection between 0.35 and 0.45 for K_3 while predictions under K_1 show more farms in this highest risk category. These figures also reveal relatively subtle differences in predicted risk under K_2 and K_4. The contrasting risk profiles explored above provide a further demonstration of the importance of kernel choice in predicting risk and therefore in terms of the design of disease control programs. Therefore, we now turn to our model selection methods.

As previously, we used DIC_1 and DIC_2 and the latent residual methodology to assess the suitability of the four kernel transmission functions in light of the available data. The latent residuals method gave a preference to K_2, followed in order of choice by K_4, K_1, and K_3, respectively, as shown in **Table 2**. From a purely model assessment perspective, there is some evidence against all the kernels using the latent residuals since $\Pr(p < 5\%)$ is greater than 5% in all cases, but this is hardly surprising given the relative simplicity of the models used. However, in terms of model selection, K_2 is preferred. DIC_1 agrees with the latent residuals method in the choice of K_2 as the preferred kernel but the assessments are not in the same order for the other kernels. By contrast, DIC_2 values do not show significant differences between kernels.

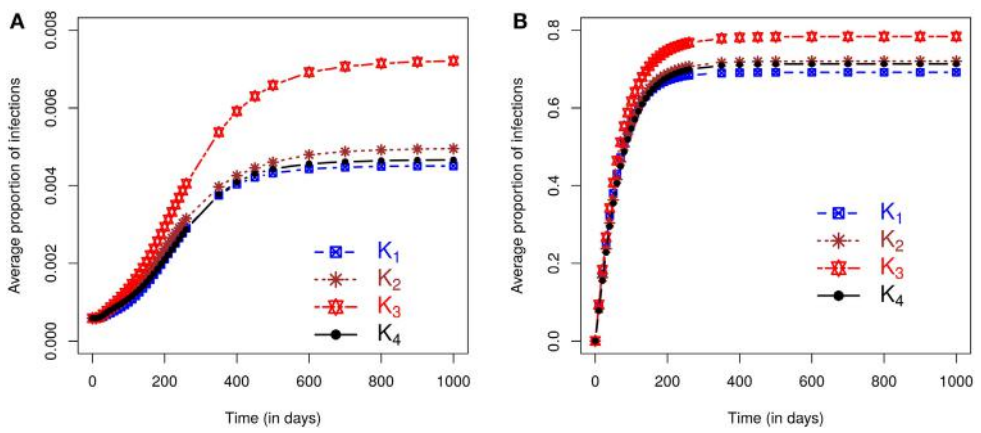


FIGURE 3 | Posterior predicted average proportion of premises infected plotted as a function of time. On each graph, the lines correspond to the results obtained when kernels $K_1 - K_4$ are fitted to the CSF data with final sizes **(A)** smaller or **(B)** larger than the mean of the final size distribution.

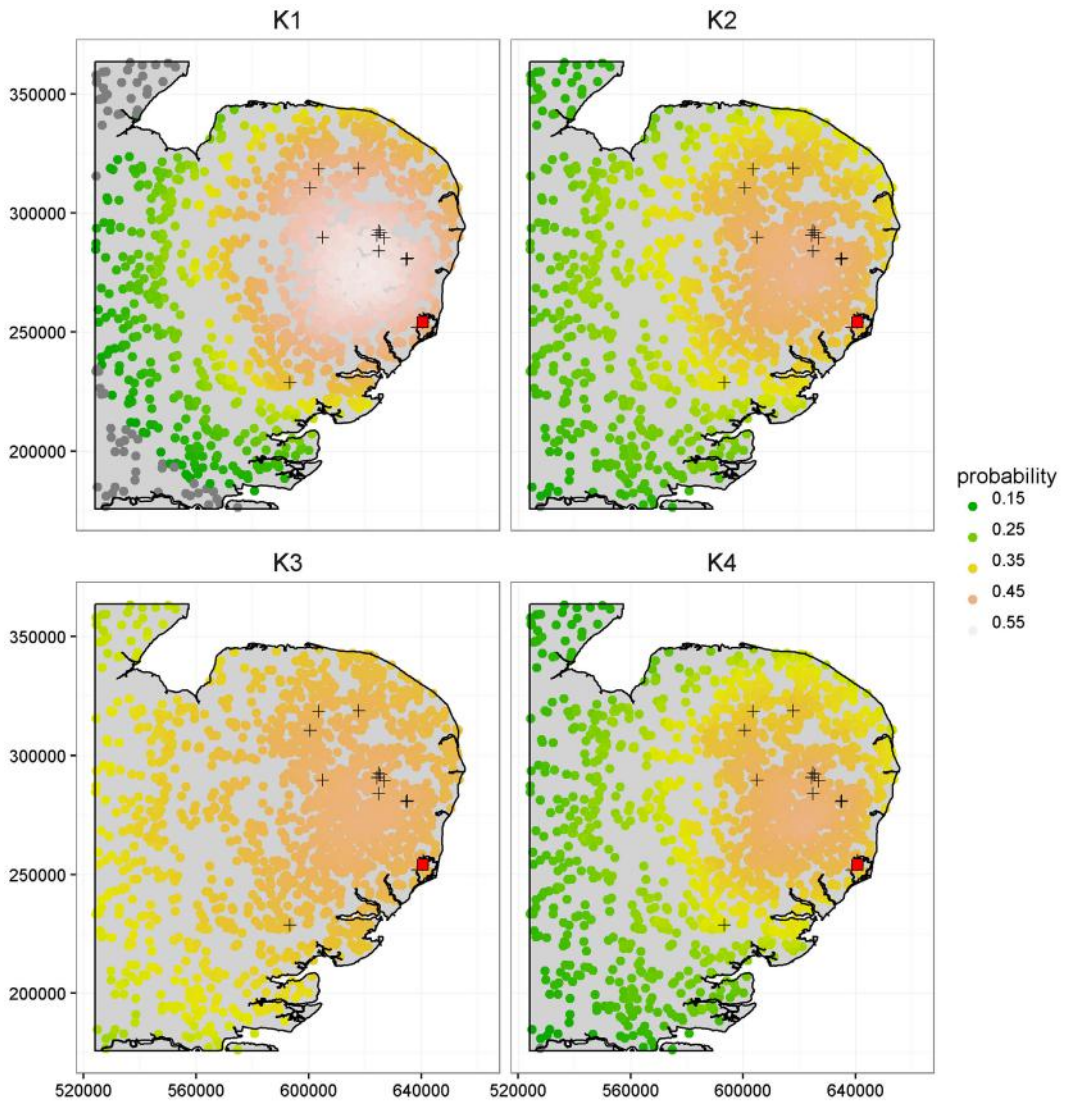


FIGURE 4 | Comparison of the risk maps using K_1, K_2, K_3, and K_4 at time $t = 90$ days, corresponding to a length of time sufficient to capture the early phase and small scale epidemics' behavior, based on the population from the CSF data. The 16 cases detected during the real outbreak are shown by the "+" symbols, along with the index case shown as a red square. The x and y axes are in meters.

4. DISCUSSION

The potential for large scale livestock epidemics to give rise to significant economic, welfare, and social costs (see, e.g., Ref (4, 5)) emphasizes the need for quantitative assessment (8, 9) of the risks associated with emerging and re-emerging pathogens. There is potential to use small localized outbreaks of emerging or re-emerging pathogens to inform such risk assessments before a large outbreak occurs. The key challenge addressed in this paper is to use data from small localized historic outbreaks (21) to inform quantitative risk assessment. We show that rigorous risk assessment based on small outbreaks can be achieved by combining state of the art methods for statistical inference in stochastic epidemic models. Moreover, this methodology was tested using simulated scenarios and by application to data on a small outbreak of CSF in East Anglia, UK (44).

We have shown that data-augmentation MCMC techniques (23, 25, 27) can be applied to continuous time models in order to generate Bayesian estimates of key characteristics of between-farm epidemics using data from small outbreaks consisting of farm locations and times at which disease is detected on farm. These estimated characteristics, impractical or difficult to measure directly, include unobserved exposure times, the distribution of times between exposure and disease detection, and the so-called transmission kernel describing the nature of spatial spread of disease between farms.

Analysis of inferences based on simulated data scenarios shows that when the fitted model has the same form as that used

TABLE 2 | Summary of model fit and risk assessments based on the CSF data: the results are provided for each kernel $K_1 - K_4$ as indicated in the first column.

	DIC_1	DIC_2	Pr($p < 5\%$)	n(risk > 0.45)	n(risk ∈ (0.35, 0.45))	n(risk < 0.15)
K_1	429	156	27.78%	866	395	72
K_2	**317**	157	**10.67%**	105	954	1
K_3	353	**156**	32.83%	1	1,466	0
K_4	411	158	19.47%	15	933	1

The next two columns indicate the computed DIC_1 and DIC_2 values with their smallest values under K_2 and K_3, respectively. The following column reports the proportion of p-values less than 5% (Pr($p < 5\%$)), resulting from testing the distribution of ILRs at each MCMC iteration, with the smallest proportion occurring under K_2. The other columns show the number of farms falling in defined intervals of probabilities of infection at time $t = 90$ days under the four kernels.
Bold font indicates the smallest values for the DICs and Pr($p < 5\%$) bringing out the selected model by each tool.

to generate the data, inferred parameter estimates are reliable even when using data from relatively small outbreaks, and the precision of such estimates increases with outbreak size. Fitted models can be used to conduct risk assessments of future outbreak scenarios that account for the uncertainty in parameter estimates. Such predicted risks are said to be drawn from a posterior predictive density and can be used to inform disease control efforts, e.g., targeting high risk farms (18, 61). Therefore, the method has the potential to inform the design and implementation of control measures such as the size of control zones used as part of wider movement restrictions and the geographically targeted use of vaccine and removal operations (17, 50, 62).

However, when analyzing real disease outbreaks all implementable models are, to varying degrees, approximations of the underlying system (63). We therefore considered scenarios in which the data were generated using a known model, and then used as the basis of inference for a set of models which varied according to the functional form of the spatial transmission kernel. Since the resulting risk assessments were found to be dependent on the structure of the model being fitted, model choice has important implications for such disease control policies.

Although risk assessment could be based on a weighted average across the set of models considered, here we focused on criteria used to select a single best-fit model. In particular, we considered two forms of the Deviance Information Criteria (DIC) and model selection based on analysis of so-called latent residuals (see methods for details). DIC is not uniquely defined for inference problems involving latent variables, e.g., missing exposure times, and we considered two variants, DIC_1 and DIC_2 (37). Latent residuals are designed to assess the fit of particular model components and we focused on Infectious Link Residuals (ILRs) to test the appropriateness of the form of the kernel density function (40). The set of fitted models included the data-generating model thus enabling an objective assessment of these model selection procedures. Multiple simulated replicate data sets for a range of outbreak sizes were generated and fitted so that coverage properties of inferences could be determined and reliability of model selection procedures assessed as a function of outbreak size.

For the scenarios considered it was found that DIC_2 was unreliable, but that model selection based on either DIC_1 or ILRs achieved high levels of reliability even when using data from relatively small outbreaks. For outbreak sizes of 11–15 and above, model selection based on ILRs out performed that based on DIC_1. For outbreaks with more than 20 cases, there is at least 90% of chance of selecting the correct kernel using ILRs, increasing to 100% for outbreaks with 31–35 cases or greater. However, reliability falls off with outbreak size and for data sets containing 6–10 cases, analysis of ILRs identified the data-generating model in a little over 50% of cases. In this very low data regime, DIC_1 was seen to give slightly more reliable model selection.

