Construction of Disease Risk Scoring Systems using Logistic Group Lasso: Application to Porcine Reproductive and Respiratory Syndrome Survey Data

Journal:	*Journal of Applied Statistics*
Manuscript ID:	CJAS-2012-0354.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	14-Nov-2012
Complete List of Authors:	Lin, Hui; Iowa State University, Department of Statistics Wang, Chong; Iowa State University, Department of Statistics Liu, Peng; Iowa State University, Department of Statistics Holtkamp, Derald; Iowa State University, Department of Veterinary Diagnostic and Production Animal Medicine
Keywords:	group lasso, multivariate logistic regression, index scoring system, PRRS, survey data, receiver operating characteristic curve, area under the curve

Journal of Applied Statistics
Vol. 00, No. 00, Month 200x, 1–23

RESEARCH ARTICLE

Construction of Disease Risk Scoring Systems using Logistic Group Lasso: Application to Porcine Reproductive and Respiratory Syndrome Survey Data

Hui Lin[a], Chong Wang[ab]*, Peng Liu[a] and Derald J. Holtkamp[b]

[a] *Department of Statistics, College of Liberal Arts and Sciences, Iowa State University, Ames, IA 50011, USA* ; [b] *Department of Veterinary Diagnostic and Production Animal Medicine, College of Veterinary Medicine, Iowa State University, Ames, IA 50011, USA*

(*Received 00 Month 200x; in final form 00 Month 200x*)

We propose to utilize the group lasso algorithm for logistic regression to construct a risk scoring system for predicting disease in swine. This work is motivated by the need to develop a risk scoring system from survey data on risk factor for porcine reproductive and respiratory syndrome (PRRS), which is a major health, production and financial problem for swine producers in nearly every country. Group lasso provides an attractive solution to this research question because of its ability to achieve group variable selection and stabilize parameter estimates at the same time. We propose to choose the penalty parameter for group lasso through leave-one-out cross validation, using the criterion of the area under the receiver operating characteristic curve. Survey data for 896 swine breeding herd sites in the United States and Canada completed between March 2005 and March 2009 is used to construct the risk scoring system for predicting PRRS outbreaks in swine. We show that our scoring system for PRRS significantly improves the current scoring system that is based on expert opinion. We also show our proposed scoring system is superior in terms of area under the curve to that developed by using multivariate logistic regression model selected based on variable significance.

Keywords: area under the curve; group lasso; multivariate logistic regression; PRRS; receiver operating characteristic curve; risk scoring system; survey data

*Corresponding author. Email: chwang@iastate.edu

ISSN: 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/0266476YYxxxxxxxx
http://www.tandfonline.com

1. Introduction

Risk scoring systems for predicting disease are widely used in medicine. Such scoring systems are usually derived from multivariate logistic regression models with disease status as the response variable. Typical approaches in the literature select potential explanatory variables (risk factors) based on variable significance, with risk scores of selected variables assigned based on estimated regression coefficients [2, 10, 13]. However, when the number of potential explanatory variables is large, such approaches may fail to produce a risk scoring system with the greatest power for predicting disease.

This paper is motivated by the need to develop a risk scoring system for porcine reproductive and respiratory syndrome (PRRS) based on survey data. PRRS, caused by the PRRS virus, is a major health, production and financial problem for swine producers in nearly every country. PRRS costs the United States swine industry around $560 million annually [8]. PRRS outbreaks in China caused pork prices to increase by 85 percent in 2006 [5]. For breeding herds, costs of clinical outbreaks of PRRS result from lost production due to abortion, mummies, stillborns, pre-wean mortality and sow deaths and from increased costs for treatment and control. Performance of observational studies to better understand the relative importance of risk factors for PRRS outbreaks have been limited by the availability of good data on a large set of farms over a relatively long period of time.

In human medicine, large datasets of information on risk factors, prevalence, incidence and clinical outcomes of disease are common. In veterinary medicine, until recently, there have been no parallel efforts to create epidemiological databases on a similar scale. The American Association of Swine Veterinarians (AASV) Production Animal Disease Risk Assessment Program (PADRAP) is a program through which a set of web-based risk assessment surveys are delivered(please visit: http://vdpambi.vdl.iastate.edu/padrap/default.aspx). It is used by veterinarians who are members of the AASV. Each of the surveys consists of a set of questions about potential risk factors for clinical outbreaks of PRRS in swine. Each question may have up to 6 possible responses. Members of the AASV use PADRAP to help producers systematically assess biosecurity factors that may be associated with clinical outcomes. As assessments are performed by veterinarians, they are added to the database of completed assessments.

Version 2 of the PRRS Risk Assessment for the Breeding Herd survey was introduced in 2005. The survey instrument was developed using expert opinion with the aid of the PRRS Risk Assessment Working Group composed of 21 veterinarians and researchers with expertise in PRRS. Initial estimates of the risk scores associated with each response were based on the consensus of expert opinion and equal weight is assigned to each question.

