Preventive Veterinary Medicine 171 (2019) 104749

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

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

Machine-learning algorithms to identify key biosecurity practices and factors associated with breeding herds reporting PRRS outbreak

Gustavo S. Silva[a,1], Gustavo Machado[b,1], Kimberlee L. Baker[a], Derald J. Holtkamp[a], Daniel C.L. Linhares[a,*]

[a] Veterinary Diagnostic and Production Animal Medicine Department, College of Veterinary Medicine, Iowa State University, Ames, Iowa, United States
[b] Department of Population Health and Pathobiology, North Carolina State University, College of Veterinary Medicine, Raleigh, North Carolina, United States

ARTICLE INFO

Keywords:
Biosecurity practices and factors
PRRSv outbreaks
Risk index
Machine learning
Decision-making

ABSTRACT

Investments in biosecurity practices are made by producers to reduce the likelihood of introducing pathogens such as porcine reproductive and respiratory syndrome virus (PRRSv). The assessment of biosecurity practices in breeding herds is usually done through surveys. The objective of this study was to evaluate the use of machine-learning (ML) algorithms to identify key biosecurity practices and factors associated with breeding herds self-reporting (yes or no) a PRRS outbreak in the past 5 years. In addition, we explored the use of the positive predictive value (PPV) of these models as an indicator of risk for PRRSv introduction by comparing PPV and the frequency of PRRS outbreaks reported by the herds in the last 5 years. Data from a case control study that assessed biosecurity practices and factors using a survey in 84 breeding herds in U.S. from 14 production systems were used. Two methods were developed, method A identified 20 variables and accurately classified farms that had reported a PRRS outbreak in the previous 5 years 76% of the time. Method B identified six variables which 5 of these had already been selected by model A, although model B outperformed the former model with an accuracy of 80%. Selected variables were related to the frequency of risk events in the farm, swine density around the farm, farm characteristics, and operational connections to other farms. The PPVs for methods A and B were highly correlated to the frequency of PRRSv outbreaks reported by the farms in the last 5 years (Pearson r = 0.71 and 0.77, respectively). Our proposed methodology has the potential to facilitate producer's and veterinarian's decisions while enhancing biosecurity, benchmarking key biosecurity practices and factors, identifying sites at relatively higher risk of PRRSv introduction to better manage the risk of pathogen introduction.

1. Introduction

Biosecurity is the combination of all practices taken to reduce the risk of introduction and spread of pathogens (Amass and Clark, 1999; Food and Agriculture Organization, 2010). In other words, biosecurity practices intend to keep diseases out of the farm as well as to manage further spread of endemic diseases in the farm. The adoption of these practices is essential to prevent introduction of infectious diseases such as porcine reproductive and respiratory syndrome virus (PRRSv) in a herd. Despite the years of research since it was first reported (Keffaber, 1989), PRRSv continues to cause productivity losses worth $664 million annually and is considered a major health problem affecting the global swine industry (Holtkamp et al., 2013; Pileri and Mateu, 2016).

The evaluation of biosecurity is usually done by the assessment of highly correlated practices (Boklund et al., 2004). PRRS-specific biosecurity assessments take into account the most likely risk events for transmission of PRRSv between farms (e.g. animal movements, pickup and deliveries of supplies from or to farms, people movement, contact with other animals, air and water) (Otake et al., 2002a; Perez et al., 2015; Zimmerman et al., 2012) and are useful for understanding how PRRSv may be introduced into a herd (Holtkamp and Linhares, 2017; Silva et al., 2018).

Supervised machine-learning (ML) algorithms are broadly used in human and animal health, as they can handle both classification and regression operations. Outcome predictions are based on the consensus of hundreds of randomized decision trees built using a large number of explanatory variables (Breiman, 2001; Machado et al., 2019, 2015). Compared with other "conventional" methods (e.g. logistic regression models) (Tuerlinckx et al., 2006), ML algorithms have shown better performance (Fountain-Jones et al., 2019; Machado et al., 2019, 2015).

* Corresponding author.
E-mail address: linhares@iastate.edu (D.C.L. Linhares).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.prevetmed.2019.104749
Received 5 April 2019; Received in revised form 12 August 2019; Accepted 19 August 2019

Generalized linear models are not as effective as in capturing such non-linear responses and complex interactions, which is an important advantage of most machine-learning algorithms in capturing these non-linear responses and complex interactions (Elith et al., 2008; Tu, 1996).

Biosecurity survey data are often complex, correlated and extensive. Thus, the characteristics of ML make it an attractive approach to identify vulnerabilities in biosecurity practices and factors associated with risk of PRRSv introduction. Identifying and ranking the most relevant biosecurity practices and factors associated with PRRSv outbreaks will generate new knowledge about the impact of biosecurity on PRRSv transmission and allows for benchmarking of key biosecurity practices and factors within and between swine production systems over time.

The objective of this study was to evaluate the use of a set of ML algorithms to identify key biosecurity practices and factors associated with reported PRRS outbreaks in swine breeding herds. In addition, the plausibility of using the positive predictive value (PPV) of these models as an indicator of risk for PRRSv introduction was explored by assessing the association between the PPV and the frequency of PRRSv outbreaks reported by the herds in the last 5 years.

2. Material and methods

For the purpose of this study, "biosecurity practices and factors" refer to all practices taken to prevent or reduce the likelihood of PRRSv introduction into a swine breeding herd. Risk events were defined as events which may result in PRRSv introduction and they are described by category in Table 1. The risk events were established based on the common events that occur on swine breeding herds and data from the literature. In addition, biosecurity factors include herd characteristics and swine density nearby the farm that is not classified as a biosecurity practice per se.

2.1. Study design and data collection

The data used in this study originated from a case-control study designed to describe biosecurity practices in breeding herds with history of low incidence of PRRS outbreaks compared to those with high incidence (Silva et al., 2018).

The target population of the study were the breeding herds that were part of MorrisonSwine Health Monitoring Project (MSHMP). MSHMP is a swine disease monitoring program where U.S. swine producers share weekly disease status and represents a convenience sample of 930 sow herds from 25 production systems including approximately 2.7 million sows (Perez et al., 2019). Swine production system was defined as the set of farms owned, managed or both by the same company.

The source population of this study was herds reporting to MSHMP and the study sample was a convenient sample of 84 breeding herds from 14 swine production systems which agreed to share historical information of PRRS outbreaks and complete a biosecurity survey. The data were collected during the summer of 2017 by one of the authors (KLB) using an electronic file (Microsoft Excel) and personal interview when needed. The definition used to report a PRRSv outbreak was the identification of a PRRS wild type virus in the breeding herd and was self-reported by the production system.