Ranking of models based on ILRs was more closely correlated, in comparison with DIC based assessments, with predicted risks under future disease incursion scenarios. For example, we found that two of the four transmission kernels considered in this study provided similar fits to the data especially when the density of farms was low and the number of short-range transmissions limited. In such scenarios, it was found that the kernel K_4 and a generalized form of the Cauchy kernel K_2 that captures short-range spread, predicted very similar risk profiles. Moreover, this similarity was reflected in model rankings based on the infectious link residuals, but not in DIC values. When farm density was higher a greater number of short-range transmissions were inferred and differences between K_4 and K_2 were evident in both rankings based on ILRs and in the posterior predicted risk profiles. In the case of field data on CSF, we found subtle differences in the risk profiles predicted under these two kernels, and this was flagged by analysis of the ILRs. However, the similarity between predictions under these models was not reflected in their DIC scores.

We have illustrated our approach using a particular set of between-farm epidemic models applicable to CSF outbreaks. However, the methodology is flexible and could be applied to a broad range of models, e.g., incorporating additional disease classes, multiple diagnostic tests, or the modeling of specific routes of infection. Here, we considered kernel transmission functions based on Euclidean distance and in many cases, the detailed information needed to parameterize more specific routes of infection may not be available. Euclidean distance-based kernel transmission functions have been extensively used by many authors (48–52) and according to Savill et al. (64), Euclidean distance is better predictor of transmission risk than shortest and quickest routes via road, and appropriate to most regions except where major geographical features intervene.

We tackled the challenging problem of extracting useful information from a knowledge of just the location of the susceptible population of farms and farm-level case detection data obtained from observations of small sized epidemic outbreaks (16 cases for the CSF epidemic). We have shown that using such limited data it is possible to perform reliable inferences

and quantify disease risks associated with individual farms. The work reported here focused on SIR epidemic models. In the case of CSF within individual animals, there is evidence of an exposed class, E, suggesting that for an individual level model an SEIR model would be more appropriate (20). However, in this paper the farm was taken to be the basic epidemiological unit since information was only available describing the infectious status of whole farms and following Stegeman et al. (58) farms were categorized according to an SIR framework. Definition of the exposed state at the whole farm level is somewhat problematic since exposed individuals may be moved between farms and therefore spread infection. We note that in modeling between-farm transmission, Boender et al. (43) do not consider an exposed farm state explicitly but do modulate for farm size. An interesting focus of future work could be a formal statistical assessment of SEIR versus SIR models of between-farm spread. However, in the small outbreak setting this would be challenging given the limited information available in the data. Other authors have sought to tackle limitations in observed case data by developing approaches that combine phylogenetic information with the case detections to increase the power of estimates (65–67). However, the methods presented here are applicable in the many situations where suitable phylogenetic data is not available.

The underlying methods described in this paper are generic and could be applied to a wide range of disease scenarios including other livestock diseases. However, while the methodology is generic it is critical to tailor models to the scenario of interest to represent key aspects of the disease dynamics and the available data. For example, such models could account for the role of wildlife in disease transmission, but in practice how this could be achieved is dependent on the extent to which data is available on the prevalence of the focal pathogen in the wildlife host (or hosts). In the absence of such data, it is likely that at best it may be possible to estimate a background infection rate from wildlife sources. Another interesting possibility would be to simultaneously model the spread of multiple pathogens including interactions between them, but again this would only yield meaningful estimates if suitable data were available.

In conclusion, we have developed a toolkit to reliably assess risks from potential future disease incursions using observational data from historic outbreaks that can be applied to support policy decisions relevant to contingency planning for emerging and re-emerging pathogens. We have shown that epidemic models based on discrete state continuous time Markov and semi-Markov processes and data-augmentation MCMC techniques enable reliable and rigorous statistical inferences and probabilistic risk assessments based on data from relatively small between-farm outbreaks. Moreover, recently introduced model assessment methodology based on latent residuals (40) enables candidate models to be ranked, on the basis of their fit to the available data, in a manner that is more reliable than standard DIC approaches (35, 37). We tested this toolkit in the data limited regime using both simulated data and by application to a real world outbreak of CSF with only 16 infected farms.

Our approach is designed to make the best possible use of the data available from even very small historic outbreaks. However, it is important to realize that such data may provide a biased view of future incursions. For example, if the region in which the historic outbreak occurred is not representative of the regions for which risk assessments are needed, then estimates obtained, e.g., of rates of transmission need to be applied with suitable caution. Nonetheless, our approach does provide a rational and statistically rigorous approach to extracting information on disease dynamics and transmission characteristics that are difficult, costly or impossible to measure directly. The quantification of such characteristics and associated uncertainty provides a practical and rational basis for the quantitative assessment of risks under future pathogen incursions.

ETHICS STATEMENT

Ethical approval not required as the data were collected from clinical examinations arising from a natural outbreak.

AUTHOR CONTRIBUTIONS

KG, GM, and TP conceptualized the ideas and formulated the goals. KG and GM developed and designed the methodology with the formal analyses carried by KG. TP provided the outbreak data and all authors contributed to the writing of the manuscript. GM secured funding for this project.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Animal and Plant Health Agency (APHA) for providing under confidentiality agreements, the 2000 CSF outbreak data in East Anglia, including details of infected premises, contact tracing investigation, and susceptible herds. The authors are also thankful to Giles Innocent at BioSS for his valuable comments. The authors thank the Editor and two referees for their positive comments which have helped to greatly improve the presentation.

FUNDING

KG, GM, and TP were funded by the Scottish Government Rural and Environment Science and Analytical Services Division (RESAS), as part of the Centre of Expertise on Animal Disease Outbreaks (EPIC); KG and GM also received funding from the RESAS Strategic Research Programme.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/fvets.2017.00016/full#supplementary-material.

REFERENCES

1. Gibbens JC, Wilesmith JW, Sharpe CE, Mansley LM, Michalopoulou E, Ryan JBM, et al. Descriptive epidemiology of the 2001 foot-and-mouth disease epidemic in Great Britain: the first five months. *Vet Rec* (2001) 149(24):729–43. doi:10.1136/vr.149.24.729
2. Dijkhuizen AA. The 1997-1998 outbreak of classical swine fever in The Netherlands. *Prev Vet Med* (1999) 42:135–7.
3. Stegeman A, Elbers A, de Smit H, Moser H, Smak J, Pluimers F. The 1997-1998 epidemic of classical swine fever in the Netherlands. *Vet Microbiol* (2000) 73(2–3):183–96. doi:10.1016/S0378-1135(00)00144-9
4. National Audit Office. *The 2001 Outbreak of Foot and Mouth Disease HC 939 Session 2001-2002*. Technical report. London: Stationary Office (2002).
5. Horst HS, de Vos CJ, Tomassen FHM, Stelwagen J. The economic evaluation of control and eradication of epidemic livestock diseases. *Rev Sci Tech* (1999) 18:367–79. doi:10.20506/rst.18.2.1169
6. Kocik J, Janiak J, Negut M. Preparedness against bioterrorism and re-emerging infectious diseases. *NATO Asi*. Amsterdam: IOS Press (2004).
7. Kapur GB, Smith JP. *Emergency Public Health: Preparedness and Response*. Sudbury, MA: Jones & Bartlett (2012). 568 p.
8. Andersson H, Britton T. *Stochastic Epidemic Models and Their Statistical Analysis*. New York: Springer (2000).
9. Jewell CP, Kypraios T, Neal P, Roberts GO. Bayesian analysis for emerging infectious diseases. *Bayesian Anal* (2009) 4(3):465–96. doi:10.1214/09-BA417
10. Woolhouse M, Chase-Topping M, Haydon D, Friar J, Matthews L, Hughes G, et al. Epidemiology: foot-and-mouth disease under control in the UK. *Nature* (2001) 411(6835):258–9. doi:10.1038/35077149
11. Haydon DT, Kao RR, Kitching RP. The UK foot-and-mouth disease outbreak – the aftermath. *Nat Rev Microbiol* (2004) 2:675–81. doi:10.1038/nrmicro960
12. Ferguson NM, Donnelly CA, Anderson RM. The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science* (2001) 292(5519):1155–60. doi:10.1126/science.1061020
13. Chowell G, Hayman JM, Bettencourt LMA, Castillo-Chavez C. *Mathematical and Statistical Estimation Approaches in Epidemiology*. Dordrecht: Springer (2009).
14. Pritchett J, Thilmany D, Johnson K. Animal disease economic impacts: a survey of literature and typology of research approaches. *Int Food Agribusiness Manage Rev* (2005) 8(1):23–45.
15. Knight-Jones TJD, Rushton J. The economic impacts of foot and mouth disease – what are they, how big are they and where do they occur? *Prev Vet Med* (2013) 112(3–4):161–73. doi:10.1016/j.prevetmed.2013.07.013
16. Pandey A, Atkins KE, Medlock J, Wenzel N, Townsend JP, Childs JE, et al. Strategies for containing Ebola in West Africa. *Science* (2014) 346(6212):991–5. doi:10.1126/science.1260612
17. Tildesley MJ, Savill NJ, Shaw DJ, Deardon R, Brooks SP, Woolhouse MEJ, et al. Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK. *Nature* (2006) 440:83–6. doi:10.1038/nature04324
18. Keeling MJ, Woolhouse MEJ, May RM, Davies G, Grenfell BT. Modelling vaccination strategies against foot-and-mouth disease. *Nature* (2003) 421:136–42. doi:10.1038/nature01343
19. Porphyre T, Boden LA, Correia-Gomes C, Auty HK, Gunn GJ, Woolhouse ME. How commercial and non-commercial swine producers move pigs in Scotland: a detailed descriptive analysis. *BMC Vet Res* (2014) 10:140. doi:10.1186/1746-6148-10-140
20. Backer JA, Hagenaars TJ, van Roermund HJW, de Jong MCM. Modelling the effectiveness and risks of vaccination strategies to control classical swine fever epidemics. *J R Soc Interface* (2009) 6:849–61. doi:10.1098/rsif.2008.0408
21. O'Dea EB, Pepin KM, Lopman BA, Wilke CO. Fitting outbreak models to data from many small norovirus outbreaks. *Epidemics* (2014) 6:18–29. doi:10.1016/j.epidem.2013.12.002
22. Gibson GJ, Renshaw E. Estimating parameters in stochastic compartmental models using Markov chain methods. *IMA J Math Appl Med Biol* (1998) 15:19–40. doi:10.1093/imammb/15.1.19
23. O'Neill P, Roberts G. Bayesian inference for partially observed stochastic epidemics. *J R Stat Soc A* (1999) 162(1):121–9. doi:10.1111/1467-985X.00125
24. Gibson GJ, Otten W, Filipe JAN, Cook A, Marion G, Gilligan CA. Bayesian estimation for percolation models of disease spread in plant populations. *Stat Comput* (2006) 16(4):391–402. doi:10.1007/s11222-006-0019-z
25. Jewell CP, Keeling MJ, Roberts GO. Predicting undetected infections during the 2007 foot-and-mouth disease outbreak. *J R Soc Interface* (2009) 6(41):1145–51. doi:10.1098/rsif.2008.0433
26. Cook A, Marion G, Butler A, Gibson G. Bayesian inference for the spatio-temporal invasion of alien species. *Bull Math Biol* (2007) 69(6):2005–25. doi:10.1007/s11538-007-9202-4
27. Neal P, Roberts G. A case study in non-centering for data augmentation: stochastic epidemics. *Stat Comput* (2005) 15(4):315–27. doi:10.1007/s11222-005-4074-7
28. Jeffreys H. Some tests of significance, treated by the theory of probability. *Proc Cambridge Philos Soc* (1935) 31(2):203–22. doi:10.1017/S030500410001330X
29. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc* (1995) 90(430):773–95. doi:10.1080/01621459.1995.10476572
30. Green PJ. Reversible jump MCMC computation and bayesian model determination. *Biometrika* (1995) 82(4):711–32. doi:10.1093/biomet/82.4.711
31. Hastie DI, Green PJ. Model choice using reversible jump Markov chain Monte Carlo. *Stat Neerl* (2012) 66(3):309–38. doi:10.1111/j.1467-9574.2012.00516.x
32. Cook AR, Otten W, Marion G, Gibson GJ, Gilligan CA. Estimation of multiple transmission rates for epidemics in heterogeneous populations. *Proc Natl Acad Sci U S A* (2007) 104(51):20392–7. doi:10.1073/pnas.0706461104
33. Knock ES, O'Neill PD. Bayesian model choice for epidemic models with two levels of mixing. *Biostatistics* (2014) 15(1):46–59. doi:10.1093/biostatistics/kxt023
34. Hsu CY, Yen AMF, Chen LS, Chen HH. Analysis of household data on influenza epidemic with bayesian hierarchical model. *Math Biosci* (2015) 261:13–26. doi:10.1016/j.mbs.2014.11.006
35. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc B* (2002) 64(4):583–639. doi:10.1111/1467-9868.00353
36. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. The deviance information criterion: 12 years on. *J R Stat Soc B* (2014) 76(3):485–93. doi:10.1111/rssb.12062
37. Celeux G, Forbes F, Robert CP, Titterington DM. Deviance information criteria for missing data models. *Bayesian Anal* (2006) 1(4):651–74. doi:10.1214/06-BA122
38. Gelman A, Meng X-L, Stern H. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Stat Sin* (1996) 6:733–807.
39. Draper D. Assessment and propagation of model uncertainty. *J R Stat Soc B* (1995) 57(1):45–97. doi:10.2307/2346087
40. Lau MSY, Marion G, Streftaris G, Gibson G. New model diagnostics for spatio-temporal systems in epidemiology and ecology. *J R Soc Interface* (2014) 11(4):20131093. doi:10.1098/rsif.2013.1093
41. Mintiens K, Laevens H, Dewulf J, Boelaert F, Verloo D, Koenen F. Risk analysis of the spread of classical swine fever virus through 'neighbourhood infections' for different regions in Belgium. *Prev Vet Med* (2003) 60(1):27–36. doi:10.1016/S0167-5877(03)00080-1
42. Staubach C, Teuffert J, Thulke HH. Risk analysis and local spread mechanisms of classical swine fever. *Epidemiol Santé Anim – Proceedings of 8th ISVEE*. Paris (1997). p. 31–2.
43. Boender GJ, van den Hengel R, van Roermund HJ, Hagenaars TJ. The influence of between-farm distance and farm size on the spread of classical swine fever during the 1997–1998 epidemic in The Netherlands. *PLoS One* (2014) 9(4):e95278. doi:10.1371/journal.pone.0095278
44. Paton D. Chapter 5.3: The reappearance of classical swine fever in England in 2000. In: Morilla A, Yoon K-J, Zimmerman JJ, editors. *Trends in Emerging Viral Infections of Swine*. Ames, IA: Iowa State Press (2002). p. 153–8.
45. Gay NJ. Modeling measles, mumps and rubella: implications for the design of vaccination programs. *Infect Control Hosp Epidemiol* (1998) 19(8):570–3. doi:10.2307/30141782
46. Shaw MW. Simulation of population expansion and spatial pattern when individual dispersal distributions do not decline exponentially with distance. *Proc R Soc B* (1995) 259:243–8. doi:10.1098/rspb.1995.0036
47. Deardon R, Brooks SP, Grenfell BT, Keeling MJ, Tildesley MJ, Savill NJ, et al. Inference for individual-level models of infectious diseases in large populations. *Stat Sin* (2010) 20(1):239–61.
48. Keeling MJ, Brooks SP, Gilligan CA. Using conservation of pattern to estimate spatial parameters from a single snapshot. *Proc Natl Acad Sci U S A* (2004) 101(24):9155–60. doi:10.1073/pnas.0400335101