The aim of this study is to use the survey data that has been collected to develop a risk scoring system with 127 survey questions (categorical explanatory variables) that outperforms the current risk scoring system based on expert opinion. Multivariate logistic regression has been used in similar studies [add the previous 1, 2, 9 references] with variables selected by significance (usually at 0.05 level) and scores assigned based on estimated regression coefficients. However, " Quasi-complete-separation" may result when there are a large number of explanatory variables which makes estimation of the coefficients unstable [1]. To stabilize the estimation of parameter coefficients, one popular approach is the lasso algorithm with l_1-norm penalty proposed by Tibshirani [12]. Because the lasso algorithm can estimate some variable coefficients to be 0, it can also be used as a variable selection tool. For models with categorical survey questions (explanatory variables), however, original lasso algorithm only selects individual dummy variables instead of sets of the dummy variables grouped by question in the survey. Another disadvantage of applying lasso to grouped variables is that the estimates are affected by the way dummy variables are encoded. Thus the group lasso [16] method has been proposed to enable variable selection in linear regression models on groups of variables, instead of on single variables. For logistic regression models, the group lasso algorithm was first studied by Kim et al. [4]. They proposed a gradient descent algorithm to solve the corresponding constrained problem, which does, however, depend on unknown constants. Meier et al. [7] proposed a new algorithm that could work directly on the penalized problem and its convergence property does not depend on unknown constants. The algorithm is especially suitable for high-dimensional problems. It can also be applied to solve the corresponding convex optimization problem in generalized linear models. The logistic group lasso involves selection of a penalty (tuning) parameter λ which can be determined by cross-validation. The group lasso estimator proposed by Meier et al. [7] for logistic regression has been shown to be statistically consistent, even with large number of categorical predictors.

In this paper, we propose to use the logistic group lasso algorithm to construct risk scoring systems for predicting clinical PRRS outbreaks in swine herds. The paper is organized as follows. In Section 2, we introduce the multivariate logistic regression and the group lasso method for logistic regression to construct risk scoring systems for clinical PRRS outbreaks and we propose to use the second one. The penalty parameter λ for group lasso is selected through leave-one-out cross validation, using the criterion of the area under the receiver operating characteristic curve . In Section 3, we discuss the application to the PRRS survey data from 896 swine breeding herd sites in the United States and Canada. We show that our scoring system for PRRS is superior to both the current scoring system based on expert opinion and that developed by using logistic regression with model selection based on variable significance. Section 4 presents a simulation study to evaluate the performance of the multivariate logistic regression and the logistic group lasso method. We conclude with some discussion in Section 5.

2. Models for risk scoring systems

Consider risk scoring system construction using a sample of n observations, with information collected for G categorical predictors and one binary response variable for each observation. Let $\mathbf{x}_{i,g}$ be the vector of dummy variables associated with the gth categorical predictor for the ith observation, where $i = 1, \cdots, n$, $g =$

$1, \cdots, G$. Let y_i (= 1, diseased; or 0, not diseased) be the binary response for the ith observation. Denote the degrees of freedom of the gth predictor by df_g, which is also the length of vector $\mathbf{x}_{i,g}$.

2.1 *Multivariate logistic regression model*

Multivariate logistic regression has been used to construct risk scoring systems for predicting disease [2, 10, 13]. Denote the probability of disease for ith subject by θ_i, the model can be formulated as

$$y_i \sim Bounoulli(\theta_i), \tag{1}$$

with

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \eta_{\boldsymbol{\beta}}(x_i) = \beta_0 + \sum_{g=1}^{G} \mathbf{x}_{i,g}^T \boldsymbol{\beta}_g, \tag{2}$$

where β_0 is the intercept and $\boldsymbol{\beta}_g$ is the parameter vector corresponding to the gth predictor.

Construction of risk scoring systems using logistic regression usually consists of two steps: selection among the G risk factors, and estimation of parameters for the selected factors. For model selection, statistical significance has been used as a criterion for inclusion and exclusion of risk factors [2, 10, 13]. Some researchers use univariate logistic regression to screen factors by significance before putting them into a multivariate logistic regression model [2, 10], whereas others [13] don't. Traditional estimation of logistic parameters $\boldsymbol{\beta} = (\beta_0^T, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, ..., \boldsymbol{\beta}_G^T)^T$ is done through maximizing the log-likelihood

$$\begin{aligned} l(\boldsymbol{\beta}) &= log[\prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i}] \\ &= \sum_{i=1}^{n}\{y_i log(\theta_i) + (1-y_i) log(1-\theta_i)\} \\ &= \sum_{i=1}^{n}\{\ y_i \eta_{\boldsymbol{\beta}}(\mathbf{x}_i) - log[1+exp(\eta_{\boldsymbol{\beta}}(\mathbf{x}_i))]\ \}. \end{aligned}$$

For logistic regression analysis with a large number of explanatory variables, complete- or quasi-complete-separation may result which makes the maximum likelihood estimation unstable [1].