The outcome of interest in this study was the presence (case) or absence (control) of at least one PRRSv outbreak in a period of five years self-reported by the farms. From the 84 breeding herds, 50 farms were classified as cases and 34 farms as controls respectively.

2.2. Biosecurity survey

The survey used to capture biosecurity practices and factors was developed through the PRRSv Outbreak Investigation Program (Canon et al., 2015) at Iowa State University, and collected information on herd characteristics, swine density, PRRS outbreak history (number of PRRSv outbreaks), frequency of risk events and bioexclusion practices associated with each risk event. The survey is described in Table A in Supplement File.

2.3. Machine learning algorithms

Supervised machine-learning (ML) algorithms: Random Forest (RF) (Breiman, 2001), Support Vector Machine (SVM) (Boser et al., 1992), and Gradient Boosting Machine (GBM) (Friedman, 2001) were used to explore the association between PRRS outbreaks and biosecurity practices and factors. In the context of this study, algorithms' capacities to classify outcomes are dependent on complexity, volume and internal structure of the data. Thus, we compared and selected the best algorithm based on higher i) accuracy, ii) sensitivity, and iii) specificity in classifying farms that had reported PRRS outbreak (Fig. 1).

First, the data were modeled with their natural imbalance nature (different proportion of case:control). Then, a down-sampling strategy

Table 1
Risk events that may result in PRRSv introduction into a breeding herd.

Categories of risk events	Risk events
Swine movements	I. Semen delivered to premises
	II. Breeding replacement animals delivered to premises
	III. Cull breeding animals hauled from premises
	IV. Weaned pigs hauled from premises
Pickup/Deliveries	I. Dead animals removed from premises
	II. Feed or feed ingredients delivered to premises
	III. Propane and fuel delivered to premises
	IV. Garbage collected from premises
	VI. Electrical meter read on premises
	VII. New tools and supplies are delivered to premises
	VIII. Tools and supplies transferred from other swine premises are delivered to premises
People movement	I. On-farm employees enter premises
	II. Repair, maintenance, electrical and plumbing personnel enter premises
	III. Veterinarians, off-site production managers, vendors and other visitors enter premises
Pork/food product entry	
Manure removal	
Wild animals, domestic animals and insects	I. Wild animals
	II. Domestic animals
	III. Insects
Air and water	I. Entry of air
	II. Entry of water

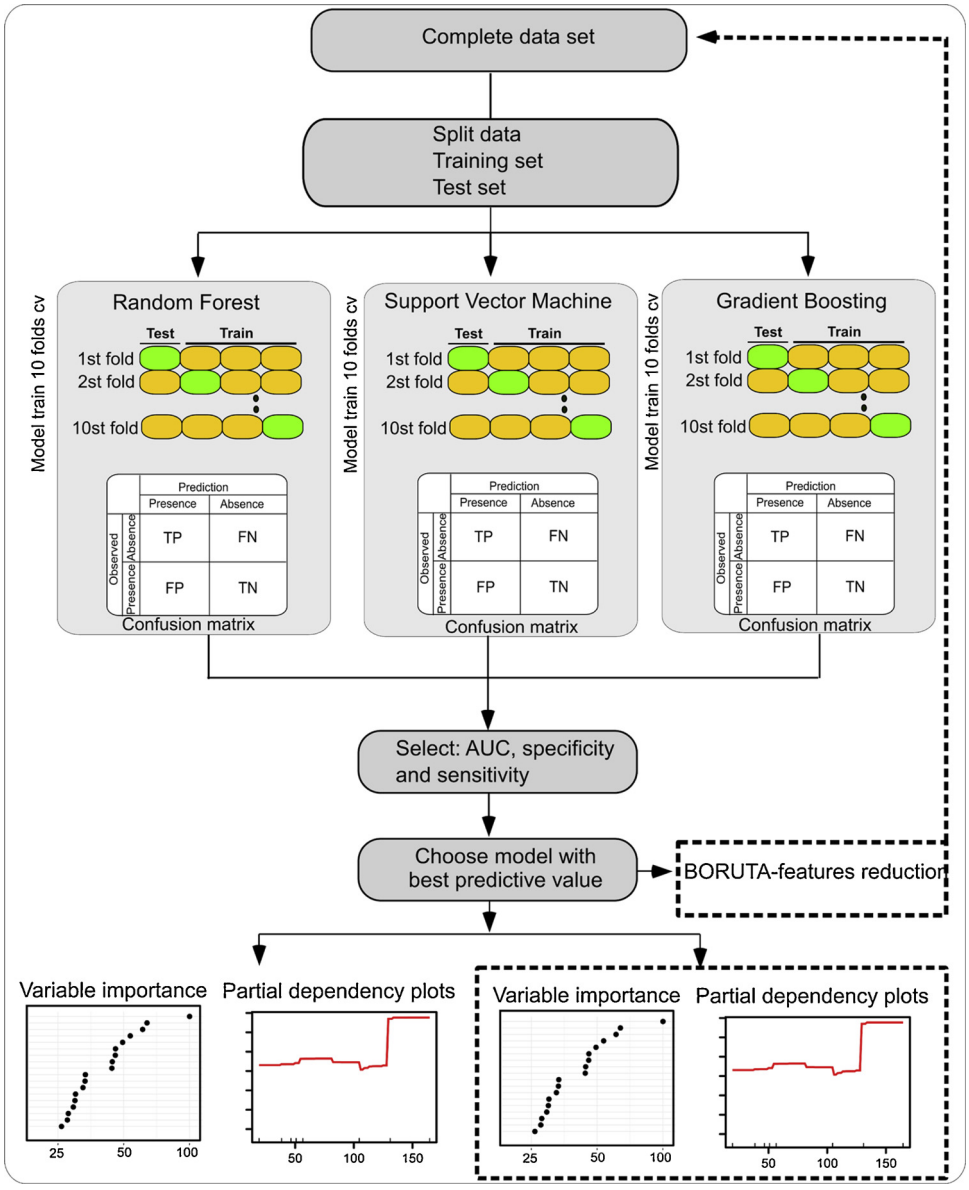

Fig. 1. Flow chart showing the steps in our machine-learning framework.

via "downSample" function in the R package *caret* (Max Kuhn Contributions from Jed Wing et al., 2018) was performed and compared with the imbalanced data given the strong likelihood that classifications trending towards the majority class (cases) would affect model performance (Haixiang et al., 2017). For the down-sampling strategy, the majority class was randomly down-sampled to match the frequency of the rarest class. Prior to down-sampling, the original data were randomly and uniformly divided into a training (80%) and an independent test set (20%).