49. Parry M, Gibson GJ, Parnell S, Gottwald TR, Irey MS, Gast TC, et al. Bayesian inference for an emerging arboreal epidemic in the presence of control. *Proc Natl Acad Sci U S A* (2014) 111:6258–62. doi:10.1073/pnas.1310997111
50. Hayama Y, Yamamoto T, Kobayashi S, Muroga N, Tsutsui T. Mathematical model of the 2010 foot-and-mouth disease epidemic in Japan and evaluation of control measures. *Prev Vet Med* (2013) 112:183–93. doi:10.1016/j.prevetmed.2013.08.010
51. Boender GJ, van Roermund HJ, de Jong MC, Hagenaars TJ. Transmission risks and control of foot-and-mouth disease in the Netherlands: spatial patterns. *Epidemics* (2010) 2:36–47. doi:10.1016/j.epidem.2010.03.001
52. Kypraios T. *Efficient Bayesian Inference for Partially Observed Stochastic Epidemics and a New Class of Semi-Parametric Time Series Models*. Ph.D. thesis, Lancaster University, Lancaster (2007).
53. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller A, Teller E. Equations of state calculations by fast computing machines. *J Chem Phys* (1953) 21:1087–91. doi:10.1063/1.1699114
54. Geman S, Geman D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* (1984) 6(6):721–41. doi:10.1109/TPAMI.1984.4767596
55. Spiegelhalter DJ, Thomas A, Best NG, Lunn DJ. WinBUGS version 1.4 user manual. *MRC Biostat Unit* Cambridge, UK (2003). Available from: http://www.mrc-bsu.cam.ac.uk/bugs/
56. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* (1977) 81(25):2340–61. doi:10.1021/j100540a008
57. Keeling MJ, Rohani P. *Modeling Infectious Diseases in Humans and Animals*. Princeton, NJ: Princeton University Press (2007).
58. Stegeman A, Elbers ARW, Smak J, de Jong MCM. Quantification of the transmission of classical swine fever virus between herds during the 1997–1998 epidemic in The Netherlands. *Prev Vet Med* (1999) 42(3–4):219–34. doi:10.1016/S0167-5877(99)00077-X
59. Le Potier M-F, Mesplède A, Vannier P. Chapter 15: Classical swine fever and other pestiviruses. In: Straw BE, Zimmerman JJ, D'Allaire S, Taylor DJ, editors. *Disease of Swine*. 9th ed. Oxford, UK: Blackwell Publishing (2006). p. 309–22.
60. Streftaris G, Gibson GJ. Bayesian inference for stochastic epidemics in closed populations. *Stat Modelling* (2004) 4(1):63–75. doi:10.1191/1471082X04st065oa
61. Jönsson B. Targeting high-risk populations. *Osteoporos Int* (1998) 8(Suppl 1):S13–6.
62. Hayama Y, Yamamoto T, Kobayashi S, Muroga N, Tsutsui T. Potential impact of species and livestock density on the epidemic size and effectiveness of control measures for foot-and-mouth disease in Japan. *J Vet Med Sci* (2016) 78(1):13–22. doi:10.1292/jvms.15-0224
63. Box GEP. Science and statistics. *J Am Stat Assoc* (1976) 71(356):791–9.
64. Savill NJ, Shaw DJ, Deardon R, Tildesley MJ, Keeling MJ, Woolhouse MEJ, et al. Topographic determinants of foot and mouth disease transmission in the UK 2001 epidemic. *BMC Vet Res* (2006) 2:3–10. doi:10.1186/1746-6148-2-3
65. Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc Biol Sci* (2012) 279:444–50. doi:10.1098/rspb.2011.0913
66. Morelli MJ, Thébaud G, Chadoeuf J, King DP, Haydon DT, Soubeyrand S. A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput Biol* (2012) 8:e1002768. doi:10.1371/journal.pcbi.1002768
67. Lau MSY, Marion G, Streftaris G, Gibson G. A systematic Bayesian integration of epidemiological and genetic data. *PLoS Comput Biol* (2015) 11:e1004633. doi:10.1371/journal.pcbi.1004633

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Veterinary Science

ORIGINAL RESEARCH
published: 02 June 2017
doi: 10.3389/fvets.2017.00082

Quantifying Preferences of Farmers and Veterinarians for National Animal Health Programs: The Example of Bovine Mastitis and Antimicrobial Usage in Switzerland

Bart H. P. van den Borne[1], Felix J. S. van Soest[2], Martin Reist[3] and Henk Hogeveen[2,4]*

[1] Vetsuisse Faculty, Veterinary Public Health Institute, University of Bern, Liebefeld, Switzerland, [2] Business Economics Group, Wageningen University & Research, Wageningen, Netherlands, [3] Federal Food Safety and Veterinary Office, Liebefeld, Switzerland, [4] Department of Farm Animal Health, Utrecht University, Utrecht, Netherlands

OPEN ACCESS

Edited by:
Moh A. Alkhamis, Kuwait Institute for Scientific Research, Kuwait

Reviewed by:
Luis Gustavo Corbellini, Universidade Federal do Rio Grande do Sul, Brazil
Flavie Luce Goutard, Agricultural Research Centre for International Development, France

***Correspondence:**
Bart H. P. van den Borne
bartvdborne@gmail.com

Specialty section:
This article was submitted to Veterinary Epidemiology and Economics, a section of the journal Frontiers in Veterinary Science

Received: *01 February 2017*
Accepted: *15 May 2017*
Published: *02 June 2017*

Citation:
van den Borne BHP, van Soest FJS, Reist M and Hogeveen H (2017) Quantifying Preferences of Farmers and Veterinarians for National Animal Health Programs: The Example of Bovine Mastitis and Antimicrobial Usage in Switzerland. Front. Vet. Sci. 4:82. doi: 10.3389/fvets.2017.00082

Bovine udder health in Switzerland is of a relatively high level. However, antimicrobial usage (AMU) seems high in comparison to other European countries also. A new udder health and AMU improvement program could improve this situation but it is uncertain whether there is support from the field. This study aimed to quantify preferences of dairy farmers and veterinarians for the start and design characteristics of a new national udder health and AMU improvement program in Switzerland. A total of 478 dairy farmers and 98 veterinarians completed an online questionnaire. Questions on their demographics and their mindset toward AMU were complemented with an adaptive choice-based conjoint interview, a novel conjoint analysis technique to quantify preferences of respondents for characteristics of a product for which multiple trade-off decisions must be made (here a bovine udder health and AMU improvement program). The conjoint analysis was followed by a multivariate multiple regression analysis to identify groups of respondents with different program design preferences. Logistic regression models were used to associate covariates with respondents' preference to start a new udder health and AMU improvement program. Most farmers (55%) and veterinarians (62%) were in favor of starting a new voluntary udder health and AMU improvement program, but the program design preferences agreed moderately between the two stakeholder groups. Farmers preferred an udder health and AMU improvement program that did not contain a penalty system for high AMU, was voluntary for all dairy herds, and aimed to simultaneously improve udder health and reduce AMU. Veterinarians preferred a program that had the veterinary organization and the government taking the lead in program design decision making, did not contain a penalty system for high AMU, and aimed to simultaneously improve udder health and reduce AMU. Differences between groups of farmers and veterinarians concerning their start preference were identified. Also, the magnitude of various program design preferences changed for farmers with different opinions toward AMU. The information obtained from this study may support the decision-making process

and the communication to the field afterward, when discussing national strategies to improve udder health and AMU in Switzerland.

Keywords: mastitis, dairy cows, adaptive choice-based conjoint analysis, multivariate multiple regression, animal disease program

INTRODUCTION

Bovine mastitis negatively affects milk quality (1, 2), animal welfare (3), the herd's profitability (4), and farmers' milking routine (5). Antimicrobial resistance (6, 7) and an increased risk of antibiotic residues in milk (8) are also associated with mastitis. Mastitis is the most common reason for applying antimicrobials to dairy cattle (9, 10). It therefore impairs the image of the dairy industry.

