BENCHMARKING COW HEALTH STATUS WITH DAIRY HERD SUMMARY DATA

K.L. Parker Gaddis[1], J.B. Cole[2], J.S. Clay[3], and C. Maltecca[1]

[1] Department of Animal Science, North Carolina State University, Raleigh, NC, USA
[2] Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD, USA
[3] Dairy Records Management Systems, Raleigh, NC, USA

SUMMARY

Genetic improvement of dairy cattle health using producer-recorded data is feasible. Estimates of heritability are typically low, indicating that genetic progress will be slow. Health improvement may also be possible through incorporation of environmental and managerial aspects into herd health programs. The objective of this study was to utilize non-parametric methodologies including support vector machines and random forests to explore prediction of cow health status from routinely collected herd summary data. Random forest models attained the highest accuracy for predicting health status in all health categories. Accuracy of prediction (*SD*) of random forest models ranged from 0.87 (0.06) to 0.93 (0.001). Results of these analyses indicate that non-parametric algorithms, specifically random forest, can be used to accurately identify individual cows likely to experience a health event of interest. Further development of predictive models into herd management programs will continue to improve dairy health.

INTRODUCTION

To fully understand complex diseases, it is important to understand relationships between genotype, environment, and phenotype. Genetic improvement of dairy cattle health has been determined to be feasible utilizing producer-recorded data by several studies (Zwald *et al.* 2004; Parker Gaddis *et al.* 2012, 2014). Low estimates of heritabilities indicate, however, that genetic progress will be slow. Variance observed in lowly heritable traits can largely be attributed to non-genetic or environmental factors. In typical genetic evaluations, adjustments for environmental effects are accomplished by considering them as fixed effects. This disregards potential effects of management and environmental conditions on genetic expression (Windig *et al.* 2005). The question is then whether more rapid phenotypic improvement can be achieved if herd health programs incorporate environmental and managerial aspects.

Recent studies have incorporated herd characteristics into statistical models in relationship to reproductive efficiency (*e.g.*, Löf *et al.* 2007), production (*e.g.*, Windig *et al.* 2006), and health (*e.g.*, Stengärde *et al.* 2012). Farm staff or Dairy Herd Information (*DHI*) Association technicians regularly report on numerous herd characteristics observed on test days (DHI-202: Dairy Records Management Systems 2014). Additional environmental information is accessible through online databases including climatic, human census, and geographical data. Large numbers of variables create analysis challenges, ranging from increased data pre-processing to increased computing time. The majority of previous studies have utilized parametric statistical models to analyse herd characteristics (e.g., Stengärde *et al.* 2012), which can suffer from multiple testing problems and colinearities of numerous variables (Sato *et al.* 2008). Alternatively, non-parametric methods have recently been investigated to better handle numerous variables (*e.g.*, Schefers *et al.* 2010). The objective of this study was to utilize non-

parametric methodologies to explore prediction of cow health status from routinely collected herd summary data.

MATERIALS AND METHODS

Data. The DHI-202 Herd Summary provides a report on herd production, reproduction, genetics, udder health, and feed cost information (www.drms.org). Data were available from 2000 through 2011 from Dairy Records Management Systems (*DRMS*; Raleigh, NC). Four months (March, June, September, and December) of collected records were available for each year. Each herd summary contained over 1,100 variables. Number of contributing herds varied from 647 to 1,418, depending on month and year of reporting. Data included Ayrshire, Brown Swiss, Guernsey, Holstein, Jersey, and crossbred herds.

Supplementary data were acquired from publicly available datasets. The National Oceanic and Atmospheric Administration National Climatic Data Center (*NCDC*) provides information regarding temperatures, precipitation, degree-days, and drought indices (NCDC, 2014). Monthly summaries of data from the weather station located closest to each herd were merged with herd characteristic data. Estimates of population size were obtained on a county-basis from the United States Census Bureau (www.census.gov) as a measure of population density. Intercensal estimates from 2000 through 2010 were produced by updating the Census 2000 counts with estimates for components of population change (United States Census Bureau, 2012).