2.2 *Group lasso for logistic regression*

In this paper, we propose to perform model selection and parameter estimation for risk scoring system construction by using the group lasso algorithm of Meier et al. [7]. Instead of maximizing the log-likelihood $l(\boldsymbol{\beta})$ in the maximum likelihood method, the logistic group lasso estimates are calculated by minimizing the covex function

$$S_\lambda(\beta) = -l(\boldsymbol{\beta}) + \lambda \sum_{g=1}^{G} s(df_g) \parallel \boldsymbol{\beta}_g \parallel_2, \tag{3}$$

where λ is a tuning parameter for the penalty and $s(\cdot)$ is a function to rescale the penalty. In lasso algorithms, selection of λ is usually determined by cross-validation using data. For $s(\cdot)$, we use the square root function $s(df_g) = df_g^{0.5}$ as suggested in Meier et al. [7].

Here we consider selection of the tuning parameter λ from a multiplicative grid of 148 values $\{0.96\lambda_{max}, 0.96^2\lambda_{max}, 0.96^3\lambda_{max}, ..., 0.96^{148}\lambda_{max}\}$, as in Meier et al. [7]. Here λ_{max} is defined as

$$\lambda_{max} = \underset{g\in\{1,\dots,G\}}{max} \left\{ \frac{1}{s(df_g)} \parallel \mathbf{x}_g^T(\mathbf{y} - \bar{\mathbf{y}}) \parallel_2 \right\}, \tag{4}$$

such that when $\lambda = \lambda_{max}$, only the intercept is in the model. When λ goes to 0, the model is equivalent to ordinary logistic regression.

The optimal value of λ is determined through leave-one-out cross validation, which is a special case of K-fold cross-validation with K being equal to n, the number of observations in the sample. In each fold, leave-one-out cross validation uses a single observation from the original sample as the validation data, and the remaining observations as the training data. This step is repeated until each observation in the sample is used once as the validation data. Predicted probabilities of disease are calculated and are compared to true observed disease status to assess the predictive power of model.

Three criteria may be used to select the optimal value of λ. The log-likelihood score used in Meier et al. [7] is taken as the average of log likelihood of the validation data over all cross-validation sets. Another one is the the maximum correlation coefficient in Yeo and Burge [15] that is defined as

$$\rho_{max} = max\{\rho_\tau | \tau \in (0,1)\}, \tag{5}$$

where $\tau \in (0,1)$ is a threshold to classify the predicted probability into a binary disease status and ρ_τ is the Pearson correlation coefficient between the true binary disease status and the predictive disease status with threshold τ.

The third criterion is through the Receiver Operating Characteristic (ROC) analysis. For a given λ value, each leave-one-out cross validation results in one pair of the predicted probability of disease and the true observed disease status for the validation data. In total, we get n such pairs from all leave-one-out cross validation. Given a cutoff value for the predicted probability of disease, we can calculate the true positive rate (sensitivity) and false positive rate (1-specificity) using the n pairs. Then when varying the cutoff value for the predicted probability of disease, different pairs of true positive rate and false positive rate are generated. Plotting true positive rate versus false positive rate results in an ROC curve. Theoretically, cutoff values can be any values on the real line. The practical cutoff values are determined from resulting scores based on our data. The value of area under the ROC curve (AUC) as well as the confidence interval of AUC can be estimated through an approach proposed by DeLong et al. [3]. One interpretation of AUC is that it is the probability for the case that a random diseased individual has larger predicted probability of disease than a random non-diseased individual [9] and it has been used to assess predictive power of risk scoring systems [2, 10, 13]. We calculate the AUCs for all λs, and the value of λ with the largest AUC is chosen as the λ used in constructing the final scoring system.

3. Application to PRRS Data

In this section, we applied the proposed group lasso method to construct a scoring system for PRRS survey data of swine breeding herd sites in the United States and Canada.

3.1 *Data Description*

Surveys in the database completed between March 2005 and March 2009 were candidates for inclusion in the analysis. To avoid multiple surveys from a single swine breeding herd site, the study dataset was limited to responses obtained from the most recently completed survey for each site. Surveys meeting these criteria were extracted from the database, and identity information was removed. Incomplete surveys were excluded.

The outcome of interest is whether a site is positive or not. Positive sites are sites with clinical PRRS outbreak in the 3 years prior to when the assessment was completed, negative sites otherwise. The information to determine the outcome was obtained from the survey. A clinical PRRS outbreak was described in the survey as an increase in one or more reproductive performance measures that exceeded normal variation with diagnostic confirmation of PRRS virus involvement.

Of the 896 sites in the United States and Canada included in the study, 499 (56%) became positive during the past 3 years. 127 survey questions were considered potential explanatory variables in the analysis. The survey questions were first converted to dummy indicator variables. All of the responses for each survey question were defined as a group of variables.