Bovine udder health in Switzerland is, from an international perspective, of a relative high level. Bulk milk and composite somatic cell counts are low and incidence rates of clinical mastitis are reported to be below estimates from other countries (11). The milk quality payment system in place largely explains this. Swiss farmers receive a penalty from their milk-processing company when their geometric bulk milk somatic cell count is ≥350,000 cells/ml. Some milk-processing companies have set lower penalty thresholds. On the other hand, Swiss farmers generally receive a bonus from their milk-processing companies when bulk milk somatic cell counts are <100,000 cells/ml. Despite a relatively good udder health, national annual failure costs of mastitis are estimated to be approximately 129 Million Swiss Francs for farmers, which equals to 198 Swiss Francs[1] per average cow per year (12). Also, antimicrobial resistance of mastitis pathogens is not uncommon, especially of coagulase-negative staphylococci species for which phenotypic resistance prevalence levels up to 47% were observed (13). Finally, approximately 70% of antimicrobial usage (AMU) in dairy cows is because of intramammary purpose (14), and there is evidence that sales of intramammary antimicrobials in Switzerland are high compared with other European countries (15). Switzerland has currently a federal strategy to improve antimicrobial resistance in the human and animal populations and the environment. However, a nation-wide udder health and AMU improvement program that could improve its situation in dairy herds, especially regarding production losses and AMU, is not existing yet. Similar programs have successfully been started in many countries, including Australia (16), Canada (17), Norway (18), and the Netherlands (19).

Designing a new national animal health control program is often a highly complex and political process in which trade-offs decisions are to be made between the epidemiological and cost-effectiveness of proposed interventions on one hand, and time restrictions, financial resources, responsibilities, and stakeholders' interests on the other hand. Issues raised are, for example, adaptation of existing legislation or payment schemes, the program's aims and tasks (what should it do?), its implementation (who should execute it?), and its financing (who should pay for it?). Designing a new animal health program is a complex task in which various stakeholders may have different program design preferences. *A priori* investigating these preferences offers a mechanism for shared decision making (20), provides understanding of stakeholders' opinions, and can be a starting point when discussing the program's final design (21). Incorporating stakeholders' preferences into the decision-making process might improve their compliance when the animal health program is implemented afterward (22). This is expected to be especially true for multifactorial animal health issues, such as bovine mastitis and AMU, where the involvement of stakeholders from the field is crucial for the success of a control program (23, 24). It is currently unclear whether a new dairy health program to improve udder health and intramammary AMU in Switzerland would be supported by the field and which components should ideally be included when an udder health and AMU improvement program is constructed.

The aim of this study was to elicit preferences of Swiss dairy farmers and veterinarians for the start and design characteristics of a new national udder health and intramammary AMU improvement program. It was also investigated whether groups of farmers and veterinarians with different start and design preferences could be identified.

MATERIALS AND METHODS

Study Population and Sample Size Estimation

Two questionnaires were conducted in this cross-sectional study; one aiming at farmers and one at veterinarians. The sampling frame for the farmer questionnaire consisted of 19,042 dairy farmers who were producing marketed milk, had ≥11 cows, and an email address deposited at the national milk quality payment organization in May 2014 (85% of all Swiss dairy herds; personal communication by TSM Trust Ltd., Bern, Switzerland). Seasonal communal pasture holdings and herds located in the Italian-speaking Canton of Ticino were excluded. Furthermore, 1,296 dairy herds randomly selected from the same sampling frame that were requested to participate in a parallel survey were excluded from the current study to avoid farmers receiving 2 questionnaires shortly after one another. The sampling frame for the veterinarian questionnaire consisted of all 438 Swiss cattle veterinarians that were registered with the Swiss Society for Ruminant Health (Schweizerische Vereinigung für Wiederkäuergesundheit, Bern, Switzerland).

Since no prior information was available on the preference of farmers and veterinarians to start a new dairy health program, sample size calculations were estimated for the proportion with the largest variance (i.e., a proportion of 0.50). A higher level of precision was accepted for veterinarians (10%) than for farmers (5%) because lower levels resulted in sampling fractions

[1]This equals to €182 or $199 (currency at May 1, 2017)

that were deemed unachievable. The sample size needed with 95% confidence in the sample frames of 19,042 dairy herds and 438 cattle veterinarians was 385 and 79, respectively, using Winepiscope 2.0. Given an expected response rate of 30% (25, 26), 1,283 dairy farmers and 264 cattle veterinarians needed to be contacted. Using a stratified (by Swiss Canton) random sampling approach, 1,300 dairy farmers (with stratum sample sizes proportional to the cantonal dairy herd population) were eventually requested to participate. All 438 registered cattle veterinarians were contacted.

Adaptive Choice-Based Conjoint (ACBC) Analysis

Elicitation of farmers' and veterinarians' preferences toward udder health and AMU improvement program characteristics was investigated using the computer-based ACBC analysis method (27) within SSI Web 8.4 (Sawtooth Software, Orem, UT, USA). In conjoint analysis, respondents make trade-off decisions between different product characteristics allowing to elicit the relative preference for each product characteristic. ACBC originates from market research and is the latest conjoint analysis technique to elicit preferences of respondents for characteristics of a product (28). Alternative conjoint techniques have been successfully applied in veterinary medicine and animal science to elicit farmers' preferences for management strategies (29–31) and as a tool for disease prioritization (32, 33).

In conjoint analysis, products are characterized by attributes and levels (e.g., the attribute color may contain the levels red, yellow and blue for the product chair). Following conjoint analysis terminology, attributes and levels for the "product" udder health and AMU improvement program were defined as program characteristics that decision makers have to consider. Program attributes and levels were initially defined based on a literature review and the authors' experience with national mastitis control programs. A draft list of program attributes and levels was then discussed with five experts involved in dairy cattle disease control in Switzerland, including two experts from the Federal Food Safety and Veterinary Office, two researchers from the Faculty of Veterinary Medicine, University of Bern, and one experienced practicing cattle veterinarian. The list of program attributes and levels was finalized after consulting an expert from Sawtooth Software to optimize methodological and statistical efficiency.

The two questionnaires consisted of four parts. The first part investigated the demography of respondents, their opinion on AMU in Switzerland (**Tables 1** and **2**), and their preference toward starting a voluntary udder health and AMU reduction program. This part included questions determining the herds' current size and the number of treated clinical mastitis cases during the previous calendar year. The following three parts of the questionnaires concerned the ACBC interview. In the second part, respondents were offered all program attributes and levels from which they had to select their most preferred level for each attribute individually. The outcome of this part was brought forward by the software to the third part of the questionnaire. This screening section allowed the identification of program levels that were systematically avoided or preferred by respondents. Here, rather than making definite choices, alternative program designs were offered to respondents for which they had to indicate whether these were a possibility for them or not. Respondents were also asked whether such program levels were completely unacceptable or an absolute requirement for them. Program design alternatives showed after this screening section satisfied those requirements by either explicitly excluding or including program levels. Respondents were presented with

TABLE 1 | Description of demographic and motivation characteristics of Swiss farmers.

Variable	Category	Frequency	
		N	%
Age (years)	0–42	161	33.7
	43–52	165	34.5
	≥53	152	31.8
Language	German	405	84.7
	French	73	15.3
Education	Certificate of competence	196	41.0
	Agricultural entrepreneur	59	12.3
	Professional degree	156	32.6
	University (of applied sciences)	14	2.9
	Other	53	11.1
Successor	Yes	153	32.0
	No	61	12.8
	Do not know yet	242	50.6
	I am the successor but have not taken over the farm yet	22	4.6
Production zone	Lowland	189	39.5
	Hilly region	96	20.1
	Mountainous region	193	40.4
Stall system	Free-stall	198	41.4
	Tie-stall	214	44.8
	Both	66	13.8
Production system	Conventional	73	15.3
	Environmental and animal friendly	345	72.2
	Organic	49	10.3
	Other	11	2.3
Dairy production is the main source of income	Yes	451	94.4
	No	27	5.7
Crop production	Yes	238	49.8
	No	240	50.2
Fruit production	Yes	91	19.0
	No	387	81.0
Poultry production	Yes	44	9.2
	No	434	90.8
Pig production	Yes	100	20.9
	No	378	79.1
Veal production	Yes	95	19.9
	No	383	80.1
Herd size (number of cows)	0–20	171	35.8
	21–30	153	32.0
	≥31	154	32.2
Incidence rate of farmer-reported treated clinical mastitis (/100 cow-years at risk)	0–12.5	162	33.9
	12.6–23.5	150	31.4
	≥23.6	166	34.7
Do you think that antimicrobial usage is too high in Swiss dairy herds?	Yes	164	34.3
	No	195	40.8
	I do not know	119	24.9

TABLE 2 | Description of demographic and motivation characteristics of Swiss ruminant veterinarians.

Variable	Category	Frequency	
		N	%
Gender	Male	68	69.4
	Female	30	30.6
Language	German	92	93.9
	French	6	6.1
Are you part of a joint practice?	Yes	38	38.8
	No	60	61.2
Number of vets working in practice	1	23	23.5
	2	18	18.4
	3	21	21.4
	≥4	36	36.7
Percentage of time allocated to dairy cows	0–55%	22	22.5
	55–99%	60	61.2
	100%	16	16.3
Practice is also covering companion animals?	Yes	74	75.5
	No	24	24.5
Practice is also covering horses?	Yes	75	76.5
	No	23	23.5
Practice is also covering pigs?	Yes	73	74.5
	No	25	25.5
Practice is also covering poultry?	Yes	23	23.5
	No	75	76.5
Practice is also covering exotic pets?	Yes	11	11.2
	No	87	88.8
Years working as a vet	0–10	26	26.5
	11–20	14	14.3
	21–30	34	34.7
	≥30	24	24.5
Veterinary specialization	National specialist or board certified	20	20.4
	No or other[a]	78	79.6
Proportion of antimicrobial sales being injectors	≥10%	38	38.8
	<10%	60	61.2
Do you think that antimicrobial usage is too high in Swiss dairy herds?	Yes	37	37.8
	No, sometimes, or I do not know	61	62.2

[a]*Complementary medicine, currently in education or other specializations (in other species for instance).*

a maximum of eight screening tasks, displaying four program alternatives each. These udder health and AMU improvement programs were then taken forward into the fourth and final part of the questionnaire to identify the overall best udder health and AMU improvement program. This part consisted of a traditional choice-based conjoint interview but with a restricted number of choice options to choose from since systematically avoided program levels in the screening section were excluded from this part of the questionnaire. Respondents had to choose one out of three presented program designs with a maximum of nine choice tasks (the exact number was conditional on respondents' answers in previous sections of the ACBC interview). This facilitated discrimination of slightly different alternative program designs from the respondents' most preferred program design (i.e., the one created in the first part of the ACBC interview). Preferred program design concepts in each choice task competed in subsequent choice tasks until the most preferred program design was identified. More detailed information on ACBC can be found in a technical paper of Sawtooth Software (28).

Data Collection

Questionnaires were being conducted in German and French. Translation from German to French was conducted by a professional translator with a background in agriculture. Farmers received a personalized email explaining the purpose of the study and an individualized link to the online questionnaire in January 2015. Cattle veterinarians were contacted by email by the Swiss Society for Ruminant Health with a description of the project and a general link to the online questionnaire the same day. Reminder emails were sent after 3 and 5 weeks. To increase response rates, vouchers for an agricultural and a veterinary wholesale were provided to 110 randomly selected farmers and 50 veterinarians. Participants were also informed that they would receive a summary of the project's results at the end of the study.