Voluntary producer-recorded health event data were available from DRMS (Raleigh, NC) from U.S. farms from 2000 through 2012. These data were merged with available production data. Health and production datasets were edited following the editing procedures described in Parker Gaddis *et al.* (2012). Health events included hypocalcemia, cystic ovaries, digestive problems, displaced abomasum, ketosis, mastitis, metritis, and retained placenta. These events were grouped into three main categories: mastitis, metabolic (hypocalcemia, digestive problems, displaced abomasum, and ketosis), and reproductive (cystic ovaries, metritis, and retained placenta) disorders. Health events were combined with herd characteristics based on date of health event occurrence.

Data pre-processing. A function was employed to determine and remove highly correlated variables by searching the correlation matrix. Editing was also performed to ensure that no variables were linear combinations of other variables (Kuhn 2013). Any variables with (near) zero variance were removed from the data. The above editing reduced the size of the dataset to approximately 3.7 million records with 829 variables. Missing records needed to be handled before statistical modeling could be performed. Variables with more than 50% missing observations (n = 70) were excluded from further analyses. Remaining missing herd characteristic records were imputed using an iterative principal component analysis algorithm (Husson and Josse 2012). Once a complete dataset was created, lactational incidence rate was calculated for each health event by herd-year as number of affected lactations per lactations at risk (Kelton *et al.* 1998).

Analyses. Analyses were performed using a binary indicator where "0" represented no incidence of a health event during a lactation and "1" represented at least one incidence of a respective health event during a lactation. Nonparametric models investigated included support vector machines (SVM) and random forests (RF). Briefly, an SVM model maps response variables to a higher-dimensional space that contains a "maximal separating hyperplane" (Sullivan 2012). The response variable should separate across this hyperplane into correct classifications (Sullivan 2012). Two different kernel

functions were investigated: a linear kernel and a radial basis kernel (RBF). The SVM^{perf} software (version 3.0) was utilized to fit SVM models (Joachims 2006).

Tree models are a data mining technique that are easily interpretable and implicitly perform feature selection, making them ideal for data with numerous variables (Kuhn and Johnson, 2013). Random forest (RF) models were utilized as a machine learning algorithm that fits many decision trees to bootstrapped samples of a dataset and then averages these decision trees to create a final predictive model (Breiman 2001). The "bigrf" package of R (R Core Team 2014) was used to fit these models (Lim *et al.* 2014). An optimal number of trees was determined prior to fitting a final model by testing a range of values for each health event category.

For all the above described models, 10-fold cross validation was used to evaluate predictive ability. Measures of predictive ability included accuracy, sensitivity, and specificity. Accuracy was calculated as the sum of true positives and true negatives divided by the sum of positive and negative incidences. Sensitivity, or true positive rate, was calculated as number of positive incidences correctly identified divided by the total number of positive incidences. Specificity, or true negative rate, was calculated as the number of negative incidences correctly identified divided by the total number of negative incidences (Fawcett 2006).

RESULTS AND DISCUSSION

The number of states reporting data ranged from 35 to 45, depending on health event. The most common herd size fell in a range of 100 to 299 cows; however, data included herds with fewer than 50 cows and a maximum herd size of over 5,500 cows. Overall median incidence rates were 24%, 8%, and 18% for mastitis, metabolic, and reproductive health events, respectively. These fall within the range of previously reported incidence rates (Parker Gaddis *et al.* 2012).

Predictive ability in training datasets were similar to those estimated for validation data, indicating that the models were not being overfit to training data. Prediction accuracies, sensitivity, and specificity for SVM models are shown in Table 1. Linear and RBF kernels performed similarly for all health event categories. These models had much higher specificity compared to sensitivity, indicating that they were more capable of identifying healthy cows.

Table 1 Summary of model performance for incidences of mastitis, reproductive, and metabolic health events averaged across 10-fold cross validation results fitting support vector machine (SVM) and random forest models

		Accuracy (Validation)	Sensitivity (Validation)	Specificity (Validation)
Mastitis	SVM (linear) c=0.01*	0.70 (0.003)	0.24 (0.002)	0.88 (0.003)
	SVM (RBF) c=10.0	0.70 (0.01)	0.39 (0.03)	0.83 (0.02)
	Random forest	0.93 (0.001)	0.82 (0.003)	0.97 (0.001)
Reproductive	SVM (linear) c=0.005	0.69 (0.002)	0.32 (0.01)	0.79 (0.004)
	SVM (RBF) c=10.0	0.77 (0.01)	0.33 (0.03)	0.88 (0.02)
	Random forest	0.92 (0.001)	0.74 (0.006)	0.97 (0.0007)
Metabolic	SVM (linear) c=0.01	0.76 (0.03)	0.12 (0.03)	0.93 (0.05)
	SVM (RBF) c=10.0	0.75 (0.01)	0.25 (0.02)	0.88 (0.01)
	Random forest	0.87 (0.061)	0.57 (0.145)	0.96 (0.04)

*c represents the error penalty tuning parameter for SVM models

The optimal number of trees for RF models was determined to be 25, regardless of health event. Random forest models had the best predictive ability across all health event categories (Table 1). Overall, sensitivity was lower than specificity; however, sensitivity was higher for RF models compared to SVM models.

Each of the models investigated herein had benefits and disadvantages. Support vector machines are a flexible class of models with several kernels that can be employed. These models require estimation of tuning parameters and results can be more difficult to interpret. Random forests were the most flexible models. They can easily handle a large number of variables, as well as missing records. Random forest models can be more difficult to interpret than a single decision tree, but tend to have better predictive performance and are capable of identifying influential variables.

This study suggests that benchmarking of cow health is feasible with routinely collected data. Improvement in predictive ability may be possible by modeling each health event as opposed to grouping events into categories. Factors that predispose a cow to retained placenta, for example, may not be the same as factors that increase a cow's risk of cystic ovaries. With continued development and incorporation of predictive models into herd management, routinely recorded herd data could be used in conjunction with genomic selection strategies to further improve dairy cattle health.

REFERENCES

Breiman L. (2001) *Mach. Learn.* **45**: 5.

Fawcett T. (2006) *Pattern Recognit. Lett.* **27**: 861.

Husson F. and Josse J. (2012) 'Handling missing values with/in multivariate data analysis (principal component methods).'

Joachims T. (2006) In 'Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining' ACM Press, New York.

Kelton D.F., Lissemore K.D. and Martin R.E. (1998) *J. Dairy Sci.* **81**: 2502.

Kuhn M. (2013) 'Classification and Regression Training.'

Kuhn M. and Johnson K. (2013) 'Applied Predictive Modeling' Springer, New York.

Lim A., Breiman L. and Cutler A. (2014) 'bigrf: Big Random Forests: Classsification and Regression Forests for Large Data Sets.'

Löf E., Gustafsson H. and Emanuelson U. (2007) *J. Dairy Sci.* **90**: 4897.

Parker Gaddis K.L., Cole J.B., Clay J.S. and Maltecca C. (2012) *J. Dairy Sci.* **95**: 5422.

Parker Gaddis K.L., Cole J.B., Clay J.S. and Maltecca C. (2014) *J. Dairy Sci.* **97**: 3190.

R Core Team (2014) 'R: A Language and Environment for Statistical Computing.'

Sato K., Bartlett P.C., Alban L., Agger J.F. and Houe H. (2008) *Acta Vet. Scand.* **50**: 4.

Schefers J.M., Weigel K.A., Rawson C.L., Zwald N.R. and Cook N.B. (2010) *J. Dairy Sci.* **93**: 1459.

Stengärde, L., Hultgren, J., Tråvén, M., Holtenius K. and Emanuelson U. (2012) *Prev. Vet. Med.* **103**: 280.

Sullivan, R. (2012) 'Introduction to Data Mining for the Life Sciences' Springer, New York.

United States Census Bureau. (2012) 'Methodology for the Intercensal Population and Housing Unit Estimates: 2000 to 2010.'

Windig J.J., Calus M.P.L., Beerda B. and Veerkamp R.F. (2006) *J. Dairy Sci.* **89**: 1765.

Windig J.J., Calus M.P.L. and Veerkamp R.F. (2005) *J. Dairy Sci.* **88**: 335.

Zwald N.R., Weigel K.A., Chang Y.M., Welper R.D. and Clay J.S. (2004) *J. Dairy Sci.* **87**: 4287.