The training set was used to train the ML algorithms via a k-fold cross-validation process and the independent test set was used for validation. All algorithms (RF, SVM and GBM) were trained (80% of data) using 260 variables related to biosecurity practices and factors. RF was performed with *randomForest* package (Liaw and Wiener, 2002), and SVM and GBM were performed with the *caret* package (Max Kuhn Contributions from Jed Wing et al., 2018). In the training steps, an internal repeated 10-fold cross-validation process to estimate model performance was implemented to prevent overfitting and artificial inflation of accuracy because training and testing data come from the same cross-sectional study. Fig. 1 describes the steps in our ML framework.

Model performance for each ML algorithm was assessed by calculating the training accuracy, specificity, and sensitivity based on the construction of a confusion matrix (Fig. 1) (Machado et al., 2019, 2015). The 20% of data set aside from the original dataset was used as an independent test set of observations. The independent test set of observations was fed into the selected ML algorithm, allowing the algorithm to predict the outcome for the new data. Accuracy was calculated as the overall proportion of observations correctly predicted. Specificity was calculated by dividing TN (true negative) by the sum of TN and FP (false positive). Sensitivity was calculated as TP (true positive) divided by the sum of TP and FN (false negative) for each ML algorithm. The threshold used to classify a farm as a case was a probability $\geq$50% (0.50).

2.4. Variable selection and ranking

To rank the importance of each variable the unscaled node impurity measured by the Gini index was calculated. Regardless of the algorithm, the variable importance score represents the relevance of each

Table 2
Classification performance using method A (p-value and Gini index) with the confusion matrix for the RF model trained with the complete set of predictor variables (n = 19), and Classification performance using method B (BORUTA) with the confusion matrix for the RF model trained with the complete set of predictor variables (n = 6).) both used average of ten repetitions of 10-fold cross-validation. Standard deviations are given in parenthesis*.

P-value and Gini index*	.	Observed	
.	.	Not reporting PRRS outbreak	Reporting PRRS outbreak
Method A-predicted	Not reporting PRRS outbreak	17.5 (1.08)	5.5 (1.64)
	Reporting PRRS outbreak	10.5 (1.08)	34.5 (1.64)
Method B -predicted	Not reporting PRRS outbreak	20.2 (0.63)	5.6 (1.42)
	Reporting PRRS outbreak	7.8 (0.63)	34.4 (1.42)

* Performance metrics method A: ACC: 76.47 (Sd. 2.49); SPE: 62.50 (Sd. 3.87) and SEN: 86.25 (Sd. 3.45) and method B ACC: 80.29 (Sd. 1.72); SPE: 72.14 (Sd. 2.25) and SEN: 86.00 (Sd. 3.57).

biosecurity practice and factor in predicting PRRSv outbreaks, with larger values representing more relevant predictors.

To reduce the number of predictors used, two approaches named method A and method B were generated. Method A used three parameters to variable selection: 1) Accuracy, calculated the accuracy when variables are permuted. 2) P-value calculated for each variable to evaluate whether it was used in the random forest more often than when variables were included in decision trees at random. P-values were based on a binomial distribution of the number of nodes split on the variable assuming that variables are randomly drawn (Ishwaran et al., 2010). 3) Node purity, where the mean decrease in the Gini index of node impurity (i.e. increase of node purity) by splits is calculated. Finally, only variables with a p-value ≤ 0.01, an increase of node purity > 0.04 points and with an increase in accuracy of ≥ 0.1, were considered.

Method B consisted of a Boruta algorithm (Kursa and Rudnicki, 2010), a more direct method to determine variable relevance by comparing the relative relevance of the real variables with that of random probes. The Boruta routine has been found to be the one of the most powerful approaches to select relevant features (Degenhardt, Seifert, & Szymczak, 2017) and does so by measuring the importance of each variable also using a tree-like algorithm (Kursa and Rudnicki, 2010). *Boruta* is an all relevant variable selection method, while most other approached are minimal optimal; this means it tries to find all variables carrying information usable for prediction, rather than finding a possibly compact subset of variables on which some classifier has a minimal error (Kursa and Rudnicki, 2010). Boruta was applied using function "Boruta" from the *boruta* package (Kursa and Rudnicki, 2010).

The influence of the most relevant variables was further analyzed via partial dependence plots. These provide insight on the marginal effect of each predictor on the likelihood of a PRRS outbreak while controlling for the effects of all other variables. The partial dependence of a variable's effect is best understood by visually examining general patterns in relation to the values of the predictor variable (Friedman, 2001). Because we modeled a binary classification, case or control, partial dependence values were reported on the "logit" scale (Cutler et al., 2007), where larger values indicate higher probability of being a case herd (Fig. 1).

2.5. Positive predictive value as a risk index

The positive predictive value (PPV) is the probability that a farm that was predicted to be a case, had reported at least one PRRS outbreak (McKenna and Dohoo, 2006) and estimated using methods A and B.

To evaluate if the PPV obtained by methods A and B could be used as a risk index to identify farms at higher risk based on the combination of key biosecurity practices and factors. The PPV for each farm (n = 84) was estimated, then the association between the PPV and the frequency of PRRS outbreaks (number of outbreaks) reported by each farm was assessed using the Pearson correlation coefficient. The correlation coefficient was estimated using the package *stats* (R Foundation for Statistical Computing, Vienna, 2011).

3. Results

3.1. Descriptive analysis

A total of 84 breeding herds from 14 production systems were investigated. Thirty-four (n = 34) herds did not report a PRRS outbreak in the previous 5 years, and 50 herds reported at least one PRRS outbreak. Negative farms had median herd size of 3,100 sows (range: 1,563.8–5,215.0), while positive farms 3,811 sows (range: 2,680.3–5,350.3).

3.2. Model selection and validation

The best-performing algorithm was selected by comparing model performance in the cross-validation step and ranking algorithms, RF without down-sample had the best performance with an accuracy of 76.4% (± 2.49), SVM with 69.11% (± 2.05), and GBM 70.14% (± 5.32). Sensitivity were 86.2% (± 3.45) for RF, 72.97% (± 8.91) for SVM, and 79% (± 5.42) GBM. Specificities were 62.5% (± 3.87) for RF, 68% (± 4.77) for SVM, and 57.50% (± 3.12) for GBM. Random forest was the best algorithm, consequently, was the ML algorithm used for all-further analysis. Table 2 describes the cross-validated confusion matrix for RF algorithm (Table 2). Thus, using the RF algorithm and applying the rules described in the method section for method A, 20 variables were selected based on accuracy improvement and the node purity. The calculated p-value for each variable's influence on PRRSv outbreak, according to accuracy and increase of node purity, is shown in Fig. 2.