Statistical Analysis

Start Preference

Preference of farmers and veterinarians to start a new voluntary udder health and AMU improvement program and their associations with potential covariates were evaluated first. Farmers' and veterinarians' start preference was assessed by a single question with three possible outcomes ("*The government, scientists, and the dairy industry would like to improve antimicrobial usage and udder health in Swiss dairy herds. A new voluntary program that would support farmers with this should be started. Would you be in favor of such a program?*" Answers: "Yes", "I do not know," and "No"). Covariates potentially associated with the start preference of farmers (**Table 1**) were therefore investigated using multinomial logistic regression models. Assuming a constant herd size, the herd level incidence rate of farmer-reported treated clinical mastitis was calculated as the number of treated clinical mastitis cases divided by the herd size and was expressed per 100 cow-years at risk. After a univariable screening, covariates deemed relevant ($P < 0.25$, based on the Type 3 test) were retained for further investigation. When covariate pairs had an absolute correlation >0.50, the biological more meaningful covariate was selected to avoid multicollinearity. Multivariable statistical modeling subsequently consisted of a stepwise backward elimination process until all covariates were significantly ($P < 0.05$) associated with the preference of farmers to start a new udder health and AMU improvement program or considered a confounder. Confounding was assumed to occur when effect estimates changed >25% upon exclusion of a covariate from the model. Interaction terms were not evaluated.

A similar model approach was used to associate covariates with the preference of veterinarians to start a new udder health and AMU improvement program. However, a low number of observations in the "No" category ($n = 9$) resulted in quasicomplete separation for several covariates. The outcome categories "No" and "I do not know" were therefore merged and start preference was modeled as a binary outcome variable using binary logistic regression models.

Part-Worth Utility Estimation

In a conjoint analysis, and thus also in an ACBC analysis, preference of respondents is quantified by part-worth utilities. Individual-level part-worth utilities represent respondent's relative preference for each level within an attribute. Part-worth utilities are zero-centered with higher values representing more preferred attribute levels (27). Individual and mean part-worth utility values were estimated using the Hierarchical Bayes estimation procedure within Sawtooth Software (34). This estimation procedure borrows information from the entire population (prior) to determine how each respondent's parameter estimate (posterior) differs from the upper-level population mean. It does this in an iterative manner, using a Monte Carlo Markov Chain procedure, to constantly update parameter estimates until convergence has achieved. A total of 40,000 iterations were computed, from which the first 20,000 were discarded, to obtain individual-level part-worth utility values.

The relative preference (RP_i) of each attribute (A_i) for each respondent was subsequently derived according to:

$$RP_i = \frac{\text{Range } A_i}{\sum_{i=1}^{n} \text{Range } A_i} \times 100\%,$$

where Range A_i represents the difference between the highest and lowest part-worth utility values of attribute *i*, with *n* being the number of attributes. The relative preference represents the preference each respondent has for this attribute. Preferred program-levels result in larger part-worth utility values and therefore also in a higher relative preference. Sum of the relative preference of all attributes is 100% for each respondent. Mean relative preference values were calculated for each attribute to elicit their relevance in the farmer and veterinarian population.

Goodness-of-fit of the final hierarchical Bayes models was assessed by the percent certainty and the Root Likelihood. Both indicators reflect how well a model performs in comparison to a chance model alone and a perfect model. Percent certainty is 0% for a chance model and 100% for a perfect model. Root Likelihood is 0.33 for a chance model (1 divided by the number of program alternatives per choice task, which was 3 in this study) and 1.0 for a perfect model (34).

Program Design Preference

The preference of farmers and veterinarians for design characteristics of a new udder health and AMU improvement program was investigated next. First, it was assessed whether farmers and veterinarians preferred certain design attributes more than others. The sum of relative preference of all attributes (eight for farmers and seven for veterinarians) for each respondent equals 100%, implying that they would be equally preferred if the relative preference of each attribute would be 12.5 or 14.3% for farmers and veterinarians, respectively. Deviation from equal preference was determined using the Wilcoxon signed rank test. The most preferred attributes were defined as having a significant mean relative preference above the equal preference threshold value. Second, differences in mean standardized part-worth utilities of program attribute levels between farmers and veterinarians were assessed using the Wilcoxon rank sum test. A Bonferroni adjustment was applied to correct for multiple comparisons.

Covariates associated with program design preferences were identified in three analytical steps. The first step was to identify a global (i.e., multivariate) association of covariates with the relative preference of the eight (seven for the veterinarians) program attributes simultaneously. The second step elicited the individual (i.e., univariate) program attributes responsible for rejection of the global null hypothesis. The third and last step was to investigate which levels within identified program attributes were associated with the covariates significant in the previous steps. Statistical modeling was performed separately for farmers and veterinarians.

Multivariate multiple regression models correct for correlation between multiple outcome variables (one for each program attribute) and multiple comparisons, thereby reducing Type 1 errors (35). Such multivariate linear regression models were used within the first step. Each potential covariate (**Tables 1** and **2** plus respondents' start preference) was tested one at a time against the relative preference of all program attributes simultaneously. The multivariate Wilks' lambda *F* statistic was used to test the global hypothesis that all regression coefficients are zero across all program attributes for the evaluated covariate. Covariates deemed relevant ($P < 0.25$) were thereafter offered to a multivariate multiple regression in which a stepwise backward elimination process was conducted to identify all significant ($P < 0.05$) global associations between covariates and program attributes. Proportion of explained variance of the global model was derived as 1 − Wilks' lambda, which is the multivariate counterpart of the univariate R^2 (35). In step 2, univariate associations of covariates with the eight (or seven) program attributes were identified for the global significant covariates using the Type 3 test commonly used for univariate linear regression models. To correct for multiple comparisons, a Bonferroni adjustment was also made in this step. Proportions of explained variance of univariate models were reported by partial eta squared [η^2_P (35)]. The first two steps revealed associations between covariates and program attributes. Associations between covariates and program levels of relevant program attributes were investigated in the third and final step. Mean part-worth utility values between covariate categories were compared using the Tukey–Kramer multiple comparison test for program levels of all univariate associations identified in the previous step.

Regression modeling to associate covariates with start and program design preferences was performed using PROC LOGISTIC and PROC GLM in SAS 9.4 (Cary, NS, USA). Statistical significance was set at $P < 0.05$, except when noted otherwise. To evaluate potential non-response bias, the number of reminders (0, 1, or 2) being sent before respondents filled in the questionnaire was evaluated separately in all models evaluating the start and design preference of farmers and veterinarians.

RESULTS

Demography

Of 1,300 farmers and 438 cattle veterinarians contacted initially, 478 farmers (36.8%) and 98 veterinarians (22.4%) filled out their

respective questionnaires completely and were included in the statistical analysis. Seventy-three farmers were French speaking (**Table 1**) while only six veterinarians filled out the questionnaire in French (**Table 2**).

A description of farmers' demography is presented in **Table 1**. Farmers' median age was 49 (range: 26–81), and many (50.6%) did not know yet whether they had a successor or not. Most herds were located either in the lowland or mountainous regions of Switzerland. The distribution of housing systems (free-stall vs. tie-stall) was approximately equal. Dairy production was the main source of income for almost all farmers but many (82.0%) had additional agricultural production systems in place. Median farmers-reported incidence rate of clinical mastitis was 18.1 (range: 0–137.5) cases per 100 cow-year at risk.

Variables describing veterinarians' demography are presented in **Table 2**. Most veterinarians filling out the questionnaire were male and working in a practice that employed multiple veterinarians. Respondents dedicated most of their time practicing dairy health but most veterinarians (96.9%) worked in practices that serviced other species too. Median years working as a ruminant veterinarian was 25 (range: 1–43).

Preference to Start

Farmers and veterinarians were offered the following question: "*The government, scientists, and the dairy industry would like to improve antimicrobial usage and udder health in Swiss dairy herds. A new voluntary program that would support farmers with this should be started. Would you be in favor of such a program?*." Farmers (55.4%; 95% CI: 50.1–59.8%; $n = 265$) and veterinarians (62.2%; 95% CI: 52.4–71.2%; $n = 61$) mostly agreed with this statement; 20.7% (farmers; 95% CI: 17.3–24.6%; $n = 99$) and 9.1% (veterinarians; 95% CI: 4.9–16.5%; $n = 9$) disagreed; and 23.8% (farmers; 95% CI: 20.2–27.9%; $n = 114$) and 28.6% (veterinarians; 95% CI: 17.9–50.7%; $n = 28$) did not know. These proportions were significantly different ($\chi^2 = 7.16$; $P = 0.03$) between the two populations with veterinarians favoring the program more.

Table 3 reports the final multivariable multinomial logistic regression model for the preference of farmers to start a new voluntary udder health and AMU improvement program. Farmers who stated that AMU in Swiss dairy herds was too high or had no opinion on this had 3.2 and 2.0 times higher odds, respectively, for preferring to start a new national voluntary program than farmers that stated that AMU was not too high. Other covariates were not statistically associated with the preference to start a new voluntary udder health and AMU improvement program.

Covariates associated with veterinarians' preference to start a new voluntary udder health and AMU improvement program in the final multivariable logistic regression model are presented in **Table 4**. Ruminant veterinarians belonging to practices that also serviced poultry farms had a lower preference to start a new voluntary program than veterinarians belonging to practices not servicing poultry farms. Veterinarians who stated that ≥10% of their antimicrobial sales were attributed to the sales of intramammary antimicrobials had 3.2 times higher odds to prefer starting a new voluntary program compared to veterinarians that had <10% of their antimicrobial sales attributable to intramammary antimicrobials.

Model Fit of ACBC

Percent certainties of the final hierarchical Bayes models estimating program design preferences of farmers and veterinarians were 46.0 and 49.5%, respectively. Root likelihood of the farmer model was 0.63 and 0.65 for the veterinarian model.

TABLE 3 | Covariates in the final multinomial logistic regression model associated with the preference (yes or I do not know vs no) of farmers to start a new udder health and antimicrobial usage (AMU) improvement program.

Covariate	Category	Preference for a new program: I do not know vs no				Preference for a program: yes vs no				*P*-value type 3 test
		OR	95% CI		Wald *P*-value	OR	95% CI		Wald *P*-value	
			Lower	Upper			Lower	Upper		
Do you think that AMU is too high in Swiss dairy herds?	Yes	1.6	0.8	3.1	0.17	3.2	1.8	5.6	<0.0001	0.0005
	I do not know	1.6	0.8	3.2	0.16	2.0	1.1	3.7	0.02	
	No	Reference				Reference				

TABLE 4 | Covariates in the final logistic regression model associated with the preference (yes vs I do not know and no) of veterinarians to start a new animal health improvement program.

Covariate	Category	Frequency	Preference for starting a new program (%)	OR	95% CI	
					Lower	Upper
Covering other species: poultry	Yes	23	47.8	0.3	0.1	0.9
	No	75	66.7	Reference		
Proportion of antimicrobial sales being injectors	≥10%	38	73.7	3.2	1.2	8.6
	<10%	60	55.0	Reference		

Relative Preference of Program Attributes and Part-Worth Utilities

Relative preference of program attributes would have been 12.5% if farmers preferred them equally (the "equal preference value"). Farmers therefore preferred the program attributes "Bonus" (22.2%; 95% CI: 21.6–22.8), "Herd" (14.9%; 95% CI: 14.4–15.4), and "Aim" (14.6%; 95% CI: 14.1–15.1) more than the other five attributes (**Figure 1**). These three attributes had a relative preference above the equal preference value. Variation in farmers' relative preference for attributes was large though. Within the three most preferred attributes, farmers assigned the highest part-worth utility values to levels representing a new program that did not contain a penalty system for high AMU, was voluntary for all dairy herds, and aimed to simultaneously improve udder health and reduce AMU (**Table 5**).

Veterinarians' equal preference value of program attributes was 14.3%. Veterinarians preferred the program attributes "Decision" (19.8%; 95% CI: 18.7–20.9), "Bonus" (18.9%; 95% CI: 17.6–20.2), and "Aim" (16.1%; 95% CI: 15.1–17.1) more than the other four attributes (**Figure 2**). Like farmers' program design preference, variation in veterinarians' preference for program attributes was large (**Figure 2**). Veterinarians assigned the highest part-worth utility values to levels representing a new program that had the veterinary organization and the government taking the lead in the program design decision-making process, did not include a penalty system for high AMU, and aimed to improve udder health and reduce AMU simultaneously (**Table 5**).