3.2 *Application of logistic group lasso*

First, leave-one-out cross validation was used to choose tuning parameter λ, as described in Section 2.2.

For each λ in the grid $\{0.96\lambda_{max}, 0.96^2\lambda_{max}, 0.96^3\lambda_{max}, ..., 0.96^{148}\lambda_{max}\}$, the values of three evaluation criteria were calculated based on cross validation. The penalty parameter for final risk scoring system was selected to be the one that optimizes AUC.

The logistic group lasso based scoring system was compared with two other systems:

(1) The current risk scoring system used in versions 2 of the PRRS risk assessment for the breeding herd that is based on expert opinion,

(2) A risk scoring system based on multivariate logistic regression model selected by variable significance.

We constructed the significance based logistic model by following the method used by Van Zee et al. [13]. Specifically, we used forward stepwise variable selection to construct the logistic regression model with 0.05 significant level. Leave-one-out cross validation was applied to the model construction by variable significance, in the same manner as described for logistic group lasso.

ROC curves are plotted for the three risk scoring systems. A point estimate as well as the 95% confidence interval for the AUC are provided. The estimated AUCs were compared by using the nonparametric approach of DeLong et al. [3] and p-values were calculated.

R package “grplasso” [6] is used to perform group lasso logistic regression. Significance-based logistic model selection is performed using the LOGISTIC procedure in SAS. All other algorithms and calculations are programmed in R language.

3.3 *Results*

3.3.1 Determination of penalty parameter λ

The AUC, maximum correlation coefficient and log-likelihood are calculated based on leave-one-out cross validation and are plotted against the penalty parameter λ in Figure 1. The trends for all three criteria are similar with a sharp increase for small values of λ and gradual decrease after reaching the maximum. The optimal values of λ selected to maximize the three criteria are 11.72, 4.22 and 11.72 for AUC, maximum correlation coefficient and log-likelihood respectively.

[Figure 1 about here.]

3.3.2 Logistic group lasso based PRRS risk scoring system

The penalty parameter maximizing AUC (i.e. $\lambda = 11.72$) from the leave-one-out cross validation was used for the group lasso estimation of the logistic regression parameters. Figure 2 show the distributions of the predicted probabilities based on cross validation for both negative and positive farms. The predicted probability for positive farms is larger than that of negative farms in stochastic order. The actual risk score can take the value of the predicted probability, the linear predictor in the logistic regression model, or any strictly increasing function of the predicted probability. This is because the ROC curve for a predictor is invariate to such transformation.

[Figure 2 about here.]

In the resulting scoring system, 74 out of 127 survey questions were estimated with 0 coefficients and were excluded from the system. PADRAP questions target internal risks (bio-management of virus already present) and external risks (bio-exclusion of virus not present). A summary of the number of questions included in the final risk scoring system in each category of risk factors in the PRRS Risk Assessment for the Breeding Herd is shown in Table 1.

[Table 1 about here.]

Three out of eight questions regarding internal risk factors remained in the scoring system, and they are all factors concerning characteristics of the herd. Fifty questions remained in external risk factor section out of the total 119 questions. In the external risks section, all of the 14 categories had at least one question remaining in the final scoring system, except that all 4 questions concerning facilities were excluded. Several categories had a large number of questions removed. In particular, 8 of 12 (66.7%) questions concerning entry of animals into the breeding herd, 18 of 31 (58.1%) questions concerning entry of semen into the breeding herd, 16 of 29 (55.2%) questions concerning transportation of live animals, and 10 of 13 (76.9%) questions concerning neighboring pig farms were excluded.

3.4 *Comparison among risk scoring systems*

The ROC curves for the three risk scoring systems are plotted in Figure 3. The ROC curves for the two scoring systems based on logistic regression analyses of the data were constructed using the results of leave-one-out cross validation. The ROC curve of logistic group lasso apparently dominates the other two scoring systems.

[Figure 3 about here.]

Point and 95% interval estimates of AUC are reported in Table 2. The risk scoring system based on the logistic group lasso has the largest AUC = 0.848. This AUC estimate is significantly higher than those based on either expert opinion (AUC = 0.696, p-value < 0.001) or logistic regression model selected by variable significance (AUC = 0.807, p-value < 0.001).

[Table 2 about here.]