Method B used six preselected variables for the final model, again RF algorithm performed best with an accuracy: 80%; sensitivity: 86%; specificity: 72%. Table 2 describes the cross-validated confusion matrix.

Analysis of the correlation matrix computed for variables selected in method A (Supplement Fig. A) suggested that a small number of variables, especially those related to the frequency of events were highly correlated. Correlation between variables used in method B was not significant (Supplement Fig. B).

3.3. Variable importance

The twenty variables were ranked according to their contribution to model prediction by the unscaled average decrease in node impurity (Fig. 3). Employees food entry, frequency of weans per month, frequency of rendering, number of employees at farm, annual employee turnover and frequency of cull sows hauled per month were the six most important variables for PRRSv classification for method A.

The variable importance matrix for method B is depicted in Fig. 4 (top). Ranked in order of importance is frequency of weans per month followed by number of finishing pigs within 3 miles radius. Subsequent rankings are as follows: PRRSv positive animals hauled on a trailer that

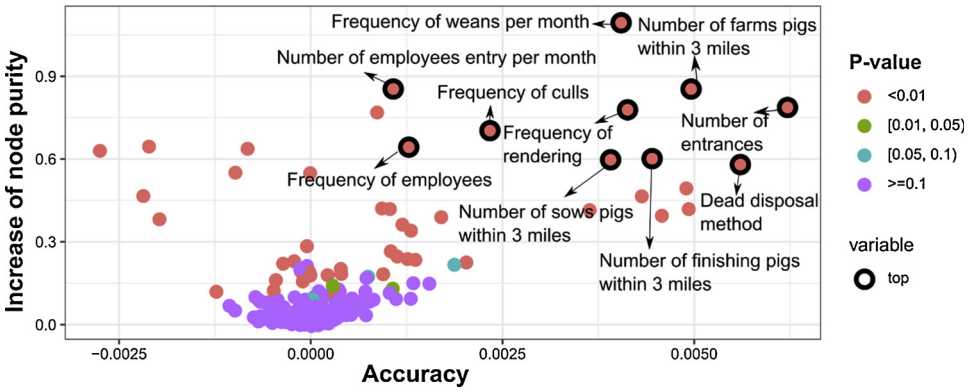

Fig. 2. Method A: Variables selection. Variable importance according to the Gini index and the accuracy metric. Variables in red represent ones that were more often used to split the nodes in the random forest. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

also hauls culled sows, type of dead disposal onsite, frequency of monthly dead animal removal and being a commercial herd.

For both variable selection methods, A and B, the key variables that were able to correctly classify the farms based on PRRS outbreak reports were related to: farm characteristics and requirements, monthly frequency of risk events, animal density around the farm and operational connections to other farms. A descriptive analysis by group for the variables identified in methods A and B can be found in Table B in the supplementary file.

3.4. Partial dependence plots

The effect of each variable on the probability of a PRRS outbreak are shown in partial dependence plots. Due to the larger number of variables retained by method A (n = 20), the full partial plots are described in Fig. C in Supplement File and for method B, the partial plots are described in Fig. 4 (bottom).

In summary, the probability of a herd being classified as having reported a PRRSv outbreak increased as frequency of risk events (e.g. frequency of rendering per month) and the swine density surrounding the farm (e.g. number of swine sites or pigs within 3 miles) increased (see Figs. 4 and C in Supplement File). Likewise, the risk of reporting a PRRSv outbreak increased when the farm did not have an onsite disposal method for dead animals, or when the farm shared trailers with other farms to transport breeding replacement animals or culled sows (see Figs. 4 and C in Supplement File). However, the risk of reporting a PRRS outbreak decreased with the increase in the frequency of weaning pigs per month, with the number of employees working daily in the farm, and with the increase in the downtime period requirement for visitors (see Fig. C in Supplement File).

3.5. PPV as risk index of PRRSv outbreak

The PPV reflects the herd's likelihood of being classified as reporting a PRRS outbreak. The PPV for methods A and B showed to be strongly correlated with the frequency of PRRS outbreak. Comparing both methods, PPV from method B had higher association, Pearson correlation coefficient of 0.77 (p-value < 0.01), compared to method A (Pearson r = 0.71, p-value < 0.01). Fig. 5 describes the relationship between the PPV and number of outbreaks for each farm. The results suggested that higher the PPV higher the number of outbreaks reported by the farm.

4. Discussion

In this study, we proposed two variable selection strategies (A and B). A large number of predictors were evaluated with the objective of identifying key biosecurity practices and factors able to correctly classify farms that had PRRSv outbreak. In comparing the performance of methods A and B with the full model that used all 260 variables (accuracy of 65.0%), both methods outperformed the full model using fewer variables. Method A had an accuracy of 76.3% using 20 variables, and method B was even better with an accuracy of 80% using only 6 variables. In addition, both models had better sensitivity than the full model. The performance of ML methods, as with other statistical methods, can be influenced by imbalanced proportions of the outcome of interest, this is widely overlooked particularly in veterinary epidemiology (Haibo He and Garcia, 2009; Liaw and Wiener, 2002; Machado et al., 2019). To address this issue, we used a down-sampling strategy and compared the results with those from the model where the imbalance was ignored. The results suggested that the imbalance did not

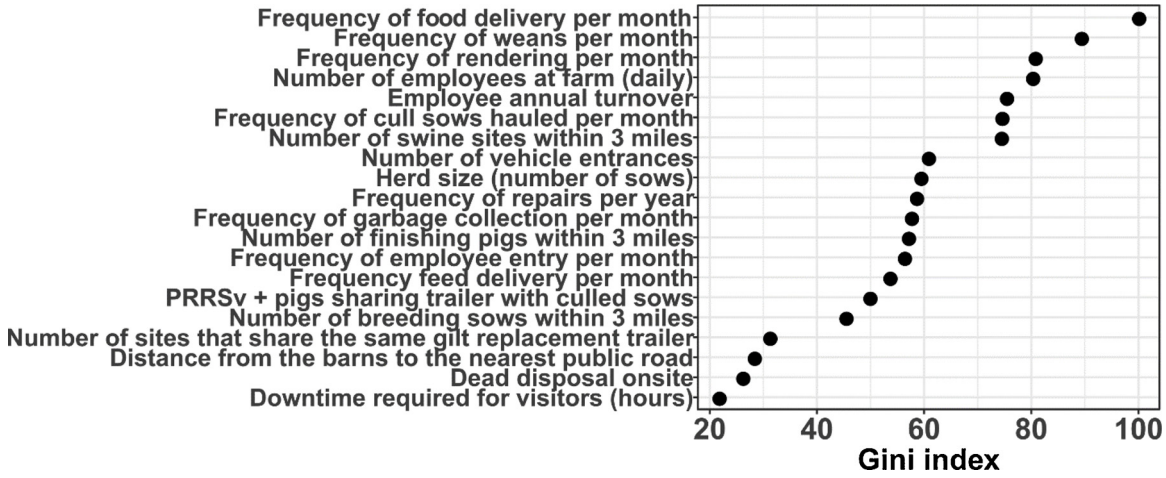

Fig. 3. Variable importance analysis performed by random forest using method A (p-value model and Gini index). The set of 20 variables used for classification are ordered by their importance as estimated by the RF model.