Except for program characteristics related to the aims and tasks of the new program, farmers and veterinarians valued most program attribute levels differently (**Table 5**). Ranking of levels within some attributes also differed between farmers and veterinarians. This included the attribute "Decision," which was the most preferred program attribute for veterinarians. Ranking of levels did not differ within the other three most preferred program attributes (i.e., "Aim," "Bonus," and "Herd").

Respondent Characteristics and Differences in Program Design Preferences

The final multivariate multiple regression model investigating farmers' relative preference of program attributes is presented in **Table 6**. Farmers' opinion on AMU in Swiss dairy herds was the only covariate globally associated with the relative preference of program attributes ($P = 0.007$). Further investigation of the univariate associations identified that farmers' opinion on AMU was associated with the attributes "Bonus," "Decision" and "Payment" (**Table 6**). Proportions of explained variance were low, being 0.08 for the multivariate model and a maximum of 0.04 for the univariate models.

Subsequently associating farmers' opinion on AMU with the program levels of each identified univariate program attribute revealed some significant relationships (**Table 7**). Farmers who had the opinion that AMU was too high in Switzerland were still not favoring the introduction of a penalty system for high AMU but their oppositions were less strong than those from farmers disagreeing with the statement that AMU was too high. Moreover, farmers agreeing with the statement that AMU was too high were slightly more in favor of the farmer organization, rather than the breeding organizations, to take the lead in decision making when designing a potential new program. Still, both groups of farmers favored the dairy industry for this. Finally, farmers who had the opinion that AMU was too high

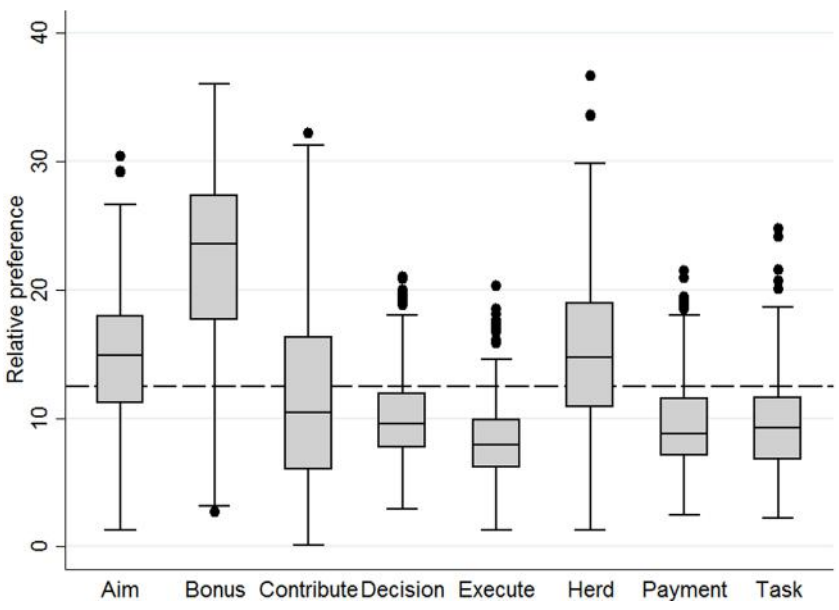


FIGURE 1 | Boxplots displaying relative preference of 478 farmers for attributes of a new Swiss animal health improvement program. The dashed line at 12.5% represents an equal preference.

TABLE 5 | Program attributes and levels evaluated in the adaptive choice-based analysis and comparison of standardized part-worth utilities for farmers and veterinarians in Switzerland.

Attribute	Description and levels	Farmers		Veterinarians		*P*-value
		Mean	SD	Mean	SD	
Aim	**Which aims should the program have?**					
	Improve udder health status and reduce AMU	50.9	28.0	54.7	26.0	0.18
	Reduce AMU while keeping udder health status constant	26.7	26.5	20.6	20.7	0.03
	Improve udder health status, no AMU improvement	−32.0	34.6	−42.0	30.9	0.004
	Reduce AMU, no udder health improvement	−45.5	25.3	−33.3	23.1	**<0.0001**
Bonus	**Should the program additionally include a bonus/malus system for AMU?**					
	No	69.6	54.0	52.3	42.4	**0.0005**
	Bonus low AMU	61.0	36.3	43.2	32.6	**<0.0001**
	Bonus low AMU and penalty high AMU	−54.2	36.2	−39.9	29.5	**<0.0001**
	Penalty high AMU	−76.5	31.7	−55.5	21.7	**<0.0001**
Decision	**Who should have the lead in decision making when designing the program?**					
	Dairy industry	23.4	26.2	2.0	42.0	**<0.0001**
	Breeding organizations	4.1	26.6	−29.0	22.5	**<0.0001**
	Farmers organization	0.3	24.3	−59.3	30.2	**<0.0001**
	Veterinary organization	−1.1	28.9	34.3	39.8	**<0.0001**
	University	−13.2	21.2	22.9	36.0	**<0.0001**
	Government	−13.5	27.2	29.0	38.0	**<0.0001**
Execute	**Who should execute the program?**					
	Dairy industry	11.9	24.8	−9.9	21.9	**<0.0001**
	Independent center of expertise	1.1	22.3	36.9	24.7	**<0.0001**
	Breeding organizations	0.8	25.3	-32.8	20.0	**<0.0001**
	Veterinary organization	−3.8	26.0	2.7	35.8	0.09
	Government	−10.0	21.1	3.2	21.0	**<0.0001**
Herd	**Which herds should the program target?**					
	Voluntary for all herds	46.8	45.5	12.9	39.6	**<0.0001**
	Compulsory for problem herds, voluntary for other herds	−2.1	40.4	9.6	31.7	**0.001**
	Compulsory for all herds	−44.7	35.2	−22.5	40.2	**<0.0001**
Payment	**Who should pay for the program?**					
	Government	21.5	29.3	5.9	27.9	**<0.0001**
	All three	11.7	23.7	23.1	18.1	**<0.0001**
	Government + dairy industry	3.6	16.4	3.5	15.9	0.85
	Government + breeding organizations	−4.9	14.0	−4.3	16.0	0.78
	Dairy industry	−6.9	28.0	7.9	21.1	**<0.0001**
	Dairy industry + breeding organizations	−8.2	17.8	−5.0	18.8	0.08
	Breeding organizations	−16.8	22.0	−31.1	19.3	**<0.0001**
Task	**What should be the main task for the program?**					
	Offering consulting for individual farmers	22.0	21.7	19.9	21.2	0.57
	Develop new knowledge	9.9	17.9	8.8	17.7	0.93
	Further education	5.0	20.1	8.3	18.5	0.14
	Honoring well-performing herds	−16.3	29.6	−12.9	18.7	0.02
	Mass communication	−20.6	27.3	−24.1	22.9	0.19
Contribute	**Mastitis in Switzerland costs on average CHF198 per cow per year. How much are you willing to contribute to the costs of the program (CHF per cow per year)?[a]**					
	CHF 0	31.7	44.3			
	CHF 1	4.3	25.7			
	CHF 2	−36.0	35.5			

Statistically significant values (after Bonferroni adjustment; P < 0.0015) are presented in bold.
AMU, antimicrobial usage.
[a]*This program attribute was only offered to farmers.*

in Switzerland opposed the option that breeding organizations should pay for the program stronger than those that did not share this opinion.

The final multivariate multiple regression model for veterinarians' design preferences identified one borderline significant association ($P = 0.05$; 1 − Wilks' lambda = 0.13) between the covariate describing whether veterinary practices were also servicing pig herds and veterinarians' relative preference of program attributes. However, further investigation of the univariate associations did not reveal any significant relationship when applying a Bonferroni adjustment (data not shown). Associations of this covariate with program attribute levels were therefore not further scrutinized.

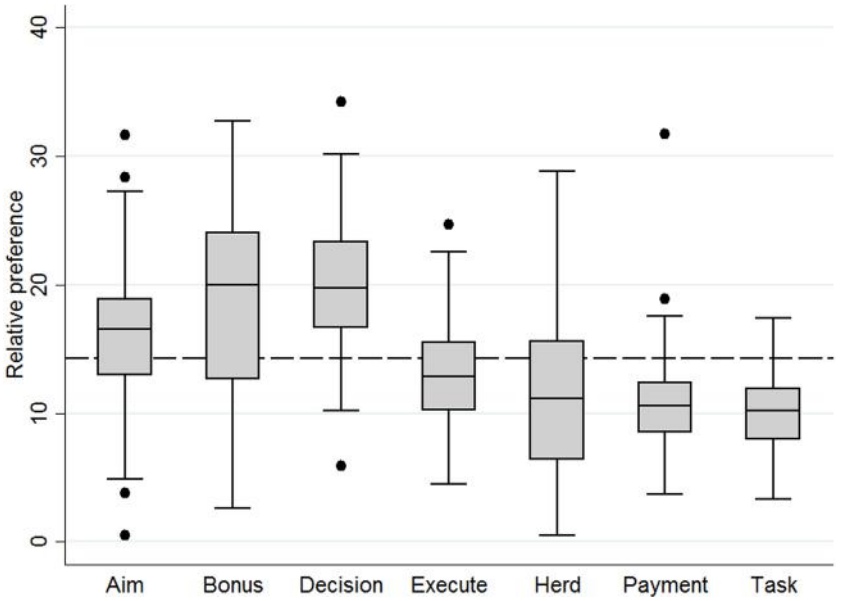


FIGURE 2 | Boxplots displaying relative preference of 98 veterinarians for attributes of a new Swiss animal health improvement program. The dashed line at 14.3% represents an equal preference.

TABLE 6 | Model output of the final multivariate multiple regression model investigating farmers' program design preferences.

	F	*P*	1 – Wilks' lambda	η^2_p
Global model: AMU opinion[a]	2.68	0.007	0.08	
Univariate models				
Attribute: aim	2.95	0.05		0.01
Attribute: bonus	8.91	0.0002		0.04
Attribute: contribute	1.42	0.24		0.01
Attribute: decision	5.24	0.006		0.02
Attribute: execute	4.50	0.01		0.02
Attribute: herd	0.94	0.39		0.00
Attribute: payment	7.01	0.001		0.03
Attribute: task	1.20	0.30		0.01

Significance in the univariate models was set at P < 0.006 to correct for multiple comparisons.

[a]*Do you think that antimicrobial usage is too high in Swiss dairy herds?*

TABLE 7 | Mean (and SE) standardized part-worth utilities of program attributes and levels for groups of farmers with a different antimicrobial usage (AMU) opinion.

Attribute	Level	Do you think that AMU is too high in Swiss dairy herds?		
		Yes	I do not know	No
Bonus	No	51.5 (4.3)[a]	73.2 (4.6)[b]	82.7 (3.7)[b]
	Bonus low AMU	62.1 (3.1)	66.0 (3.0)	57.0 (2.5)
	Bonus low AMU and penalty high AMU	−45.0 (3.0)[a]	−56.8 (3.1)[b]	−60.3 (2.5)[b]
	Penalty high AMU	−68.7 (2.7)[a]	−82.4 (2.7)[b]	−79.4 (2.1)[b]
Decision	Dairy industry	25.0 (2.2)	22.1 (2.3)	22.8 (1.8)
	Farmers organization	3.4 (2.0)[a]	−4.2 (1.9)[b]	0.5 (1.7)[ab]
	Breeding organizations	0.6 (2.2)[a]	3.6 (2.6)[ab]	7.5 (1.7)[b]
	Veterinary organization	−0.8 (2.4)	−1.0 (2.7)	−1.5 (1.9)
	Government	−13.8 (2.3)	−9.0 (2.6)	−16.0 (1.7)
	University	−14.3 (1.7)	−11.4 (2.0)	−13.4 (1.4)
Payment	Government	20.0 (2.6)	24.2 (2.7)	21.0 (1.9)
	All three	14.7 (2.0)	11.6 (2.0)	9.1 (1.6)
	Government + dairy industry	4.5 (1.4)	3.9 (1.5)	2.7 (1.1)
	Government + breeding organizations	−5.3 (1.1)	−5.2 (1.3)	−4.3 (0.9)
	Dairy industry	−6.0 (2.3)	−8.8 (2.5)	−6.6 (2.0)
	Dairy industry + breeding organizations	−8.0 (1.6)	−9.1 (1.6)	−7.7 (1.1)
	Breeding organizations	−19.9 (1.8)[a]	−16.6 (2.0)[ab]	−14.2 (1.5)[b]

Mean part-worth utilities within a row with various superscripts differ significantly.