4. Simulation Study

A simulation study was performed to demonstrate the performance of group lasso logistic regression and compare it to ordinary forward stepwise logistic regression. We simulated 800 farms (i.e. n=800) and 120 survey questions (i.e. G=120) in each dataset, mimicking the real PADRAP data that motivates this paper. There were three possible answers for each question. The outbreak status for the i^{th} farm is generated from a $Bernoulli(1,\ p_i)$ distribution with p_i being a function of the question answers: $ln(\frac{p_i}{1-p_i}) = \beta_0 + \sum_{g=1}^{G} \mathbf{x_{i,g}^T}\beta_\mathbf{g}$, where β_0 is the intercept, $\mathbf{x_{i,g}}$ is a three dimentional indication vector for question answer and $\beta_\mathbf{g}$ is the parameter vector corresponding to the g^{th} predictor. Three types of questions were considered regarding their effects on the outcome. The first forty survey questions were important questions such that the coefficients of the three answers to these questions were all different:

$$\beta_\mathbf{g} = (1, 0, -1) \times \gamma, g = 1, ..., 40,$$

where the coefficient γ in the above simulation was set to control the strength of the questions' effect on the outcome. The second forty survey questions were also important questions but only one answer had a coefficient that was different from the other two answers:

$$\beta_\mathbf{g} = (1, 0, 0) \times \gamma, g = 41, ..., 80.$$

The last forty survey questions were unimportant questions such that all three answers had the same coefficients:

$$\beta_\mathbf{g} = (0, 0, 0) \times \gamma, g = 81, ..., 120.$$

The baseline coefficient β_0 was set to be $-\frac{40}{3}\gamma$ so that on average a farm have 50% of chance to have an outbreak. In this simulation study, we considered the situations where $\gamma = 0.1, 0.25, 0.5, 1$ and 2. For each value of γ, 20 datasets were simulated. We applied the logistic group lasso procedure described in Section 2.2 and the forward stepwise logistic regression to fit the model for each simulated data and calculate the AUC for each fitted model.

[Table 3 about here.]

Results for the simulation study are shown in Table 3. The mean AUC is increasing with the value of γ for both methods. The Wilcoxon signed-rank test result in the last column of Table 3 shows that AUC's from group lasso are significantly larger than those from logistic regression, especially for $\gamma \geq 0.25$.

5. Discussion

The risk scoring system for disease developed using the logistic group lasso algorithm significantly improves upon the current risk scoring system based on expert opinion for predicting whether a swine breeding site experienced a PRRS outbreak. The simuation study explores the performance of the scoring systems with different settings of coefficients. The logistic group lasso based scoring system is superior to the scoring system constructed through logistic regression selected by variable significance.

One advantage of group lasso is that it can be used as variable selection tool. It not only helps to find important explanatory factors in predicting the response variables but also identifies questions that could be removed from the survey without affecting the survey's ability for classifying herds according to whether they report clinical PRRS outbreaks in the previous 3 years.

Seventy-four of the 127 questions analyzed were excluded from the final risk scoring system based on logistic group lasso. The questions in the survey were assigned to the internal and external risk sections, in part, on the basis of possible routes of transmission of the PRRS virus. That questions in all except one of the external risk sections were included in the risk scoring system suggests that nearly all of the routes of transmission that were considered in Version 2 of the PRRS Risk Assessment for the Breeding Herd survey are important enough that excluding them would result in risk scoring system that performed significantly worse. This is consistent with the body of research demonstrating the importance of multiple routes by which PRRS virus is transmitted [17]. The analysis and results demonstrate how a program like PADRAP, that is supported by a professional association and used by a community of veterinarians, can generate valuable data that contributes to our understanding of the relative importance of risk factors and areas of risk factors for clinical outcomes. The results may also be used to decrease the reliance upon expert opinion to identify questions that should remain in the survey and those that may be eliminated to iteratively increase the value of the program and the data.

Acknowledgments

The authors thank AASV, National Pork Board, Boehringer Ingelheim Vetmedica Inc. and USDA (USDA NIFA Award 2008-55620-19132) for supporting PADRAP financially.

References

[1] P. D. Allison, *Convergence problems in logistic regression*, Micah Altman, Jeff Gill, and Michael McDonald (eds.), New York: Wiley-Interscience (2004), pp. 247-262.

[2] E. Barranger, C. Coutant, A. Flahault, Y. Delpech, E. Darai, and S. Uzan, *An axilla scoring system to predict non-sentinel lymph node ststus in breast cancer patients with sentinel lymph node involvement*, Breast Cancer Res. Tr. 91 (2005), pp. 113-119.

[3] E.R. DeLong, D.M. DeLong and D.L. Clarke-Pearson, *Comparing the areas under two or more correlated receiver operating characteristics curves: a non-parametric approach*, Biometrics 44 (1988), pp. 837-845.

[4] Y. Kim, J. Kim and Y. Kim, *Blockwise sparse regression*, Statist. Sin. 16 (2006), pp. 375-390.

[5] Y. Li, X. Wang, K. Bo, X. Wang, B. Tang, B. Yang, W. Jiang and P. Jiang, *Emergence of a highly pathogenic porcine reproductive and respiratory syndrome virus in the Mid-Eastern region China*, Vet. J. 174 (2007), pp. 557-584.

[6] L. Meier, *grplasso: Fitting user specified models with Group Lasso penalty.*, R package version 0.4-2. (2009) http://cran.r-project.org/package=grplasso

[7] L. Meier, S. van de Geer, and P. Buhlmann, *The group lasso for logistic regression*, J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (2008), pp. 53-71.

[8] E.J. Neumann, J.B. Kliebenstein, C.D. Johnson, J.W. Mabry, E.J. Bush, A.H. Seitzinger, A.L. Green and J.J. Zimmerman, *Assessment of the economic impact of porcine reproductive and respiratory syndrome on swine production in the United States*, J. Am. Vet. Med. Assoc. 227 (2005), pp. 385-392.

[9] M.S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, 2003.

[10] E. Rhatigan, I. Tyrmpas, G. Murray and J. N. Plevris, (2010) *Scoring system to identify patients at high risk of oesophageal cancer*, Brit. J. Surg. 97 (2010), pp. 1831-1837.

[11] C. Terpstra, G. Wensvoort G and J. Pol, *Experimental reproduction of porcine epidemic abortion and respiratory syndrome (mystery swine disease) by infection with Lelystad virus: Koch's postualtes fulfilled*, Vet. Quart. 13 (1991), pp. 131-136.

[12] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1996), pp. 267-288.

[13] K.J. Van Zee, D. E. Manasseh, J.L.B. Bevilacqua, S.K. Boolbol, J.V. Fey, L.K. Tan, P.I. Borgen, H.S. Cody III, and M.W. Kattan, *A nomogram for predicting the likelihood of additional nodal metastases in breast cancer patients with a positive sentinel node biopsy*, Ann. Surg. Oncol. 10 (2003), pp. 1140-1151.

[14] G. Wensvoort, C. Terpstra, J. Pol, E.A. Ter Laak, M. Bloemraad, E.P. DeKluyver, C. Kragten and L. Van Buiten, *Mystery swine disease in the Netherlands: the isolation of Lelystad virus*, Vet. Quart. 13 (1991), pp. 121-130.

[15] G.W. Yeo and C.B. Burge, *Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals*, J. Computnl Biol. 11 (2004), pp. 475-494.

[16] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. R. Stat. Soc. Ser. B Stat. Methodol. 68 (2007), pp. 49-67.

[17] J.J. Zimmerman , D.A. Benfield , S.A. Dee , P. Murtaugh , T. Stadejek , G.W. Stevenson and M. Torremorell, *Porcine reproductive and respiratory virus (Porcine Aterivirus)*, (2012) In: J.J. Zimmerman , L.A. Karriker, A. Ramirez, K.J. Schwartz, G.W. Stevenson (eds). Diseases of Swine, 10th edition. Wiley-Blackwell, Hoboken NJ, pp. 461-486.

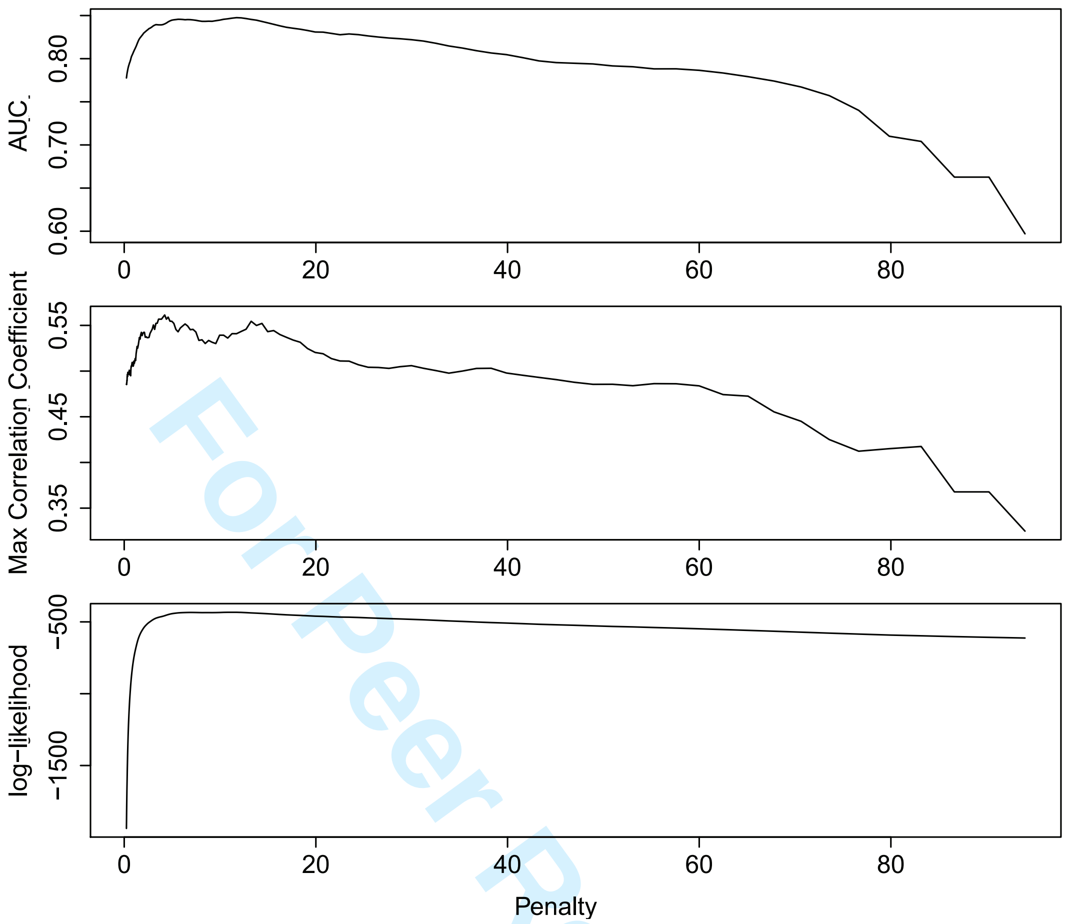

Figure 1. Three criteria for choice of penalty parameter λ

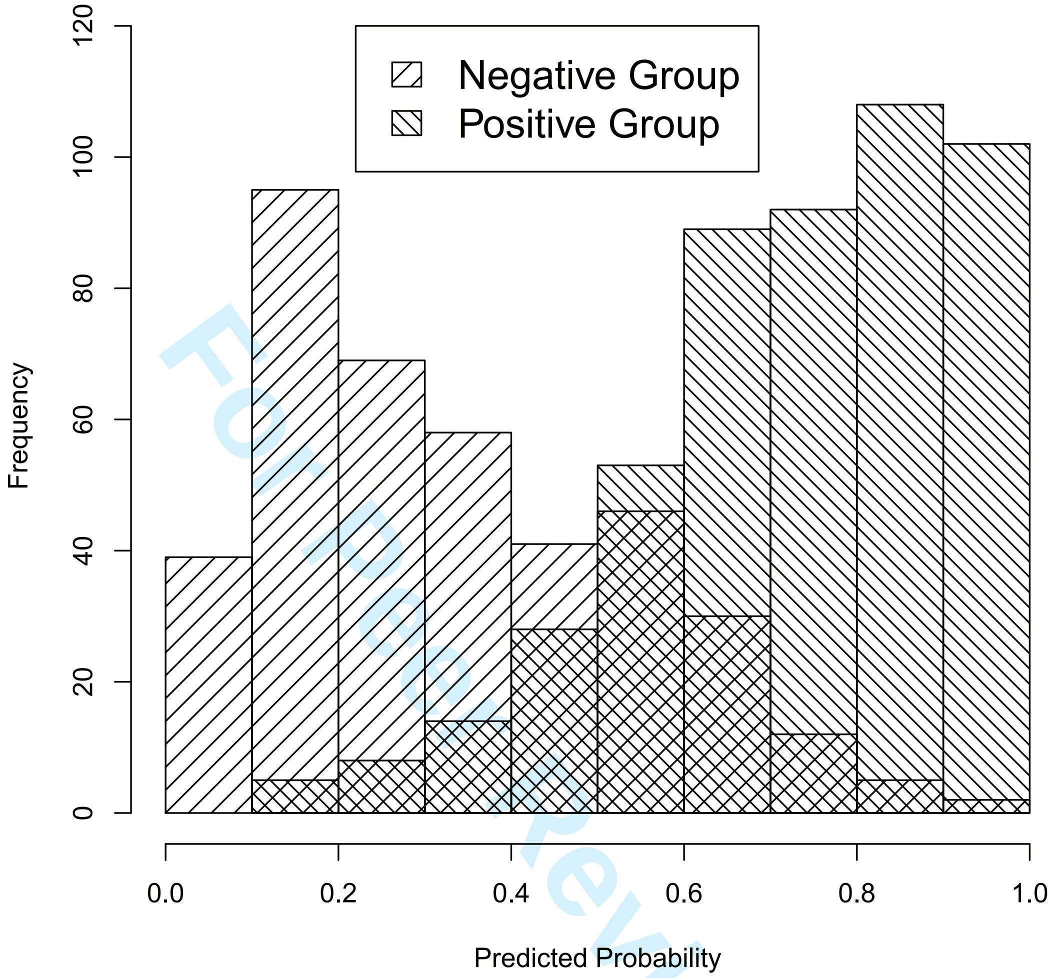

Figure 2. Distributions of estimated probabilities for both negative and positive groups

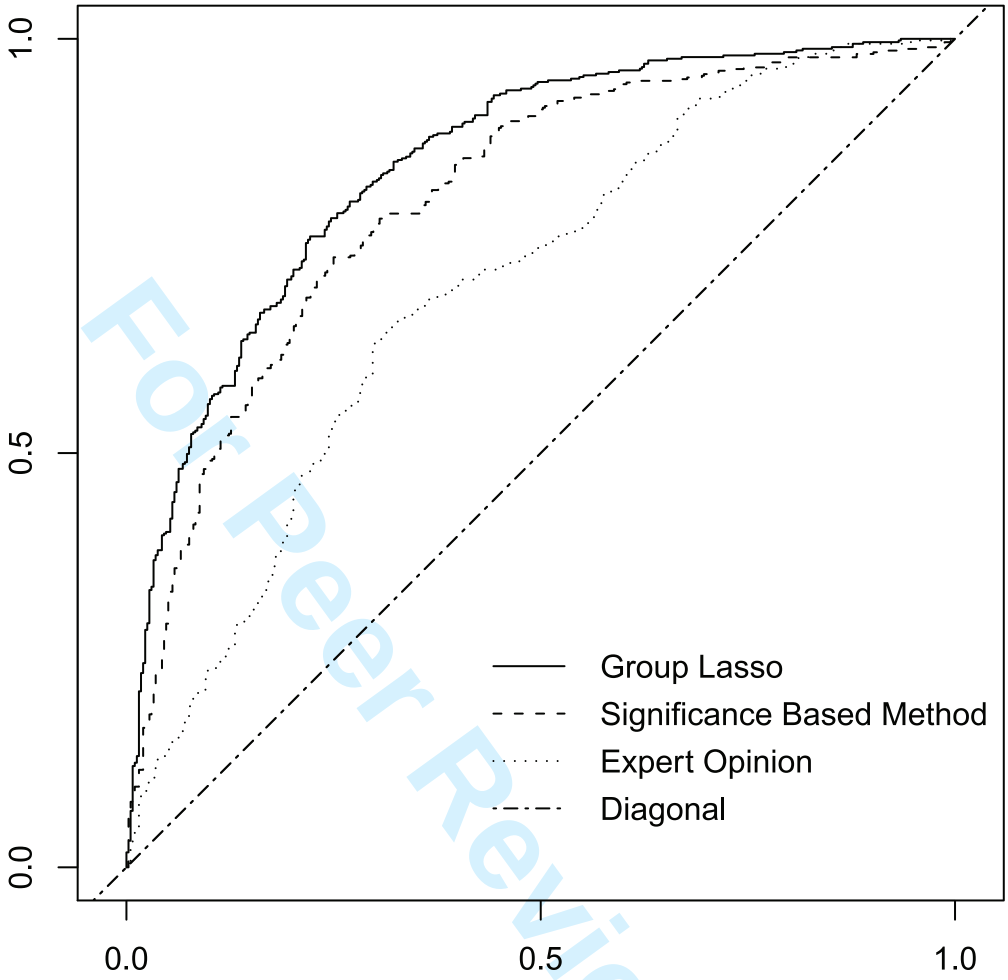

Figure 3. ROC curves for three risk scoring systems

Table 1. Summary of number of questions in the final risk scoring system by category of risk factors

Category of risk factors	Questions		Dummy Variables	
INTERNAL RISKS	Included	Total	Included	Total
Circulation Risk				
Characteristics of the herd	3	4	9	11
Characteristics of the site	0	2	0	5
Management practices	0	2	0	9
Total	***3***	***8***	***9***	***25***
EXTERNAL RISKS				
Pig Related				
Entry of replacement animals into the breeding herd	4	12	18	40
Entry of semen into the breeding herd	13	31	47	104
Non-Pig Related				
Transportation of live animals	13	29	38	71
Transportation of feed	1	1	2	2
Employee and service vehicles	1	2	3	6
Disposal of dead animals and waste management	2	8	3	10
Employees and visitors	5	9	15	19
Entry of supplies	1	1	3	3
Facilities	0	4	0	11
Biovectors	1	1	2	1
Density of pig farms in the area	3	3	10	10
Neighboring pig farms	3	13	12	28
Distance to pork industry infrastructure	2	4	5	11
Topography and forestation of surrounding area	1	1	3	3
Total	***50***	***119***	***161***	***319***

Table 2. AUC estimations for three risk scoring systems

Model Names	AUC	95% CI
Group Lasso	0.848	(0.822, 0.873)
Significance Based Method	0.807	(0.773, 0.841)
Expert Opinion	0.696	(0.661, 0.731)

Table 3. Simulation study result with various values of coefficient γ. Reported are mean and standard deviation of AUC for both methods, mean difference and p value from Wilcoxon signed rank test.

Coefficient γ	Group Lasso (mean±sd)	Logistic Regression (mean±sd)	p value
0.1	0.57 ± 0.03	0.54 ± 0.06	0.040
0.25	0.71 ± 0.02	0.64 ± 0.04	< 0.001
0.5	0.91 ± 0.03	0.78 ± 0.03	< 0.001
1	0.92 ± 0.01	0.82 ± 0.02	< 0.001
2	0.95 ± 0.01	0.84 ± 0.02	< 0.001