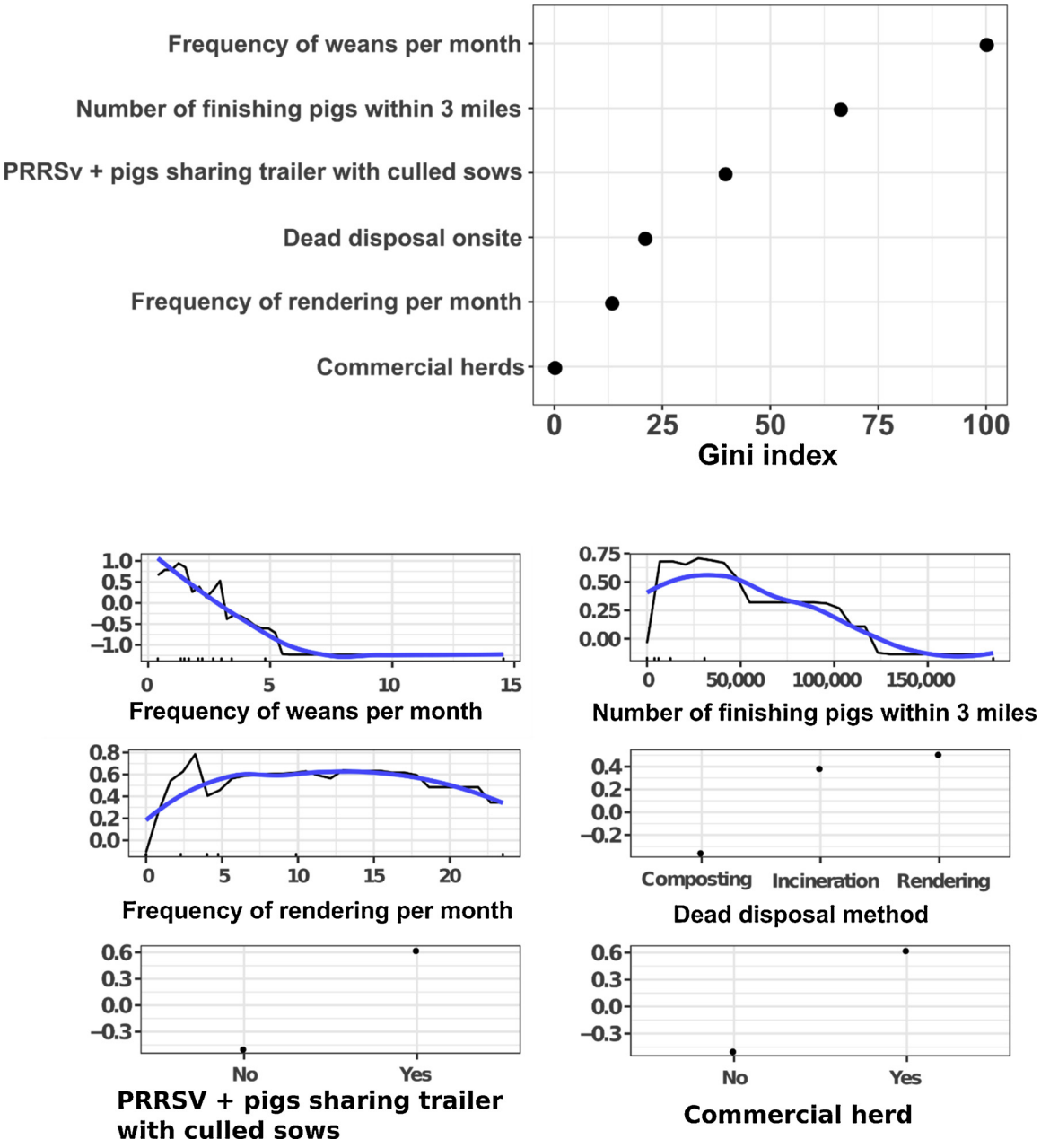

Fig. 4. Random forest variable importance and partial dependence plots using method B (BORUTA-features reduction). Upper - Biosecurity aspects in order of importance. Bottom - Partial dependence plot indicating marginal effect for PRSS prediction.

impact on model performance (results not shown).

The ranking of the final variables (in descending order of importance) (Fig. 3 and 4) suggests that the main variables involved in PRRS classification may be grouped into four categories: frequency of risk events in the farm, swine density around the farm, farm characteristics, and restrictions on visitors/operational connections to other sites. The identified association between variables and PRRSv outbreak has been previously reported elsewhere (Desrosiers, 2011; Pileri and Mateu, 2016)

Among the variables identified in this study, the frequency of risk events was the most relevant. The frequency of food entry was the most important factor and was highly correlated with the number of employees. Other events that increased the risk were the frequency of rendering, frequency of garbage collection, frequency of feed delivery, frequency of repairs and frequency of cull sows removal. Previous research has indicated that trucks, trailers, animals products, feed, fomites, contaminated equipment and personnel have the potential to increase PRRSv spread (Dee et al., 2004; Dewey et al., 2014; Magar and Larochelle, 2004; Otake et al., 2002b; Pitkin et al., 2009). Even though some of these events may not represent high risk, the probability of pathogen introduction increases as the frequency of risk events increases (Romagosa, 2017).

On the other hand, the risk was lower in farms with a high frequency of employee entry. This finding was also previously reported in other studies (Boklund et al., 2004; Postma et al., 2015; Ribbens et al., 2008; Van Steenwinkel et al., 2011), with a possible explanation being that farms with high number of employees tend to have greater concerns about biosecurity training and compliance. In addition, herds with a higher frequency of weaning pigs per month had a lower risk of having the disease and the same hypothesis can be speculated here.

From the 22 variables selected by both methods, four were related to herd size and swine density around the farm. Our results suggest that

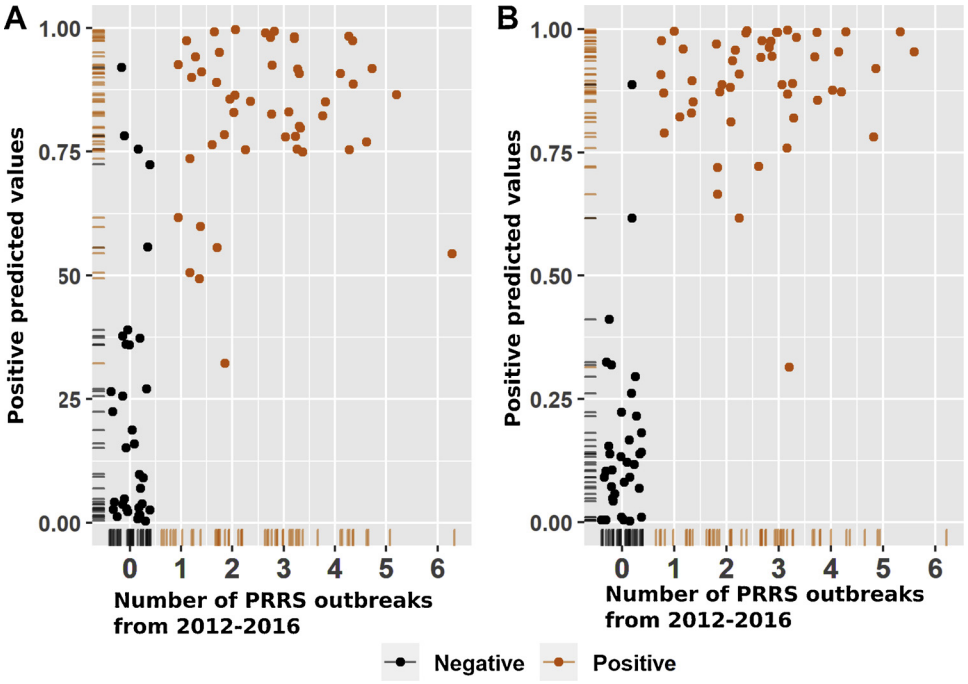

Fig. 5. Scatter plot describing the relationship between the PPV and the number of outbreaks reported for each farm. Note that some farms reported more than one outbreak during 2012–2016 and this number was used to assess this relationship. Black dots represent the farms that had not had reported an outbreak (negative farms) and orange dots represent the farms that reported a PRRSv outbreak. A – PPV obtained using method A (p-value and Gini index) and B - PPV obtained using method B (Boruta).

the chances of a farm being classified as reporting PRRSv outbreak increases linearly with an increase in herd size and swine density around the farm (Figs. 4 and C Supplement File). We believe that higher the herd size, more likely to have operational connections to other sites, which may increase the likelihood of reporting a PRRSv outbreak.

The indirect transmission of PRRSv via air is well known (Dee et al., 2009; Mortensen et al., 2002; Velasova et al., 2012). Given the importance of airborne transmission, the swine industry has adopted filtration of air in breeding herds which has been shown to reduce the number of PRRSv outbreaks (Alonso et al., 2013). Despite this trend, the use of air filtration was not ranked among the top predictors in our study and this may be due to the small percentage of farms that reported air filtration between groups (11 farms in case and 11 farms in control group).

Farms that had offsite dead disposal were also associated with higher risk. Velasova et al. (2012) described that farms that had dead pigs collected had higher odds (OR = 5.6, 95%CI: 1.7–18.3)1.7–18.3) of active PRRS infection compared to farms that had onsite dead disposal. Onsite disposal eliminates the need for rendering trucks, a potential risk event, since they visit multiple sites and carry potentially contaminated carcasses. In addition, our results suggested that the greater the distance of the farms from public roads the lower the risk (Fig. C Supplement File). Farms close to public roads may be more susceptible to airborne transmission/infection because these are used to transporting animals from different sources and health status.

The findings related to requirements for visitors and operational connections between sites showed that herds which share trailers were more likely to be classified as reporting PRRSv outbreak, reinforcing the importance of transport biosecurity practices. Previous studies have demonstrated that pigs may be infected with PRRSv through contact with contaminated vehicles and that a contaminated trailer can serve as a link between different sites and act as a source of infection (Bottoms et al., 2012; Dee et al., 2004; Pitkin et al., 2009). Our results showed that the risk of reporting a PRRS outbreak increases in herds in which PRRSv-positive pigs share the trailer with culled sows. In addition, there was a linear relationship between the number of sites that shared the same gilt replacement trailer and the probability of reporting PRRS outbreak. Monthly pig movements among herds not only facilitate PRRSv spread but also regularly introduce groups of susceptible animals (nursery pigs and gilts). Given the connectivity of the swine industry and the need to share supplies, effective protocols to clean and disinfect are necessary to reduce the risk of disease transmission by transport (Dee and Deen, 2006). Another well-established biosecurity practice identified as risk factor was the downtime required for visitors (Amass and Clark, 1999; Pitkin et al., 2011). Our findings also suggest that the longer the downtime, the lower the risk.

We identified that PPV would be a good candidate to serve as an index to evaluate the risk of PRRSv outbreak. Here, the PPV represents the combination of the selected biosecurity practices and factors and the higher the PPV, the greater the frequency of PRRSv outbreaks in the past 5 years. In this study, we identified that such an approach has the potential to become an important tool to be directly used by producers and field veterinarians. Importantly, our approach uses a small number of questions in order to avoid collecting excessive information. This may increase the success rate of obtaining biosecurity information from producers with the goal of benchmarking biosecurity practices and factors over time (Boynton and Greenhalgh, 2004; Dean, 2015).

Importantly, benchmarking key biosecurity practices and factors would facilitate the identification of common differences between farms that have had outbreaks and would similarly identify common practices and factors of those farms that keep themselves free of PRRSv over time. Moreover, production systems may use this tool to identify herds at higher risk of PRRSv introduction in order to prioritize their investments.

One point that must be emphasized is that benchmarking a smaller set of variables does not negate the importance and the need of the other biosecurity practices. Due to the lack of variability, some practices were classified as unimportant and were judged by the model as not contributing to the correct classification of the farms. This can be explained by the fact that most of the farms usually apply a large number of biosecurity practices, leaving no variability within farms. Because of that, the variables selected can be seen as a *proxy* of the biosecurity practices related to each event. This suggests that producers should paid attention to these events, given that their frequency of occurrence was associated with the occurrence of PRRSv outbreaks.

The limitations of this study include the possibility of information bias with the use of survey and interviews, even if conducted directly with the producers (Dean, 2015). The use of convenience sampling

restricts the scope of the study and the extrapolation of the results should be done with caution. Another limiting factor of the study was the sample size. The relatively small number of observations and the low variability of some variables may have affected variable selection expected to be relevant. Lastly, the use of a data-driven model did not account for biological plausibility between the outcome and the independent variables; this only identified the variables most important for the classification and may be subject to finding spurious associations. However, because our interest was to identify and reduce the number of predictors, we believe that we have achieved the goal proposed by the study.

5. Conclusion

The results support the concept that there is correlation in biosecurity practices and factors of swine breeding herds. Our findings indicate that it is possible to identify the most relevant biosecurity practices and factors by 'asking' fewer questions, and them to predict PRRSv risk of outbreak. Methods A and B have classified farms' PRRS status as positive when diseases occurred with an accuracy of accuracy of 76% and 80%, respectively. With this study we developed the initial concept for a future tool with the capacity to help producers and veterinarians to measure and benchmark key biosecurity practices and factors more frequently, identify sites at relatively higher risk of PRRSv introduction to better manage the risk of virus introduction over time.

Declaration of Competing Interest

The authors declare that there are no conflict of interests.

Acknowledgements

The authors have received funding from Veterinary Diagnostic and Production Animal Medicine (VDPAM-ISU), Zoetis and Swine Health Information Center (SHIC). In addition, was partially funded by the Department of Population Health and Pathobiology, College of Veterinary Medicine, North Carolina State University, which provided startup funds for GM.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.prevetmed.2019.104749.

References

Alonso, C., Murtaugh, M.P., Dee, S.A., Davies, P.R., 2013. Epidemiological study of air filtration systems for preventing PRRSV infection in large sow herds. Prev. Vet. Med. 112, 109–117. https://doi.org/10.1016/j.prevetmed.2013.06.001.

Amass, S.F., Clark, L.K., 1999. Biosecurity considerations for pork production units. Swine Heal. Prod 7, 217–228.

Boklund, A., Alban, L., Mortensen, S., Houe, H., 2004. Biosecurity in 116 Danish fattening swineherds: descriptive results and factor analysis. Prev. Vet. Med. 66, 49–62. https://doi.org/10.1016/j.prevetmed.2004.08.004.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT' 92. ACM Press, New York, New York, USA. pp. 144–152. https://doi.org/10.1145/130385.130401.

Bottoms, K., Poljak, Z., Dewey, C., Deardon, R., Holtkamp, D., Friendship, R., 2012. Investigation of strategies for the introduction and transportation of replacement gilts on southern Ontario sow farms. BMC Vet. Res. 8, 217. https://doi.org/10.1186/1746-6148-8-217.

Boynton, P.M., Greenhalgh, T., 2004. Selecting, designing, and developing your questionnaire. BMJ 328, 1312–1315. https://doi.org/10.1136/bmj.328.7451.1312.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Canon, A., Gerardy, K., Karriker, L., Holtkamp, D., 2015. outbreak investigation program – a systematic approach to PRRS outbreak investigations. In: Proc. 2015 North American PRRS Symposium. Chicago, IL.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88, 2783–2792. https://doi.org/10.1890/07-0539.1.

Dean, R.S., 2015. The use and abuse of questionnaires in veterinary medicine. Equine Vet. J. 47, 379–380. https://doi.org/10.1111/evj.12429.

Dee, S., Deen, J., Burns, D., Douthit, G., Pijoan, C., 2004. An assessment of sanitation protocols for commercial transport vehicles contaminated with porcine reproductive and respiratory syndrome virus. Can. J. Vet. Res. 68, 208–214.

Dee, S., Otake, S., Oliveira, S., Deen, J., 2009. Evidence of long distance airborne transport of porcine reproductive and respiratory syndrome virus and Mycoplasma hyopneumoniae. Vet. Res. 40, 39. https://doi.org/10.1051/vetres/2009022.

Dee, S.A., Deen, J., 2006. Evaluation of an industry-based sanitation protocol for transport vehicles contaminated with porcine reproductive and respiratory syndrome virus. J. Swine Heal. Prod. 14, 126–132.

Desrosiers, R., 2011. Transmission of swine pathogens: different means, different needs. Anim. Heal. Res. Rev. 12, 1–13. https://doi.org/10.1017/S1466252310000204.

Dewey, C., Bottoms, K., Carter, N., Richardson, K., 2014. A qualitative study to identify potential biosecurity risks associated with feed delivery. J Swine Health Prod.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Food and Agriculture Organization, 2010. Good Practices for Biosecurity in the Pig Sector - Issues and Options in Developing and Transition countries., FAO Animal Production and Health Paper No. 169. Food and Agriculture Organization of the United Nations/World Organisation for Animal Health/World Bank.

Fountain-Jones, N.M., Machado, G., Carver, S., Packer, C., Recamonde-Mendoza, M., Craft, M.E., 2019. How to make more from exposure data? An integrated machine learning pipeline to predict pathogen exposure. bioRxiv 569012. https://doi.org/10.1101/569012.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Source: The Annals of Statistics.

Haibo He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284. https://doi.org/10.1109/TKDE.2008.239.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: review of methods and applications. Expert Syst. Appl. 73, 220–239. https://doi.org/10.1016/J.ESWA.2016.12.035.

Holtkamp, D., Linhares, D., 2017. Understanding How Pathogens Infect Pig Farms, in: 48th AASV Annual Meeting. American Association of Swine Veterinarians, Denver, CO, US.

Holtkamp, D.J., Kliebenstein, J.B., Neumann, E.J., Zimmerman, J.J., Rotto, H.F., Yoder, T.K., Wang, C., Yeske, P.E., Mowrer, C.L., Haley, Ca., 2013. Assessment of the economic impact of porcine reproductive and respiratory syndrome virus on United States pork producers. J. Swine Heal. Prod. 21, 72–84.

Ishwaran, H., Kogalur, U.B., Gorodeski, E.Z., Minn, A.J., Lauer, M.S., 2010. High-dimensional variable selection for survival data. J. Am. Stat. Assoc. 105, 205–217. https://doi.org/10.1198/jasa.2009.tm08622.

Keffaber, K.K., 1989. Reproductive failure of unknown etiology. Am. Assoc. Swine Pract. Newsl. 1 (2), 1–10 1.

Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the **boruta** package. J. Stat. Softw. 36, 1–13. https://doi.org/10.18637/jss.v036.i11.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.

Machado, G., Mendoza, M.R., Corbellini, L.G., 2015. What variables are important in predicting bovine viral diarrhea virus? A random forest approach. Vet. Res. 46, 85. https://doi.org/10.1186/s13567-015-0219-7.

Machado, G., Vilalta, C., Recamonde-Mendoza, M., Corzo, C., Torremorell, M., Perez, A., VanderWaal, K., 2019. Identifying outbreaks of Porcine Epidemic Diarrhea virus through animal movements and spatial neighborhoods. Sci. Rep. 9, 457. https://doi.org/10.1038/s41598-018-36934-8.

Magar, R., Larochelle, R., 2004. Evaluation of the presence of porcine reproductive and respiratory syndrome virus in pig meat and experimental transmission following oral exposure. Can. J. Vet. Res. 68, 259–266.

Max Kuhn Contributions from Jed Wing, A., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Core Team the R, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T., Max Kuhn, M., 2018. Package "caret" Title Classification and Regression Training Description Misc Functions for Training and Plotting Classification and Regression Models.

McKenna, S.L.B., Dohoo, I.R., 2006. Using and interpreting diagnostic tests. Vet. Clin. North Am. Food Anim. Pract. 22, 195–205. https://doi.org/10.1016/j.cvfa.2005.12.006.

Mortensen, S., Stryhn, H., Søgaard, R., Boklund, A., Stärk, K.D., Christensen, J., Willeberg, P., 2002. Risk factors for infection of sow herds with porcine reproductive and respiratory syndrome (PRRS) virus. Prev. Vet. Med. 53, 83–101. https://doi.org/10.1016/S0167-5877(01)00260-4.

Otake, S., Dee, S., Pijoan, C., 2002a. Transmission of PRRSV: recent research reports. Int. Pigletter 22 (37), 40–42.

Otake, S., Dee, S.A., Rossow, K.D., Deen, J., Soo Joo, H., Molitor, T.W., Pijoan, C., 2002b. Transmission of porcine reproductive and respiratory syndrome virus by fomites (boots and coveralls). J. Swine Health Prod.

Perez, A.M., Davies, P.R., Goodell, C.K., Holtkamp, D.J., Mondaca-Fernández, E., Poljak, Z., Tousignant, S.J., Valdes-Donoso, P., Zimmerman, J.J., Morrison, R.B., 2015. 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. 246, 1304–1317. https://doi.org/10.2460/javma.246.12.1304.

Perez, A.M., Linhares, D.C.L., Arruda, A.G., VanderWaal, K., Machado, G., Vilalta, C., Sanhueza, J.M., Torrison, J., Torremorell, M., Corzo, C.A., 2019. Individual or common good? Voluntary data sharing to inform disease surveillance systems in food animals. Front. Vet. Sci. 6, 194. https://doi.org/10.3389/fvets.2019.00194.

Pileri, E., Mateu, E., 2016. Review on the transmission porcine reproductive and respiratory syndrome virus between pigs and farms and impact on vaccination. Vet. Res. 47, 1–13. https://doi.org/10.1186/s13567-016-0391-4.

Pitkin, A., Deen, J., Dee, S., 2009. Further assessment of fomites and personnel as vehicles for the mechanical transport and transmission of porcine reproductive and respiratory syndrome virus. Can. J. Vet. Res. 73, 298–302.

Pitkin, A., Otake, S., Dee, S., 2011. A one-night downtime period prevents the spread of porcine reproductive and respiratory syndrome virus and Mycoplasma hyopneumoniae by personnel and fomites (boots and coveralls). J. Swine Heal. Prod. 19, 345–348.

Postma, M., Backhans, A., Collineau, L., Loesken, S., Sjolund, M., Belloc, C., Emanuelson, U., Grosse Beilage, E., Stark, K.D.C., Dewulf, J., 2015. The biosecurity status and its associations with production and management characteristics in farrow-to-finish pig herds. Animal 1–12. https://doi.org/10.1017/S1751731115002487.

R Foundation for Statistical Computing, Vienna, 2011. R development core team. R a lang. A.I. 3-900051-07-0. Environ. Stat. Comput. 55, 275–286.

Ribbens, S., Dewulf, J., Koenen, F., Mintiens, K., De Sadeleer, L., de Kruif, A., Maes, D., 2008. A survey on biosecurity and management practices in Belgian pig herds. Prev. Vet. Med. 83, 228–241. https://doi.org/10.1016/j.prevetmed.2007.07.009.

Romagosa, A., 2017. Applied review on evidence-based biosecurity. 48th Annual Meeting of the American Association of Swine Veterinarians 5–11.

Silva, G.S., Corbellini, L.G., Linhares, D.L.C., Baker, K.L., Holtkamp, D.J., 2018. Development and validation of a scoring system to assess the relative vulnerability of swine breeding herds to the introduction of PRRS virus. Prev. Vet. Med. https://doi.org/10.1016/J.PREVETMED.2018.10.004.

Tu, J.V., 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J. Clin. Epidemiol. 49, 1225–1231.

Tuerlinckx, F., Rijmen, F., Verbeke, G., De Boeck, P., 2006. Statistical inference in generalized linear mixed models: a review. Br. J. Math. Stat. Psychol. 59, 225–255. https://doi.org/10.1348/000711005X79857.

Van Steenwinkel, S., Ribbens, S., Ducheyne, E., Goossens, E., Dewulf, J., 2011. Assessing biosecurity practices, movements and densities of poultry sites across Belgium, resulting in different farm risk-groups for infectious disease introduction and spread. Prev. Vet. Med. 98, 259–270. https://doi.org/10.1016/j.prevetmed.2010.12.004.

Velasova, M., Alarcon, P., Williamson, S., Wieland, B., 2012. Risk factors for porcine reproductive and respiratory syndrome virus infection and resulting challenges for effective disease surveillance. BMC Vet. Res. 8, 184. https://doi.org/10.1186/1746-6148-8-184.

Zimmerman, J.J., Benfield, D., Dee, S.A., Murtaugh, M.P., Stadejek, T., Stevenson, G.W., Torremorell, M., 2012. porcine reproductive and respiratory syndrome virus (porcine arterivirus). In: Zimmerman, J.J., Karriker, L.A., Ramirez, A., Schwartz, K.J., Stevenson, G.W. (Eds.), Diseases of Swine. Wiley-Blackwell, pp. 461–486.