Non-response Bias

Farmers' preference to start a new udder health and AMU improvement program was 64.3% for respondents not receiving an email reminder (early respondents) and significantly ($P = 0.04$) decreased to 49.4% for farmers receiving two reminding emails (late respondents). Such an association was not identified for the veterinarian dataset ($P = 0.55$). The number of reminders being sent was not globally associated with farmers' ($P = 0.33$) or veterinarians' (P = 0.96) program design preferences either.

DISCUSSION

This study identified that more than half of the respondents favored starting a new voluntary national udder health and AMU improvement program. Approximately every fourth respondent was undecided and 10 (veterinarians) to 20% (farmers) disapproved of this idea. There does seem to be support from the field to initiate a new voluntary udder health and AMU improvement program when the current Swiss strategy to improve AMU and

antimicrobial will be extended to dairy herds. It is unclear, however, whether these proportions are high enough to warrant an actual start of a new program. Its acceptance is likely to improve if it is accompanied with a communication campaign promoting its potential benefits (36). Moreover, the new national udder health and intramammary AMU improvement program referred to a voluntary program in which farmers would be supported with their activities improving udder health and AMU. Respondents may have been less positive if the program would have involved compulsorily activities or a more restrictive legislation. Also, only 37% of farmers answered the questionnaire, and some indication for non-response bias was found when evaluating farmers' start preference but not for their program design preferences. The response rate was in agreement with other studies (25, 37), and demographics of farmers agreed with two previous studies investigating the same target population (25, 38) with one exception. The median farmer-reported incidence rate of clinical mastitis was higher than observed previously (11). The latter may have been a result of farmers becoming more sensitive to the topic of udder health and AMU because of a higher awareness in the farming community and society in general. They, therefore, may diagnose CM more often. Controlling bodies may have become stricter also, resulting in a potential higher reporting rate.

Response rate of veterinarians was lower at 22%, but years of experience and proportion of male respondents were in agreement with a previous survey conducted among Swiss cattle veterinarians (26). Moreover, the range in years of experience working as a dairy cattle veterinarian indicated that there was no age bias and that also older generations were reached by the online questionnaire. Indications for non-response bias were not identified in the veterinarian dataset either. It is therefore believed that the responding veterinarians represented their target population well.

Preferences of farmers and veterinarians for program design characteristics agreed moderately. Both stakeholder groups preferred a program that aimed to improve udder health and AMU simultaneously and did not include a penalty system for high AMU. These aims are in line with the strategy of the Federal Food Safety and Veterinary Office to intervene on farmers and veterinarians with a high antimicrobial consumption. Moreover, such achievements can potentially be made through the quality payment system existing in Switzerland. Farmers generally receive a bonus for their milk price when their bulk milk somatic cell count is below 100,000 cells/ml. There are no such thresholds on AMU currently. Incorporating such a threshold (after setting up a national database to register AMU on herd level) as an extra criterion for farmers to receive a financial bonus is expected to result in an improvement on AMU. Previous research has shown that farmers are more sensitive to penalties than to bonuses, as investigated for bulk milk somatic cell counts in the Netherlands (39). A penalty system might therefore be more effective in improving udder health and AMU. Nonetheless, stricter criteria for farmers to receive a bonus are also expected to result in an AMU improvement because it would take away the financial incentive to use antimicrobials to achieve a better milk price. Moreover, they are expected to be perceived less negative than receiving a penalty for high AMU (40) but this was not evaluated in the current study.

Besides a change in the milk quality payment scheme or another change in legislation, there are also other means to improve both udder health and AMU. This is evident by examples from national mastitis control programs successfully conducted elsewhere (16–19). Unfortunately, improving udder health on a voluntary basis has been proven difficult for Swiss dairy herds as identified in a recently conducted multiarm randomized field trial (24). The intervention in which farmers formed peer study group meetings to study mastitis-related topics was able to reduce AMU though while keeping the herds' udder health status constant (24). A more realistic aim of a potential new national udder health and AMU improvement program would therefore be to reduce AMU while keeping the country's udder health status constant. This was only the second preferred aim of both farmers and veterinarians but such achievements are feasible (24, 41). AMU and udder health are highly correlated (9, 10, 14), and any efforts to control mastitis by enhancing prevention and non-antimicrobial intervention strategies are therefore assumed to result in a decrease in AMU. Reducing AMU in dairy herds and improving its udder health status should therefore not be seen separately and be targeted simultaneously in a national control program. The program should then, however, only be evaluated on its improvement in AMU and not on its udder health improvement other than keeping it constant.

Farmers and veterinarians differed in their preferences concerning other program design characteristics. Farmers preferred a program that is voluntary. That agreed with the preference of veterinarians but this stakeholder group deemed this attribute to be of less importance. Veterinarians, on the other hand, preferred a program that had the veterinary organization and the government taking the lead in the program design decision-making process whereas the farmers preferred the dairy industry to have the lead. However, farmers generally gave a lower relative preference value to this attribute, implying that this attribute was less important to them. It is therefore believed that they would not mutually exclude each other. A voluntary program with the veterinary organization and the government having the lead in the decision-making process would still satisfy the preferences of both stakeholder groups. Other preference differences between both stakeholder groups were also identified. But those concerned again program attributes and levels that were preferred less. Not much opposition from these two stakeholder groups is therefore expected if these aspects are not fully met during the decision-making process. Moreover, incorporating these aspects in the communication of the final program design is expected to take away some of the opposition.

Farmers' mindset toward AMU, as assessed by one single question, was the only covariate associated with farmers' preference for starting a new udder health and AMU improvement program and for their preferred program design characteristics. None of the demographic variables explained any of the variation in start and program design preferences of farmers. Also, explained variation of the final multivariate multiple regression model was low. It can thus be hypothesized that farmers' start and program design preferences for a new udder health and

AMU improvement program may be more explained by their mindset toward udder health, AMU, or national disease control programs in general than by their demographic characteristics. Further research is needed to scrutinize this underlying socio-psychological construct. Nonetheless, some differences in start and program design characteristics between groups of farmers with various mindsets toward AMU were identified. Farmers had a stronger preference to start a new voluntary udder health and AMU reduction program when stating that AMU was too high in Swiss dairy herds or when they had no opinion on this. Those farmers, an approximate 60% of the population, therefore not only acknowledged the problem of high AMU (or were indifferent) but also supported national strategies to improve it. This included a less strong, but still existing, opposition toward the introduction of a penalty system for high AMU. Current debates in society and other strategies facilitating the recognition of high AMU could further contribute to a less strong opposition of stricter milk quality payment legislation. There were also some subtle, but significant, differences in preferences for sectoral organizations that should have a lead in decision making and that should pay for the new udder health and AMU improvement program between groups of farmers with various mindsets of toward AMU. The identification of such associations adds to the communication to the field after decision makers have discussed program alternatives.

For veterinarians, on the other hand, differences in their preference to start a new udder health and AMU improvement program according to their demographics were identified. First, the observation that veterinarians, who are earning ≥10% from their antimicrobial sales from intramammary antimicrobials, are more motivated to start a new udder health and AMU improvement program than veterinarians earning less sounds contradictory at first. An improved on-farm udder health and AMU, resulting from implementing preventive mastitis management measures as advised by a new program, are expected to result in a decreased, rather than an increased, sales of intramammary antimicrobials at the veterinary practice level. However, herd health management is not commonly applied by Swiss cattle veterinarians, and practices selling more intramammary antimicrobials are therefore assumed to have such high sales because they attempt to improve udder health in their dairy herds by treating more (e.g., subclinical) mastitis cases to lower the infectious pressure within the herd (42, 43). Increased AMU levels in Swiss dairy herds trying to improve udder health have been observed before (24). Considering the second covariate, there are only very few veterinarians servicing poultry farms in Switzerland given the small but highly organized nature of this production system. The proportion of veterinarians working at practices also servicing poultry farms (23%) therefore should be interpreted as the proportion of veterinarians working at practices servicing backyard flocks rather than specialized poultry farms. Such veterinarians may thus be less specialized in cattle health, resulting in a lower motivation to start a new national udder health and AMU improvement program. However, interpretation of both covariates identified in the final logistic regression models remains speculation, and no causal conclusions can be drawn either from this study given its cross-sectional study design. Interpretation should therefore be cautious. No global associations between covariates and veterinarians' program design characteristics were identified, which is probably a result of the smaller sample size of this dataset (35).

Adaptive choice-based conjoint analysis was used to elicit respondents' preferences for design characteristics of a new udder health and AMU improvement program. ACBC is a novel quantitative methodology in the field of veterinary medicine and animal science. The novel aspect lies in the adaptive nature of the interview in comparison to a standard choice-based conjoint interview (27). Respondents participating in an ACBC interview first select program characteristics that they consider most important. This consideration set is then brought forward in the remaining part of the interview in which they are jointly evaluated with alternative program designs (28). In a standard (non-adaptive) choice-based conjoint interview, a fixed number of program alternatives are offered to respondents including program attributes that may not be relevant to them (27). Results of a choice-based conjoint interview subsequently may not reflect the information that is relevant for the respondent's situation when evaluating program design alternatives because the latter may not be close to the respondents' ideal (28). Because ACBC interviews are more personalized than choice-based conjoint interviews, it makes them more engaging for respondents (28).

This study was limited by its cross-sectional design. Preferences of farmers and veterinarians were assessed once but may change over time resulting from discussions in society and actions implemented by governmental bodies, industry, and others. Moreover, this study assessed stakeholders' preferences for starting a new udder health and AMU improvement program and its design. Stakeholders' preferences of potential interventions, e.g., the creation of peer study group meetings, financial support for culling mastitic cows, more affordable diagnostics, etc., were not scrutinized. Further research is thus needed to investigate stakeholders' preferences for the actual implementation of a program. Moreover, it was not the aim of this study to identify the perceived monetary and non-monetary benefits or disadvantages of a new udder health and AMU improvement program. Preferred design characteristics of the new udder health and AMU improvement program may thus differ from the most beneficial or practical design. Nonetheless, the results of this study facilitate discussions among decision makers. It should be noted also that this study investigated the start and design preferences of stakeholders in the Swiss context (e.g., concerning legislation and the sectoral organization of the dairy industry). Results may therefore be difficult to apply in other countries or regions, except when having a similar dairy industry. This study serves as an example on how to assess stakeholders' preferences for new national animal health control programs.

In conclusion, most farmers and veterinarians enrolled in this survey preferred starting a new voluntary udder health and AMU improvement program in Switzerland. Particularly, they preferred a new program that aims to improve udder health and AMU simultaneously, does not contain a penalty system for high AMU, is voluntary for all dairy herds, and have the veterinary organization and the government taking the lead in the program design decision-making progress. Differences between groups of farmers and veterinarians

concerning their start and program design preferences were also identified. The results of this study were not communicated with decision makers, yet they may support the decision-making process and to its communication afterward, when designing a new udder health and AMU improvement program for Switzerland.

ETHICS STATEMENT

According to Swiss legislation, no ethical approval was required for this study since no sensitive data were collected. All participants were involved voluntarily, were informed about the research objectives of the study, and gave their consent for participation in the study. All participants were assured anonymity.

AUTHOR CONTRIBUTIONS

BB conceived and designed the study, collected and analyzed the data, and drafted the manuscript. All other authors provided input on the design of the study, helped interpreting study results, and critically revised the manuscript. All authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

The input of the five experts on defining attributes and levels and the assistance of Anna-Alita Schwendner (Veterinary Public Health Institute, Liebefeld, Switzerland) with sending the online questionnaires was highly appreciated. TSM Trust Ltd. (Bern, Switzerland) and the Swiss Society for Ruminant Health (Bern, Switzerland) are kindly thanked for providing contact details of farmers and veterinarians, respectively.

FUNDING

This study was financed by grant 1.13.17 of the Federal Food Safety and Veterinary Office (Liebefeld, Switzerland). Funding was received by BB. Funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

1. Schukken YH, Wilson DJ, Welcome F, Garrison-Tikofsky L, Gonzalez RN. Monitoring udder health and milk quality using somatic cell counts. *Vet Res* (2003) 34:579–96. doi:10.1051/vetres
2. Johler S, Weder D, Bridy C, Huguenin M-C, Robert L, Hummerjohann J, et al. Outbreak of staphylococcal food poisoning among children and staff at a Swiss boarding school due to soft cheese made from raw milk. *J Dairy Sci* (2015) 98:2944–8. doi:10.3168/jds.2014-9123
3. Leslie KE, Petersson-Wolfe CS. Assessment and management of pain in dairy cows with clinical mastitis. *Vet Clin North Am Food Anim Pract* (2012) 28:289–305. doi:10.1016/j.cvfa.2012.04.002
4. Hogeveen H, Huijps K, Lam TJGM. Economic aspects of mastitis: new developments. *N Z Vet J* (2011) 59:16–23. doi:10.1080/00480169.2011.547165
5. Jansen J, van den Borne BHP, Renes RJ, van Schaik G, Lam TJGM, Leeuwis C. Explaining mastitis incidence in Dutch dairy farming: the influence of farmers' attitudes and behaviour. *Prev Vet Med* (2009) 92:210–23. doi:10.1016/j.prevetmed.2009.08.015
6. Saini V, McClure JT, Scholl DT, DeVries TJ, Barkema HW. Herd-level association between antimicrobial use and antimicrobial resistance in bovine mastitis *Staphylococcus aureus* isolates on Canadian dairy farms. *J Dairy Sci* (2012) 95:1921–9. doi:10.3168/jds.2011-5065
7. Oliver SP, Murinda SE. Antimicrobial resistance of mastitis pathogens. *Vet Clin North Am Food Anim Pract* (2012) 28:165–85. doi:10.1016/j.cvfa.2012.03.005
8. van Schaik G, Lotem M, Schukken YH. Trends in somatic cell counts, bacterial counts, and antibiotic residue violations in New York State during 1999-2000. *J Dairy Sci* (2002) 85:782–9. doi:10.3168/jds.S0022-0302(02)74136-2
9. Saini V, McClure JT, Léger D, Dufour S, Sheldon AG, Scholl DT, et al. Antimicrobial use on Canadian dairy farms. *J Dairy Sci* (2012) 95:1209–21. doi:10.3168/jds.2011-4527
10. Kuipers A, Koops WJ, Wemmenhove H. Antibiotic use in dairy herds in the Netherlands from 2005 to 2012. *J Dairy Sci* (2016) 99:1632–48. doi:10.3168/jds.2014-8428
11. Gordon PF, van den Borne BH, Reist M, Kohler S, Doherr MG. Questionnaire-based study to assess the association between management practices and mastitis within tie-stall and free-stall dairy housing systems in Switzerland. *BMC Vet Res* (2013) 9:200. doi:10.1186/1746-6148-9-200
12. Heiniger D, van den Borne BHP, Lechner I, Tschopp A, Strabel D, Steiner A, et al. Kosten-Nutzen-Analyse einer Intervention zur Verbesserung der Eutergesundheit in Schweizer Milchviehbetrieben. *Schweiz Arch Tierheilkd* (2014) 156:473–81. doi:10.1024/0036-7281/
13. Frey Y, Rodriguez JP, Thomann A, Schwendener S, Perreten V. Genetic characterization of antimicrobial resistance in coagulase-negative staphylococci from bovine mastitis milk. *J Dairy Sci* (2013) 96:2247–57. doi:10.3168/jds.2012-6091
14. Menéndez González S, Steiner A, Gassner B, Regula G. Antimicrobial use in Swiss dairy farms: quantification and evaluation of data quality. *Prev Vet Med* (2010) 95:50–63. doi:10.1016/j.prevetmed.2010.03.004
15. European Medicines Agency. *Sales of Veterinary Antimicrobial Agents in 26 EU/ EEA Countries in 2013 (Fifth ESVAC Report)*. London, United Kingdom (2015).
16. Brightling P, Dyson R, Hope A, Penry J. A national programme for mastitis control in Australia: countdown downunder. *Vet J* (2009) 62(Suppl 4):S52–8. doi:10.1186/2046-0481-62-S4-S52
17. Reyher KK, Dufour S, Barkema HW, Des Côteaux L, Devries TJ, Dohoo IR, et al. The national cohort of dairy farms – a data collection platform for mastitis research in Canada. *J Dairy Sci* (2011) 94:1616–26. doi:10.3168/jds.2010-3180
18. Østerås O, Sølverød L. Norwegian mastitis control programme. *Vet J* (2009) 62(Suppl 4):S26–33. doi:10.1186/2046-0481-62-S4-S26
19. Lam TJGM, van den Borne BHP, Jansen J, Huijps K, van Veersen JCL, van Schaik G, et al. Improving bovine udder health: a national mastitis control program in the Netherlands. *J Dairy Sci* (2013) 96:1–11. doi:10.3168/jds.2012-5958
20. Sidani S, Epstein D, Miranda J. Eliciting patient treatment preferences: a strategy to integrate evidence-based and patient-centered care. *Worldviews Evid Based Nurs* (2006) 3:116–23. doi:10.1111/j.1741-6787.2006.00060.x
21. Koelen MA, van den Ban AW. *Health Education and Health Promotion*. Wageningen: Wageningen Academic Publishers (2004).
22. Preference Collaborative Review Group. Patients' preferences within randomised trials: systematic review and patient level meta-analysis. *BMJ* (2008) 337:a1864. doi:10.1136/bmj.a1864
23. Green MJ, Leach KA, Breen JE, Green LE, Bradley AJ. National intervention study of mastitis control in dairy herds in England and Wales. *Vet Rec* (2007) 160:287–93. doi:10.1136/vr.160.9.287
24. Tschopp A, Reist M, Kaufmann T, Bodmer M, Kretzschmar L, Heiniger D, et al. A multiarm randomized field trial evaluating strategies for udder health improvement in Swiss dairy herds. *J Dairy Sci* (2015) 98:840–60. doi:10.3168/jds.2014-8053
25. Gordon PF, Kohler S, Reist M, van den Borne BHP, Menéndez Gonzáles S, Doherr MG. Baseline survey of health prophylaxis and management practices on Swiss dairy farms. *Schweiz Arch Tierheilkd* (2012) 154:371–9. doi:10.1024/0036-7281/a000367
26. Becker J, Reist M, Friedli K, Strabel D, Wüthrich M, Steiner A. Current attitudes of bovine practitioners, claw-trimmers and farmers in Switzerland to pain and painful interventions in the feet in dairy cattle. *Vet J* (2013) 196:467–76. doi:10.1016/j.tvjl.2012.12.021

27. Orme B. *Getting Started with Conjoint Analysis – Strategies for Product Design and Pricing Research*. Manhattan Beach, CA: Research Publishers LLC (2014).
28. Sawtooth Software. *ACBC Technical Paper*. Orem, UT: Sawtooth Software (2014).
29. van Soest FJS, Mourits MCM, Hogeveen H. European organic dairy farmers' preference for animal health management within the farm management system. *Animal* (2015) 9:1875–83. doi:10.1017/S175173111500141X
30. Huijps K, Hogeveen H, Lam TJGM, Huirne RBM. Preferences of cost factors for mastitis management among Dutch dairy farmers using adaptive conjoint analysis. *Prev Vet Med* (2009) 92:351–9. doi:10.1016/j.prevetmed.2009.08.024
31. van Schaik G, Dijkhuizen AA, Huirne R, Benedictus G. Adaptive conjoint analysis to determine perceived risk factors of farmers, veterinarians and AI technicians for introduction of BHV1 to dairy farms. *Prev Vet Med* (1998) 37:101–12. doi:10.1016/S0167-5877(98)00102-0
32. Stebler N, Schuepbach-Regula G, Braam P, Falzon LC. Weighting of criteria for disease prioritization using conjoint analysis and based on health professional and student opinion. *PLoS One* (2016) 11:e0151394. doi:10.1371/journal.pone.0151394
33. Ng V, Sargeant JM. Prioritizing zoonotic diseases: differences in perspectives between human and animal health professionals in North America. *Zoonoses Public Health* (2016) 63:196–211. doi:10.1111/zph.12220
34. Sawtooth Software. *The CBC/HB System for Hierarchical Bayes Estimation*. Sequim, WA: Sawtooth Software (2009).
35. Dattalo P. *Analysis of Multiple Dependent Variables*. New York, NY: Oxford University Press (2013).
36. Jansen J, Lam TJGM. The role of communication in improving udder health. *Vet Clin North Am Food Anim Pract* (2012) 28:363–79. doi:10.1016/j.cvfa.2012.03.003
37. Garforth C, McKemey K, Rehman T, Tranter R, Cooke R, Park J, et al. Farmers' attitudes towards techniques for improving oestrus detection in dairy herds in South West England. *Livest Sci* (2006) 103:158–68. doi:10.1016/j.livsci.2006.02.006
38. Schwendner AA, Cousin M-E, Bodmer M, Lam T, Schuepbach G, van den Borne B. Knowledge, attitudes and practices of Swiss farmers and veterinarians towards the usage of intramammary antimicrobials in dairy cows. In: Prodpecan O, editor. *Proceedings of the XV. Middle European Buiatric Congress*; 2015 June 10–13; Maribor, Slovenia (2015).
39. Huijps K, Hogeveen H, Antonides G, Valeeva NI, Lam TJGM, Oude Lansink AGJM. Sub-optimal economic behaviour with respect to mastitis management. *Eur Rev Agric Econ* (2010) 37:553–68. doi:10.1093/erae/jbq036
40. Kahneman D, Knetsch JL, Thaler RH. The endowment effect, loss aversion, and status quo bias. *J Econ Perspect* (1991) 5:193–206. doi:10.1257/jep.5.1.193
41. Santman-Berends IMGA, Swinkels JM, Lam TJGM, Keurentjes J, van Schaik G. Evaluation of udder health parameters and risk factors for clinical mastitis in Dutch dairy herds in the context of a restricted antimicrobial usage policy. *J Dairy Sci* (2016) 99:2930–9. doi:10.3168/jds.2015-10398
42. van den Borne BHP, Halasa T, van Schaik G, Hogeveen H, Nielen M. Bioeconomic modeling of lactational antimicrobial treatment of new bovine subclinical intramammary infections caused by contagious pathogens. *J Dairy Sci* (2010) 93:4034–44. doi:10.3168/jds.2009-3030
43. Barlow JW, Zadoks RN, Schukken YH. Effect of lactation therapy on *Staphylococcus aureus* transmission dynamics in two commercial dairy herds. *BMC Vet Res* (2013) 9:28. doi:10.1186/1746-6148-9-28

Conflict of Interest Statement: None of the authors of this paper have a financial or personal relationship with other people or organizations that could inappropriately influence or bias the content of the paper.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership