J. Dairy Sci. 99:1298–1314
http://dx.doi.org/10.3168/jds.2015-9840

Benchmarking dairy herd health status using routinely recorded herd summary data

K. L. Parker Gaddis,*[1] J. B. Cole,† J. S. Clay,‡ and C. Maltecca§
*Department of Animal Sciences, University of Florida, Gainesville 32611
†Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350
‡Dairy Records Management Systems, Raleigh, NC 27603
§Department of Animal Science, North Carolina State University, Raleigh 27695

ABSTRACT

Genetic improvement of dairy cattle health through the use of producer-recorded data has been determined to be feasible. Low estimated heritabilities indicate that genetic progress will be slow. Variation observed in lowly heritable traits can largely be attributed to nongenetic factors, such as the environment. More rapid improvement of dairy cattle health may be attainable if herd health programs incorporate environmental and managerial aspects. More than 1,100 herd characteristics are regularly recorded on farm test-days. We combined these data with producer-recorded health event data, and parametric and nonparametric models were used to benchmark herd and cow health status. Health events were grouped into 3 categories for analyses: mastitis, reproductive, and metabolic. Both herd incidence and individual incidence were used as dependent variables. Models implemented included stepwise logistic regression, support vector machines, and random forests. At both the herd and individual levels, random forest models attained the highest accuracy for predicting health status in all health event categories when evaluated with 10-fold cross-validation. Accuracy (SD) ranged from 0.61 (0.04) to 0.63 (0.04) when using random forest models at the herd level. Accuracy of prediction (SD) at the individual cow level ranged from 0.87 (0.06) to 0.93 (0.001) with random forest models. Highly significant variables and key words from logistic regression and random forest models were also investigated. All models identified several of the same key factors for each health event category, including movement out of the herd, size of the herd, and weather-related variables. We concluded that benchmarking health status using routinely collected herd data is feasible. Nonparametric models were better suited to handle this complex data with numerous variables. These data mining techniques were able to perform prediction of health status and could add evidence to personal experience in herd management.
Key words: herd health status, producer-recorded data, prediction, benchmarking

INTRODUCTION

To fully understand complex diseases, it is important to understand relationships between genotype, environment, and phenotype. Complex causal relationships have been identified between different diseases, culling, and production (Dhakal et al., 2015). Increased production of dairy cattle has resulted in a subsequent decline in health and fertility traits (Esposito et al., 2014). Concurrently, concern over animal welfare and use of antibiotics has steadily increased (Nyman et al., 2007). Understanding these relationships may help us to better describe the disease process (Dhakal et al., 2015). Genetic improvement of dairy cattle health has been determined to be feasible using producer-recorded data by several studies (Zwald et al., 2004; Parker Gaddis et al., 2014, 2012). Low estimated heritabilities of health events indicate, however, that genetic progress will be slow.

Variance observed in lowly heritable traits can largely be attributed to nongenetic or environmental factors. In typical genetic evaluations, adjustments for environmental effects are accomplished by considering them fixed effects. This disregards effects of management and environmental conditions on genetic expression (Windig et al., 2005). It also ignores any associations that exist between genetic and environmental effects. Dechow and Goodling (2008) showed that heritabilities estimated using data from high-performing herds were higher than those from typical or poor-performing herds, suggesting that rates of genetic gain may be higher when cows are provided with favorable environments for production. In addition, research has indicated that genetic correlations, such as between fertility and milk production, will depend upon herd environment (Win-

Received May 20, 2015.
Accepted September 25, 2015.
[1]Corresponding author: klpgaddis@ufl.edu

dig et al., 2006). The question then arises as to whether more rapid improvement can be achieved if herd health programs incorporate environmental and managerial aspects.

Previous studies have investigated the effect of environmental characteristics on dairy cattle health. An early study was able to establish 5 farm "health profiles" according to the incidence levels of health disorders and farm structure data (Faye, 1992). Health disorders included infectious diseases of the foot, uterus, and teat, and calving disorders; farm structure was represented as traditional, intensive, or intermediate. Data were collected throughout 1979 from 83 dairy farms in France and included 25 specific health events in addition to herd management variables. Hierarchical classification was used to group the farms into similar classes and confirmed a relationship between farm type and herd health profile (Faye, 1992). Path analysis and multiple logistic regression were utilized to evaluate interrelationships between herd management practices and postpartum health disorders on 32 farms located in New York State (Correa et al., 1990). Disorders included dystocia, retained placenta, metritis, cystic ovary, milk fever, ketosis, left displaced abomasum, and mastitis. Management characteristics were collected through a questionnaire provided to the person primarily responsible for care of the herd. A 2-stage analysis was performed to identify management factors and develop a path model of interrelationships between herd management and herd incidence rate (Correa et al., 1990).

More recent studies have been conducted incorporating herd characteristics in relationship to reproductive efficiency (Löf et al., 2007; Schefers et al., 2010), production (Windig et al., 2005, 2006; Simensen et al., 2010), and health (Svensson et al., 2006; Green et al., 2007; Stengärde et al., 2012). Many of these studies have utilized surveys or questionnaires to assess herd characteristics (Correa et al., 1990; Sato et al., 2008; Hill et al., 2009), which can limit the amount of data that can be collected. Data collected from a designed study may not always reflect common management practices, thus limiting applicability (Coppa et al., 2013). Data can also be limited by the chosen analysis method. The majority of previous studies have used parametric statistical models to analyze herd characteristics (Svensson et al., 2006; Löf et al., 2007; Stengärde et al., 2012), which can suffer from problems with multiple testing and collinearities with numerous variables (Sato et al., 2008). Alternatively, nonparametric methodologies have recently been investigated, such as principal component analysis (Windig et al., 2006) or common factor analysis (Enevoldsen et al., 1996), as well as regression-based decision trees (Schefers et al., 2010) to better handle numerous variables.

Farm staff or DHIA technicians report numerous herd characteristics regularly on farm test days. These reports include data on herd production, reproduction, genetics, udder health, and feed costs (Dairy Records Management Systems, 2014). Additional environmental data can be accessed through online databases such as the National Climatic Data Center (www.ncdc.noaa.gov), the United States Census Bureau (www.census.gov), and the United States Geographical Survey (www.usgs.gov). The availability of numerous variables from field data presents analysis challenges ranging from increased data preprocessing to increased computing time. Although the majority of prior research has been conducted with parametric statistical methods (Windig et al., 2005, 2006; Sato et al., 2008), a more flexible approach might be possible when analyzing large numbers of variables utilizing data mining techniques. Data mining allows patterns to be explored and is increasingly employed because of the explosion of data availability in many fields (Sullivan, 2012). The objective of this study was to utilize parametric and nonparametric methods to explore prediction of herd health status. Routinely collected herd summary data were used for benchmarking health status at the individual and herd level.

MATERIALS AND METHODS

Data

The DHI-202 Herd Summary provides a "comprehensive herd analysis and management report including production, reproduction, genetics, udder health, and feed cost information" (Dairy Records Management Systems, 2014). Data are collected by farm staff or DHI technicians and compiled each test day. Categories of data include production, income, and feed cost summary; miscellaneous herd information; reproductive summary of current breeding herd; reproductive summary of total herd; birth summary; yearly reproductive summary; cows to be milking, dry, or calving by month; stage of lactation profile; identification and genetic summary; production by lactation summary; current SCC summary; dry cow profile; yearly summary of cows entered and left the herd; and yearly production and mastitis summary. An example DHI-202 Herd Summary report is included in the supplementary material (Supplementary Figure S1; http://dx.doi.org/10.3168/jds.2015-9840); detailed information on the report can be found at www.drms.org. Data were available from 2000 through 2011 from Dairy Records

Management Systems (Raleigh, NC). Four months of collected records were included for each year: March, June, September, and December. Each herd summary contained over 1,100 variables. Number of contributing herds varied from 647 to 1,418, depending on year and month of reporting.

Supplementary data were acquired from publicly available databases. The National Oceanic and Atmospheric Administration (**NOAA**) National Climatic Data Center (**NCDC**) provides information regarding temperatures, precipitation, degree-days, and drought indices from 1895 through the present (Diamond et al., 2013). Monthly summaries were obtained from NCDC Quality Controlled Local Climatological Data (http://cdo.ncdc.noaa.gov/qclcd/) from land-based data sets for each month and year of available herd summary data. The NCDC provides geographic coordinates for each land-based station. Geographic coordinates were approximated for each herd based on herd Zip code using the R (R Development Core Team, 2012) package "zipcode" (Breen, 2012). Weather data from the weather station located closest to each herd was merged with herd characteristic data. The land-based weather station located nearest to each herd was determined utilizing the "geosphere" package (Hijmans et al., 2012) of R (R Core Team, 2014) and based on distance between geographic coordinates.

Estimates of population size were obtained on a county basis from the United States Census Bureau website (www.census.gov). Intercensal estimates from 2000 through 2010 were produced by updating the Census 2000 counts with estimates for components of population change. Components of population change included factors such as births to US women, deaths of US residents, and migration. Estimated population change was reconciled with counts from the 2010 Census to produce a consistent time-series of population estimates from 2000 to 2010 (United States Census Bureau, 2012). Census data were combined with herd characteristic data based on reported county of herd.

Voluntary producer-recorded health event data were available from Dairy Records Management Systems (Raleigh, NC) from US farms from 2000 through 2012. These data were matched to available production data. Both health and production data sets were edited following the general editing procedures described in Parker Gaddis et al. (2012). For a herd-year to be considered as reporting a health event, each herd-year had to report at least one incidence. Cows with at least one occurrence of a health event were coded as "1" for the respective event, and "0" otherwise. Health events included for analyses were hypocalcemia, cystic ovaries, digestive problems, displaced abomasum, ketosis, mastitis, metritis, and retained placenta. These events were grouped into 3 main categories based on their relationship (see Dhakal et al., 2015): mastitis, metabolic (hypocalcemia, digestive problems, displaced abomasum, ketosis), and reproductive (cystic ovaries, metritis, and retained placenta). Health events were combined with herd characteristics based on date of health event occurrence. Date of each health event was rounded to the nearest month using the R package "lubridate" (Grolemund and Wickham, 2011). Events occurring in January, February, or March were merged with herd characteristic summaries from March; events occurring in April, May, or June were merged with herd characteristic summaries from June; events occurring in July, August, or September were merged with herd characteristic summaries from September; and events occurring in October, November, or December were merged with herd characteristic summaries from December. This is not a perfect method; however, we believe it to be a reasonable approach for combining all the available data.

Data Preprocessing

A correlation analysis was performed as an initial step to reduce the dimensionality of the data and eliminate high correlations between variables. This was applied to each section of the DHI-202 Herd Summary, as well as the weather data, using the R package "caret" (Kuhn, 2013). Briefly, a function was used to determine highly correlated variables by searching the correlation matrix. When 2 variables had a correlation >0.90, the function removed the variable that had the largest absolute correlation averaged across all variables. Additional variable editing was performed to ensure that no variables were linear combinations of other variables in the data set. The "caret" package (Kuhn, 2013) of R was also used for this; it uses the QR decomposition of the matrix to determine sets of linear combinations. Fifteen variables were removed to eliminate any linear combinations within the data set. Also, any variables with near zero (at least 95% of records with the same value) or zero variance were removed from the data. The above editing reduced the size of the final data set to 3,693,778 cow records with 829 variables.

Missing records also had to be handled before statistical modeling could be performed. The distribution of missing records within each variable was examined to estimate a reasonable threshold of missing data beyond which a variable would be excluded. Based on this, variables with >50% missing observations (n = 70) were excluded from further analyses. Remaining missing records were imputed using an iterative princi-

pal component analysis algorithm (Husson and Josse, 2012). Briefly, missing values were initialized with the overall mean of each variable. A principal component analysis was then performed on this data set iteratively until convergence was reached to estimate missing values. Once a complete data set was created, lactational incidence rate (LIR) was calculated for each health event by herd-year as number of affected lactations per lactations at risk:

$$LIR = \frac{LAC_d}{LAC_t},$$

where LAC_d indicated number of first occurrences of a specific health event in a lactation, and LAC_t indicated number of lactations at risk (Kelton et al., 1998). Lactations at risk were considered the total number of cow lactation records that had the potential to experience a health event of interest. For this, the herd had to be considered actively recording the health event during the cow's lactation.

Analyses

Analyses were performed at both the herd and individual levels. For herd-based analyses, the objective was not to estimate herd disease incidence precisely. Conversely, it may be more informative to predict whether a herd has incidence below or above average, and which variables affect this incidence. To evaluate each model's ability to classify herds in this way, herd incidence was converted to a binary indicator. Preliminary analyses investigated several methods of splitting the data. Herds with event incidence below the median incidence of all herds were classified as having "low" incidence; herds with event incidence above the median incidence of all herds were classified as having "high" incidence. Analyses performed at the individual level used 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.

To evaluate the predictive ability of each model fairly, data were divided into training and prediction subsets. Cross-validation was performed using 2 different splitting methods. The first cross-validation scheme split the data into approximately 75% training and 25% validation based on year of health event occurrence. This was done to replicate data accumulation as it typically occurs in the dairy industry. Training data consisted of records through 2009; validation data consisted of records from 2010 and later. True 10-fold cross-validation was also performed to have a more statistically sound evaluation for comparative purposes.

Initial analyses fit a parametric model for each event category using forward and reverse stepwise regression. The "step" function of R (R Core Team, 2014) was used to test variables and determine the best final model based on the Akaike information criterion (AIC) using the training data set. Final models were then fit with the selected variables following the model shown below:

$$\lambda = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + e,$$

where λ represents a vector of unobserved liabilities to the given disease category, β_0 represents the intercept, β_i represents the regression coefficient for trait i, x_i is the observed value for the ith trait, and e represents random residual, modeled following $\boldsymbol{N}(0, \mathbf{I})$, fixing the variance at 1 to attain identifiability. Prediction ability of this model was evaluated by fitting the model with the validation data set(s).

Although logistic models are often favored for their simplicity, they do have disadvantages. Logistic regression is a form of linear model, which assumes that model residuals follow a normal distribution, which may not always be a valid assumption. They also encounter difficulties when multicollinearities exist. Because of these disadvantages, several nonparametric algorithms were explored. Support vector machines (**SVM**) were selected as a nonparametric classification algorithm. Support vector machines were developed from foundations of robust regression (Kuhn and Johnson, 2013). Briefly, an SVM model maps response variables to a higher-dimensional space that contains a "maximal separating hyperplane." The response variable should separate across this hyperplane into correct classifications (Sullivan, 2012). Different kernel functions can be used in an SVM model (e.g., linear, polynomial, radial basis function), allowing for great flexibility (Kuhn and Johnson, 2013). Two different kernel functions were used in these analyses: a linear kernel and a radial basis kernel (**RBF**). The SVM^{perf} software (version 3.0) was utilized to fit SVM models (Joachims, 2006). The software consists of a learning module (svm_perf_learn) and a classification module (svm_perf_classify) that were used for training and prediction, respectively.

Several machine learning algorithms were also explored before analysis. Tree models are one of the most widely implemented data mining techniques (Sullivan, 2012). The inherent structure of these models lends them easy interpretation. Tree models also implicitly perform feature selection, making them ideal for data with many variables (Kuhn and Johnson, 2013). One such algorithm first proposed by Breiman (2001) is ran-

Table 1. Summary statistics for each individual health event including total number of records, total number of herds reporting, median lactational incidence rate (LIR), and total number of states reporting

Health event	No. of cows	No. of records	No. of herd-years	Median LIR	No. of states reporting
Hypocalcemia	142,319	242,010	887	0.026	35
Cystic ovaries	216,115	363,919	2,065	0.119	39
Digestive problems	188,119	310,862	981	0.049	39
Displaced abomasum	215,064	370,833	1,267	0.050	36
Ketosis	141,625	234,597	679	0.065	31
Mastitis	284,957	494,506	2,403	0.245	45
Metritis	225,357	378,097	1,528	0.118	39
Retained placenta	204,962	316,490	1,278	0.083	41

dom forest (**RF**), which was used as a machine learning algorithm herein. Random forest is an ensemble algorithm that fits many decision trees to bootstrapped samples of a data set and then averages these decision trees to create a final predictive model (Breiman, 2001). These models are also ideal as they are robust to over-fitting (González-Recio and Forni, 2011). The "bigrf" package (Lim et al., 2014) of R (R Core Team, 2014) was used to fit RF models to the data.

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). A common measure that combines both sensitivity and specificity is the receiver operating characteristic (**ROC**) curve. A perfect model would have 100% sensitivity and specificity, which when viewed graphically, is a single step from 0% sensitivity and specificity to 100% sensitivity and specificity. The area under the ROC curve (**AUC**) would be equal to 100%. Alternatively, a model that performs no better than random chance would have a perfectly diagonal ROC curve and AUC would be equal to 50%. Receiver operating characteristic curves were produced as an alternative to examine model predictive ability.

RESULTS AND DISCUSSION

Summary statistics for each individual health event are included in Table 1. Data encompassed years 2000 through 2011 and included Ayrshire, Brown Swiss, Guernsey, Holstein, Jersey, and crossbred herds. The number of states reporting data ranged from 35 to 45, depending on the health event. The most common herd size fell in a range of 100 to 299 cows; however, data included herds with <50 cows and herds with >1,000 cows, with a maximum herd size of over 5,500 cows. A total of 2,403 herd-years reported mastitis incidences. The overall median incidence of mastitis was 24%, which falls within the range of previously reported incidence rates (Parker Gaddis et al., 2012). There were 2,290 herd-years reporting health events in the metabolic category, with a median incidence rate equal to 8%. The reproductive category had 3,191 herd-years reporting, with a median incidence rate of 18%. Because these health event data were extracted from producer-recorded information, we cannot conclusively discern between clinical and subclinical incidences.

Logistic Regression—Herd Level

Stepwise logistic regression was used to identify significant variables to be included in a final logistic model for each category based on Akaike information criterion. These parametric models are commonly used for binary traits and served as a base for comparison with the nonparametric models. The number of variables selected for each model differed depending on health category: 82, 111, and 145 for mastitis, reproductive, and metabolic events, respectively. This is a substantial reduction from the initial number of variables. Descriptions of the selected variables are further discussed below.

Logistic regression had the lowest prediction accuracy among models for mastitis, reproductive, and metabolic events when data were split into training and validation sets by year. It is interesting to note that logistic regression models had the highest accuracy in the training data sets for all event categories. The logistic regression models were not capable, however, of predicting new records accurately, possibly due in part to over-parameterization. When a model is over-parameterized, it is prone to overemphasizing patterns that are only characteristic of the training data (Kuhn and Johnson, 2013). Despite very high predictive ability with training data, performance declines drastically

Table 2. Summary of model performance[1] for discretized herd incidence of mastitis, reproductive, and metabolic health events when data are split by year

Health events	Accuracy (Training)	Sensitivity (Training)	Specificity (Training)	Accuracy (Validation)	Sensitivity (Validation)	Specificity (Validation)
Mastitis						
Stepwise regression	0.71	0.65	0.76	0.43	0.25	0.82
SVM (linear) c = 0.01	0.66	0.61	0.70	0.49	0.41	0.66
SVM (RBF) c = 70.0	0.68	0.68	0.69	0.55	0.54	0.58
Random forest[2]	0.63	0.53	0.73	0.48	0.32	0.81
Reproductive						
Stepwise regression	0.71	0.67	0.75	0.42	0.21	0.83
SVM (linear) c = 0.01	0.67	0.64	0.70	0.47	0.29	0.83
SVM (RBF) c = 60.0	0.63	0.65	0.62	0.56	0.51	0.65
Random forest[2]	0.66	0.58	0.73	0.43	0.19	0.92
Metabolic						
Stepwise regression	0.78	0.74	0.81	0.46	0.40	0.61
SVM (linear) c = 0.01	0.67	0.62	0.71	0.55	0.55	0.54
SVM (RBF) c = 60.0	0.67	0.62	0.69	0.54	0.47	0.71
Random forest[3]	0.67	0.57	0.76	0.53	0.82	0.40

[1]SVM (linear) = support vector machine with linear kernel; SVM (RBF) = support vector machine with radial basis function kernel; c = regularization parameter.

[2]Random forest model for health event category used 25 trees.

[3]Random forest model for health event category used 40 trees.

when the model is exposed to new data. When data were evaluated using true 10-fold cross-validation, logistic regression was typically neither the best nor the worst model. Predictive ability when data were split by year of occurrence is included in Table 2; predictive ability when data were evaluated by true 10-fold cross-validation is included in Table 3. Predictive ability depicted by ROC curves is shown in Figures 1 and 2 for cross-validation performed based on year and for true 10-fold cross-validation, respectively. Because accuracy reflects a combination of prediction measures, sensitivity and specificity measures may be more informative. Logistic regression models tended to have higher specificity compared with sensitivity for all health event categories when data were split based on year (Table 2). Logistic regression models were better able to identify low incidence herds versus high incidence herds for all health event categories when data were split based on year.

The accuracy of prediction when fitting a logistic regression model for herd-level incidence was slightly below 50% for all health event categories when data

Table 3. Summary of model performance for discretized herd incidence of mastitis, reproductive, and metabolic health events averaged across 10-fold cross-validation results (SD)

Health events	Accuracy (Training)	Sensitivity (Training)	Specificity (Training)	Accuracy (Validation)	Sensitivity (Validation)	Specificity (Validation)
Mastitis						
Stepwise regression	0.70 (0.01)	0.69 (0.01)	0.72 (0.01)	0.59 (0.04)	0.57 (0.06)	0.61 (0.06)
SVM (linear) c = 0.01	0.67 (0.006)	0.63 (0.01)	0.70 (0.01)	0.61 (0.04)	0.57 (0.06)	0.65 (0.04)
SVM (RBF) c = 70.0	0.66 (0.009)	0.68 (0.03)	0.63 (0.03)	0.55 (0.02)	0.59 (0.06)	0.52 (0.06)
Random forest[2]	0.58 (0.02)	0.58 (0.03)	0.58 (0.03)	0.61 (0.04)	0.62 (0.06)	0.61 (0.05)
Reproductive						
Stepwise regression	0.72 (0.02)	0.71 (0.02)	0.72 (0.02)	0.59 (0.02)	0.58 (0.04)	0.59 (0.02)
SVM (linear) c = 0.01	0.65 (0.003)	0.65 (0.009)	0.66 (0.01)	0.61 (0.04)	0.60 (0.03)	0.62 (0.05)
SVM (RBF) c = 60.0	0.64 (0.008)	0.64 (0.03)	0.64 (0.04)	0.55 (0.02)	0.55 (0.06)	0.56 (0.06)
Random forest[2]	0.59 (0.02)	0.59 (0.02)	0.59 (0.03)	0.62 (0.04)	0.62 (0.05)	0.63 (0.04)
Metabolic						
Stepwise regression	0.75 (0.01)	0.74 (0.01)	0.76 (0.01)	0.61 (0.04)	0.60 (0.05)	0.61 (0.05)
SVM (linear) c = 0.01	0.67 (0.006)	0.65 (0.01)	0.69 (0.01)	0.61 (0.04)	0.59 (0.07)	0.64 (0.04)
SVM (RBF) c = 60.0	0.67 (0.01)	0.62 (0.02)	0.73 (0.03)	0.58 (0.01)	0.52 (0.04)	0.64 (0.04)
Random forest[3]	0.60 (0.02)	0.60 (0.04)	0.60 (0.02)	0.63 (0.04)	0.63 (0.06)	0.62 (0.06)

[1]SVM (linear) = support vector machine with linear kernel; SVM (RBF) = support vector machine with radial basis function kernel; c = regularization parameter.

[2]Random forest model for health event category used 25 trees.

[3]Random forest model for health event category used 40 trees.

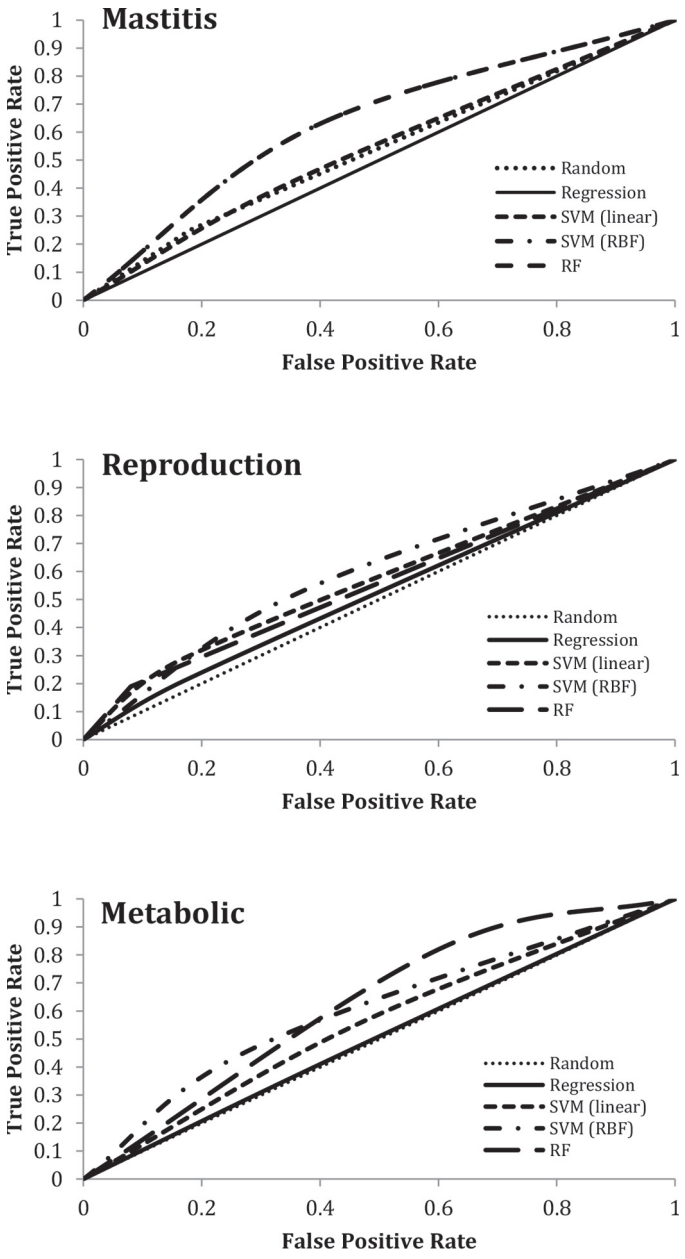

Figure 1. Receiver operating characteristic (ROC) curves for discretized herd incidence of mastitis, reproductive, and metabolic health events when data are split by year.

were split based on year. It is important to note here that accuracy is a combination of sensitivity and specificity. The low accuracy in this case was due primarily to poor prediction of low-health herds, as shown in Table 2. These models were able to correctly identify high-health herds 60 to 80% of the time. When data were split for true 10-fold cross-validation, accuracy was approximately 0.60 for all health events (Table 3). Low accuracy was expected because logistic regression had the least flexibility of the models used. Values for specificity and sensitivity for all health events in true cross-validation were very similar.

In addition to prediction of health status, logistic regression models were able to identify significant variables in determining health status. Logistic regression models accomplish this by evaluating statistical significance of each variable. The values were averaged over cross-validation folds. The 25 variables with largest absolute effect size selected by each model are shown in Supplementary Table S1 (http://dx.doi.org/10.3168/jds.2015-9840). In general, the data set variables had

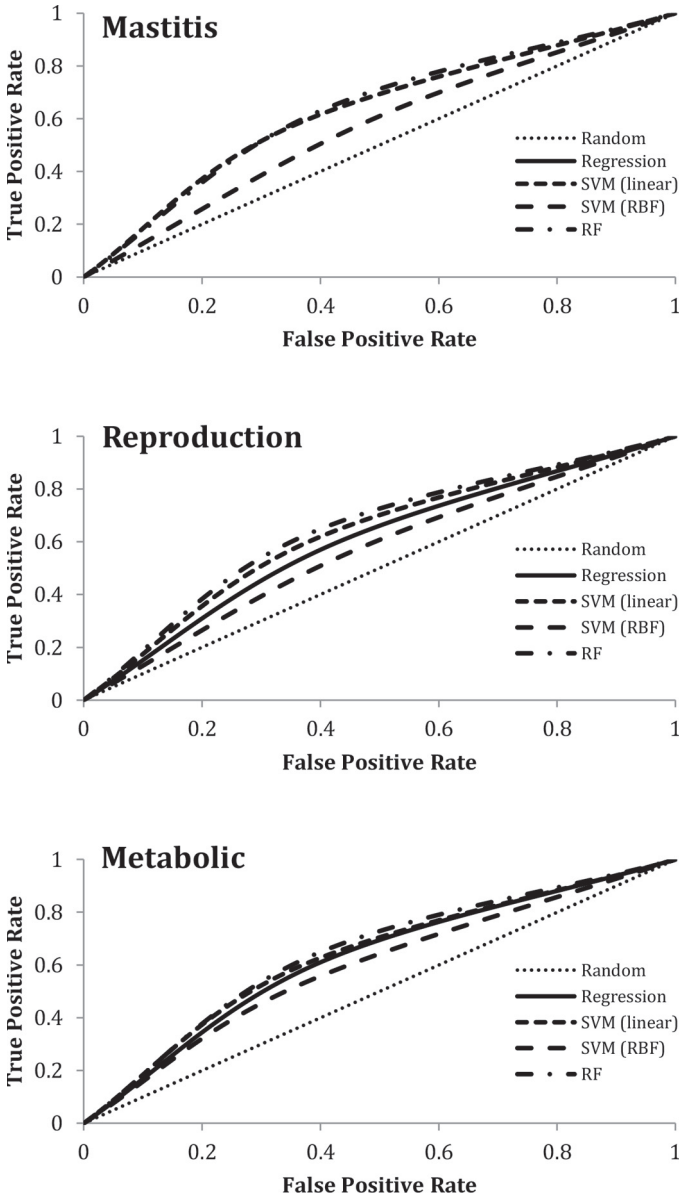

Figure 2. Receiver operating characteristic (ROC) curves for discretized herd incidence of mastitis, reproductive, and metabolic health events averaged across 10-fold cross-validation results.

lengthy, descriptive names. To better discern important factors identified by the models, key words from significant variables were combined in a "word cloud" or weighted list (Fellows, 2014). The size of each key word in the word cloud corresponds to the number of times that particular word was present in significant variables. Word clouds for key words from the top 25 variables selected by logistic regression models are included in the supplementary materials (Supplementary Figure S2; http://dx.doi.org/10.3168/jds.2015-9840).

At the herd level, key words identified most often for mastitis using the logistic regression model included SCS and mastitis, calving, and herd turnover. Herd turnover reflects all cows leaving the herd, not necessarily cows leaving the herd due to illness. Key words identified most often for metabolic events included herd turnover and production stage. Finally, key words identified most often for reproductive events included SCS, herd turnover, and days open. Although there is overlap in variables selected for each category, such as animals leaving the herd, some key words were identified that were intuitively expected (e.g., SCS and mastitis).

Logistic Regression—Individual Level

Logistic regression models were also attempted for records at the individual level. Due to the large size of the data set, a variable selection procedure was performed before stepwise logistic regression to further reduce the dimensionality of the data. The top 100 variables were provided to the stepwise procedure as described for the herd-level analyses. Despite this additional preprocessing, these models still required an unrealistic amount of time to determine a final model (with one model running for more than 90 d without identifying a final model). Due to the amount of time these models required, we determined that fitting these models for practical applications would be unfeasible and results will not be discussed further.

Support Vector Machine—Herd Level

Support vector machine models improved predictive performance over logistic regression for mastitis and reproductive events when data were split based on year, as shown in Table 2. Accuracy of prediction when splitting data based on year of occurrence and fitting a linear kernel was 0.49, 0.47, and 0.55 for mastitis, reproductive, and metabolic events, respectively. Figure 1 includes ROC curves for SVM models with linear and RBF kernels when data were split based on year of occurrence. Accuracy of prediction when data were split based on year increased when using a RBF kernel for mastitis and reproductive events but not metabolic events. Accuracy of prediction was higher for all categories when data were randomly split and a linear kernel was used for the SVM model (Table 3). Evaluation of true positive and negative rate revealed that linear SVM models had higher specificity in all but one case (linear SVM with metabolic events) when data were split based on year, ranging from 0.54 for metabolic events to 0.83 for reproductive events in the prediction data set. This indicates that the models were better able to correctly classify herds with a low incidence of health events. Lower results for sensitivity with the SVM models, ranging from 0.29 for reproductive events to 0.55 for metabolic events in the prediction data split by year, indicate that the models were less capable of identifying herds with high event incidence. The SVM performance indicates that these models may be best utilized in identifying characteristics common to herds with a low incidence of health events. Its utility in identifying herds with a high incidence of health events will be poor due to low model sensitivity. Poor sensitivity could be the result of grouping the diseases into categories. The factors that predispose a cow to one reproductive disorder—cystic ovaries, for example—may not be the same factors that predispose a cow to a different reproductive disorder; for example, metritis.

True cross-validation performance was also evaluated for each SVM model; 10-fold cross-validation was performed and averaged across folds to compare with when data were split based on year. Accuracy (SD) for each health event category when performing true cross-validation is included in Table 3. Figure 2 includes ROC curves from SVM models averaged over the 10-folds used for validation in each health category. Accuracy of prediction for SVM models improved or remained the same for most health categories by using true cross-validation. The exception was the SVM model with RBF kernel for reproductive events. Increase in prediction accuracy may be the result of using more data for training (approximately 90%) and repeating prediction multiple times. The decrease that occurs after training a model and using the model to evaluate a new set of data was less when using true cross-validation compared with when data were validated based on year. This is likely due to the difference in data used for training, but it remains important to be cognizant of this when these models are used with real data. Based on these results, the best approach to estimate model predictive ability in this type of scenario is to perform random k-fold cross-validation.

Support Vector Machine—Individual Level

Models would not converge when data were split based on year, thus SVM model results are not included in

Table 4. Summary of random forest model performance for individual incidence of mastitis, reproductive, and metabolic health events when data are split by year

Health event	Accuracy (Training)	Sensitivity (Training)	Specificity (Training)	Accuracy (Validation)	Sensitivity (Validation)	Specificity (Validation)
Mastitis[1]	0.94	0.84	0.97	0.61	0.11	0.88
Reproductive	0.93	0.75	0.98	0.52	0.26	0.62
Metabolic	0.90	0.63	0.97	0.70	0.01	0.99

[1]Random forest model for each event category used 25 trees.

Table 4. Results from SVM models fit at the individual level are shown in Table 5 for both linear and RBF kernels when data were split randomly. For all categories of health events, specificity was higher than sensitivity, regardless of kernel. This indicates that the SVM models can better identify healthy cows than cows with an incidence of disease. Similar to that reported at the herd level, this could be used to identify characteristics influencing healthy cows. Figures 3 and 4 include ROC curves for SVM models at the individual level.

Random Forest—Herd Level

Finally, RF models were fit to data from each health category to evaluate prediction of a nonparametric tree-based method. For each event category, an optimal number of trees was determined before fitting a final model by testing a range of values. The optimal number of trees was approximately 25 for mastitis and reproductive events and 40 for metabolic events when tested across a range of values, regardless of the cross-validation method utilized. Random forest models when data were split based on year did not improve prediction accuracy above that of models previously discussed. Figure 1 includes ROC curves for RF models fit at the herd level with year-based cross-validation for each health event category. For metabolic health events, the RF model had the highest sensitivity compared with other models. In this case, the RF model was more capable of identifying high-risk herds for metabolic events. Conversely, the RF model had the highest specificity for mastitis and reproductive events when data were validated based on year.

True 10-fold cross-validation was also performed for each random forest model. Accuracy was averaged across each fold and is provided in Table 3, along with standard deviations. The ROC curves for 10-fold cross-validation of RF models at the herd level are shown in Figure 2 for each health event category. As with the SVM models, accuracy of predictive performance improved for all health categories by using true cross-validation. Random forest models for mastitis and reproductive events had the best predictive performance based on accuracy and sensitivity. The metabolic RF model had the highest accuracy and sensitivity, but specificity was found to be lower than in SVM models.

Similarly to logistic regression, random forest models can provide a measure of variable importance using change produced in the Gini index. The Gini index is an alternative performance measure used for classification trees (Breiman et al., 1984). Rather than focusing on accuracy of prediction, the Gini index provides a

Table 5. Summary of model performance for individual incidence of mastitis, reproductive, and metabolic health events averaged across 10-fold cross-validation results fitting support vector machine (SVM) and random forest models (SD in parentheses)

Health event	Accuracy (Training)	Sensitivity (Training)	Specificity (Training)	Accuracy (Validation)	Sensitivity (Validation)	Specificity (Validation)
Mastitis						
SVM (linear) c = 0.01	0.70 (0.001)	0.24 (0.002)	0.88 (0.002)	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)	0.70 (0.01)	0.39 (0.03)	0.83 (0.02)
Random forest[2]	0.93(0.0002)	0.82 (0.001)	0.97 (0.0004)	0.93 (0.001)	0.82 (0.003)	0.97 (0.001)
Reproductive						
SVM (linear) c = 0.005	0.69 (0.001)	0.33 (0.009)	0.79 (0.003)	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)	0.77 (0.01)	0.33 (0.03)	0.88 (0.02)
Random forest	0.92 (0.0002)	0.73 (0.002)	0.97 (0.0002)	0.92 (0.001)	0.74 (0.006)	0.97 (0.0007)
Metabolic						
SVM (linear) c = 0.01	0.77 (0.001)	0.11 (0.02)	0.95 (0.004)	0.76 (0.03)	0.12 (0.03)	0.93 (0.05)
SVM (RBF) c = 10.0	0.75 (0.01)	0.26 (0.01)	0.88 (0.01)	0.75 (0.01)	0.25 (0.02)	0.88 (0.01)
Random forest	0.89 (0.0002)	0.61 (0.003)	0.97 (0.001)	0.87 (0.061)	0.57 (0.145)	0.96 (0.04)

[1]SVM (linear) = support vector machine with linear kernel; SVM (RBF) = support vector machine with radial basis function kernel; c = regularization parameter.

[2]Random forest model for each event category used 25 trees.

measure of "node purity" (Kuhn and Johnson, 2013). A variable that results in a high decrease in Gini index plays a larger role in partitioning the data. The top 25 variables with the greatest mean decrease in Gini index were identified from each random forest model. Their importance measures are depicted graphically in Figures 5, 6, and 7 for mastitis, reproductive, and metabolic events, respectively.

Variables identified as important in the RF model indicate those that are influential in herd incidence of each respective health event category. Across all health event categories, environmental characteristics such as temperature and weather were identified. Although environmental conditions such as temperature and weather cannot be controlled, measures can be taken by producers to minimize the effect of these factors. For example, several methods have been identified to reduce

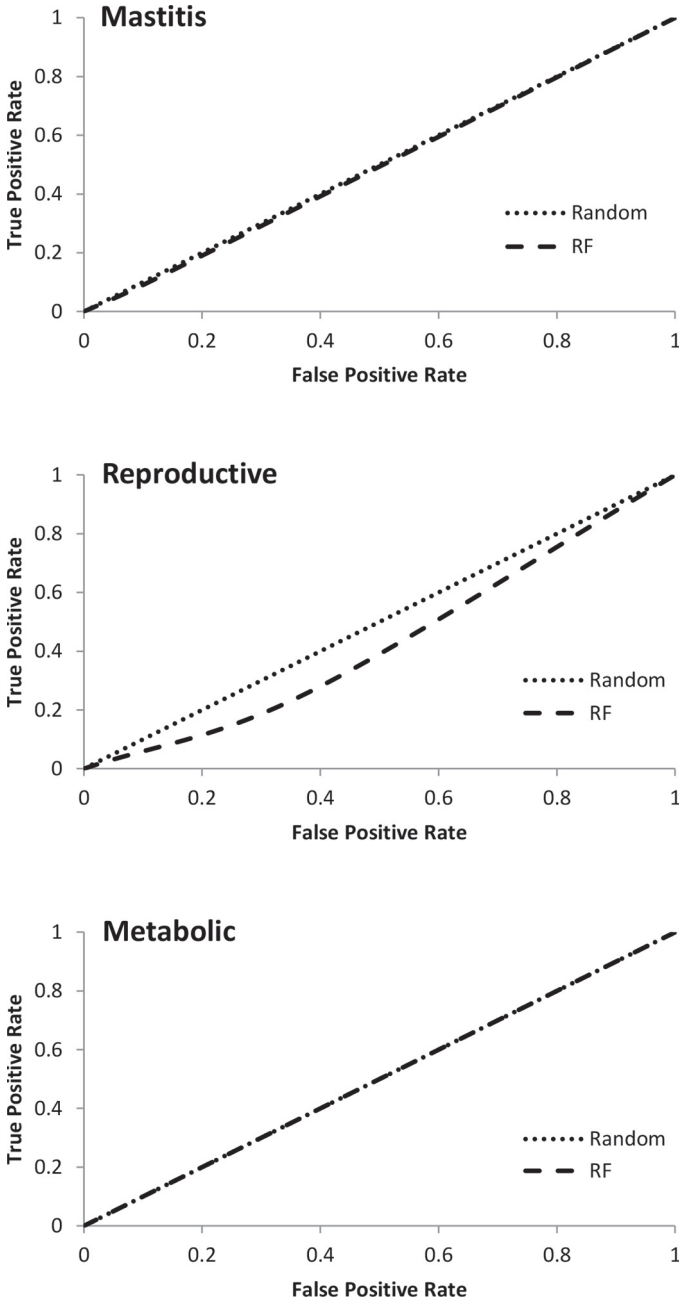

Figure 3. Receiver operating characteristic (ROC) curves for individual incidence of mastitis, reproductive, and metabolic health events when data are split by year.

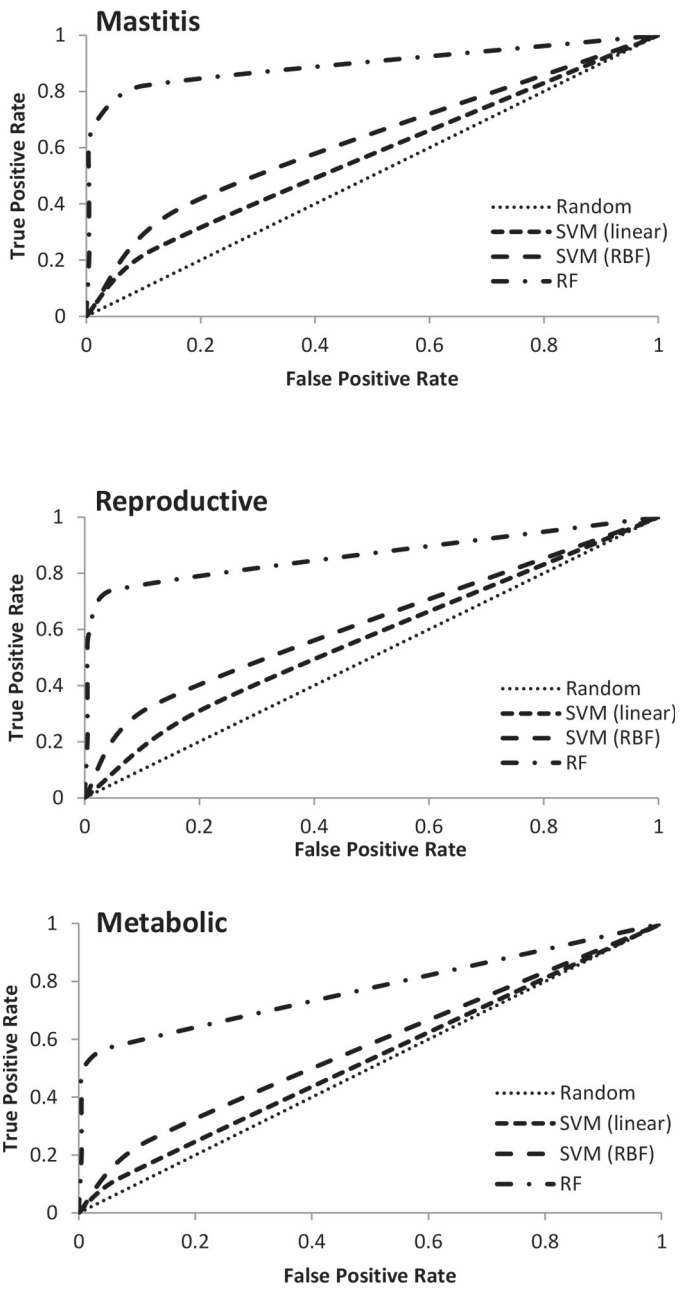

Figure 4. Receiver operating characteristic (ROC) curves for individual incidence of mastitis, reproductive, and metabolic health events averaged across 10-fold cross-validation results.

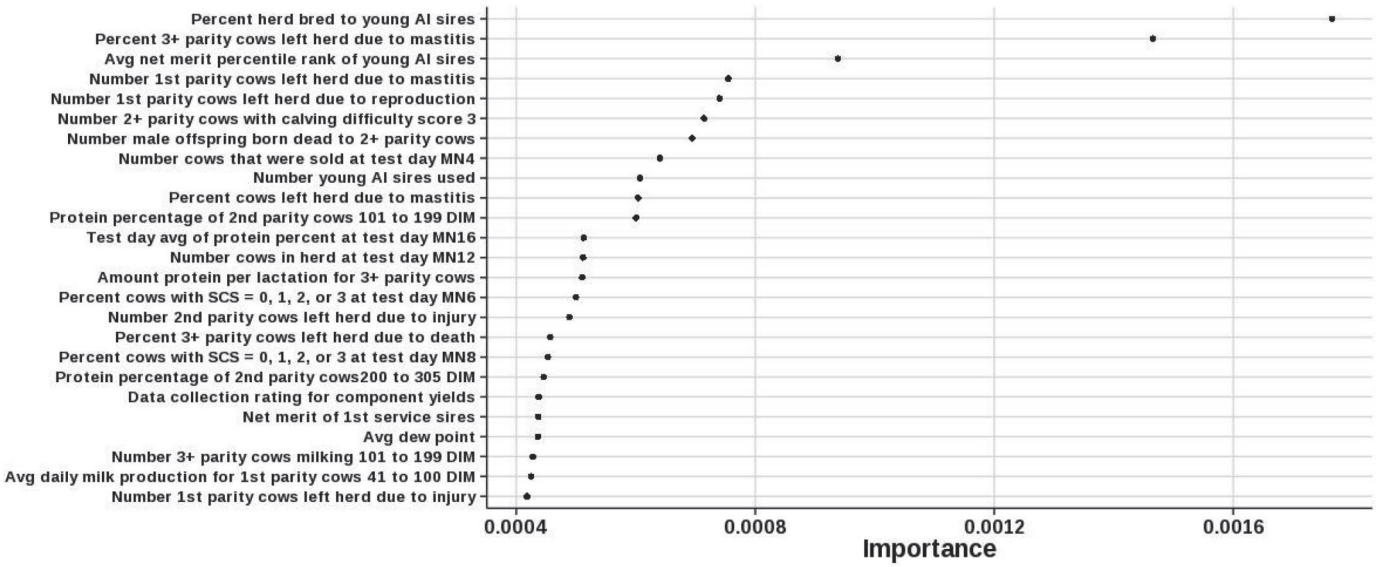

Figure 5. Variable importance plot for the 25 most important variables determined by random forest models of mastitis at the herd level. Importance averaged over cross-validation folds. Avg = average; MN# = previous test day in DHI-202 Herd Summary (http://www.drms.org/PDF/materials/202Fact.pdf).

the effects of heat stress (Bucklin et al., 1991). Influence of heat stress on reproductive performance has been identified previously by several authors (Wilson et al., 1998; Ravagnolo and Misztal, 2002; Caraviello et al., 2006). Additional key words identified in the RF model for mastitis included protein, mastitis, and herd turnover. Key words identified in the RF model for metabolic events at herd level included herd turnover, third-parity cows, and number of cows. Number of cows within a herd has been previously identified as a significant risk factor in the incidence of metabolic events including ketosis and displaced abomasum (Stengärde et al., 2012). This may be indicative of an underlying risk factor, such as less time spent per cow in larger herds (Agger and Alban, 1996). Higher parity has also been identified as a risk factor for metabolic diseases by

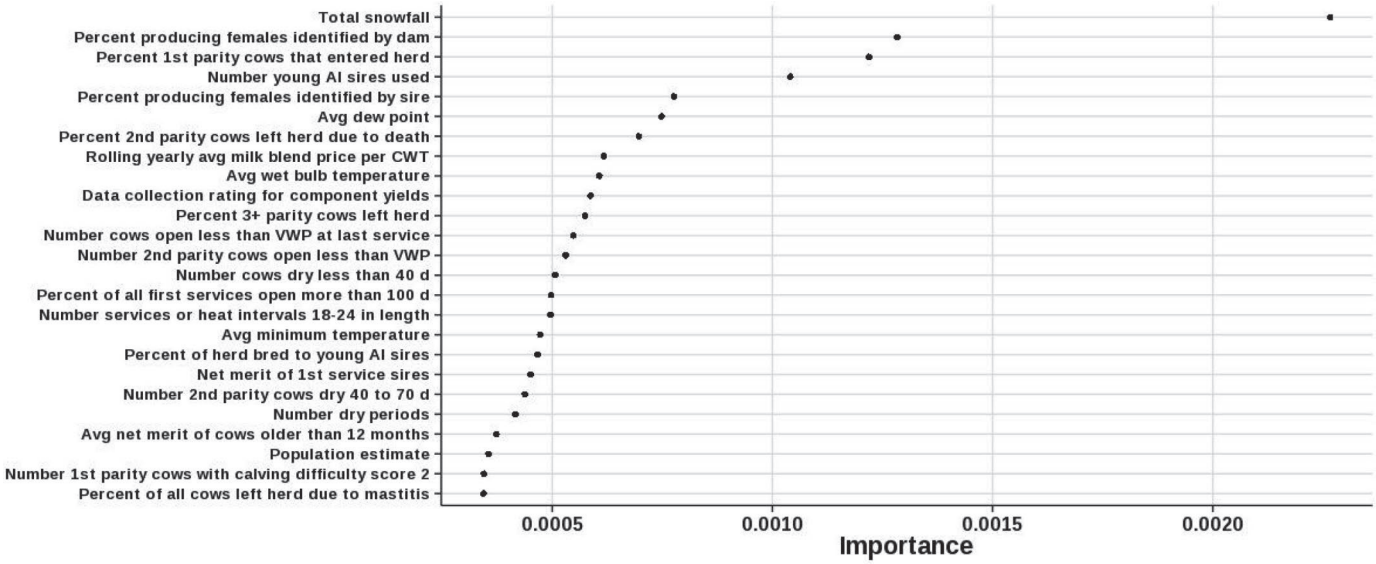

Figure 6. Variable importance plot for the 25 most important variables determined by random forest models of reproductive events at the herd level. Importance averaged over cross-validation folds. Avg = average; CWT = hundred-weight; VWP = voluntary waiting period.

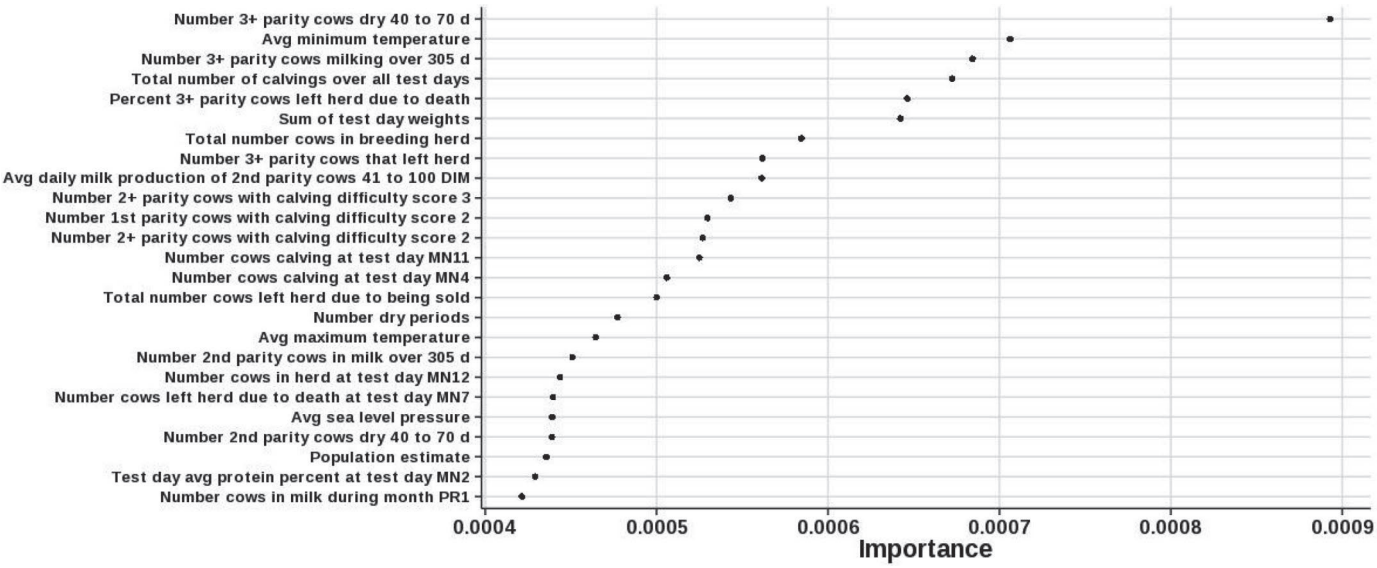

Figure 7. Variable importance plot for the 25 most important variables determined by random forest models of metabolic events at the herd level. Importance averaged over cross-validation folds. Avg = average; MN# = previous test day in DHI-202 Herd Summary (http://www.drms.org/PDF/materials/202Fact.pdf); PR# = month in DHI-202 Herd Summary.

other authors (Rasmussen et al., 1999). The RF model for reproductive health events at the herd level identified PTA and third or greater parity cows most often. It should be noted that the key word "PTA" reflected several variables, including net merit of service sires and average net merit of cows; it does not reflect PTA for a particular trait. Reproductive problems are a well-known reason why cows are removed from the herd (De Vries et al., 2010).

Random Forest—Individual Level

For each event category, an optimal number of trees was determined before fitting a final model. The optimal number of trees was approximately 25 for all health categories when tested across a range of values, regardless of cross-validation technique. High accuracy was attained for all health events when training data with RF models. Receiver operating characteristic curves for these models with year-based cross-validation are included in Figure 3. Receiver operating characteristic curves from 10-fold cross-validation are shown in Figure 4. When models were validated based on year, however, accuracy decreased drastically for prediction, with prediction accuracy ranging from 0.52 to 0.70. This was mostly due to poor sensitivity, ranging from 0.01 for metabolic events to 0.26 for reproductive events. Specificity was higher for all models, indicating that these models were better capable of identifying healthy cows. Poor prediction accuracy when data are split based on year may indicate that risk in previous years is not highly correlated with risk in later years. It will be important to remain cognizant of this for optimization of data collection.

High predictive accuracy was obtained for all health event categories when RF models were applied to individual data and cross-validation was performed across 10 randomized folds. Predictive accuracy was highest for mastitis, at 0.93 (SD = 0.001); lowest prediction accuracy was for metabolic events, at 0.87 (SD = 0.061; Table 5). Prediction accuracy was higher when data were validated using 10-folds compared with validation based on year. Overall, sensitivity was lower than specificity; however, the sensitivity was highest for RF models compared with the other methods. Improved predictive ability can be seen from the ROC curves for RF models at the individual level in Figure 4.

The same procedure was conducted for RF models at the individual level to ascertain variable importance. Variable importance plots including the top 25 variables are shown in Figures 8, 9, and 10 for mastitis, reproductive, and metabolic events, respectively. Similar to the logistic regression models, word clouds were constructed with key words from variables selected in RF models for each health event category in the same manner as the stepwise regression models (see Supplementary Figures S3 and S4; http://dx.doi.org/10.3168/jds.2015-9840). The number of cows was identified for all health events at the individual level. As stated previously, this could be indicative of an underlying risk factor due to the size of herd. Herd turnover was also identified for all health events. This was for a range of

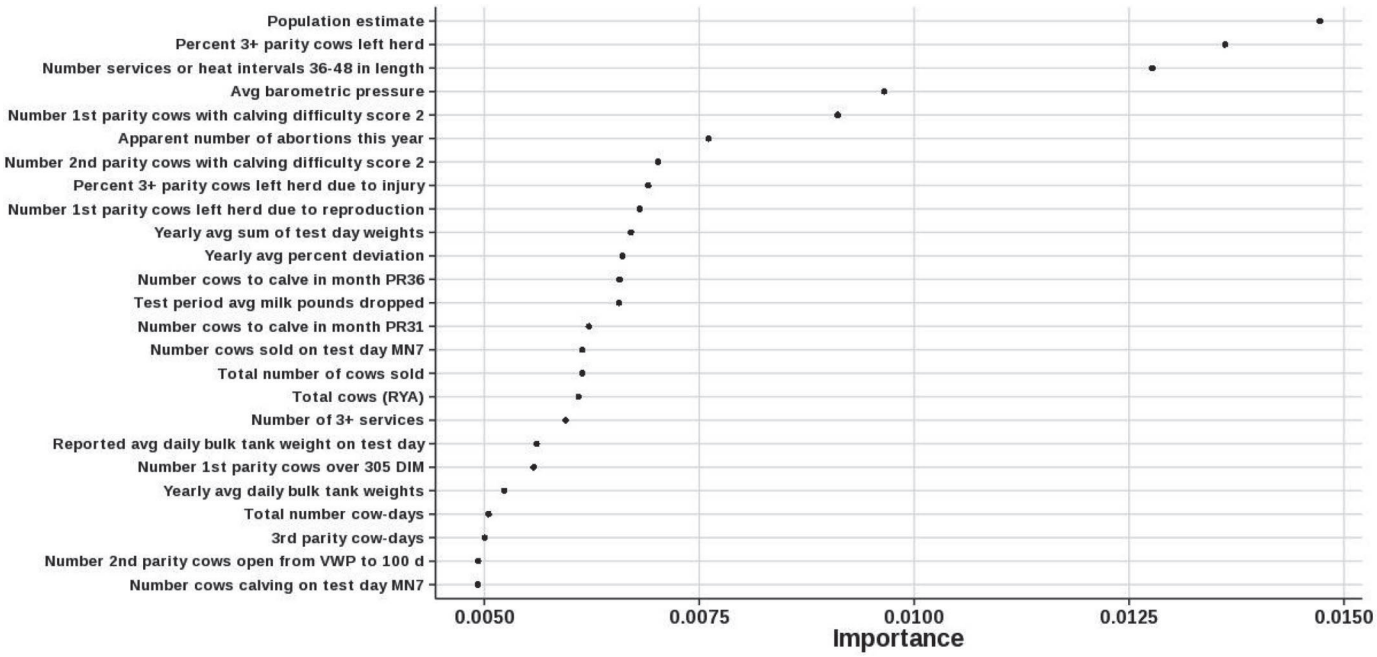

Figure 8. Variable importance plot for the 25 most important variables determined by random forest models of mastitis at the individual level. Importance averaged over cross-validation folds. Avg = average; MN# = previous test day in DHI-202 Herd Summary (http://www.drms.org/PDF/materials/202Fact.pdf); PR# = month in DHI-202 Herd Summary; RYA = rolling yearly average; VWP = voluntary waiting period.

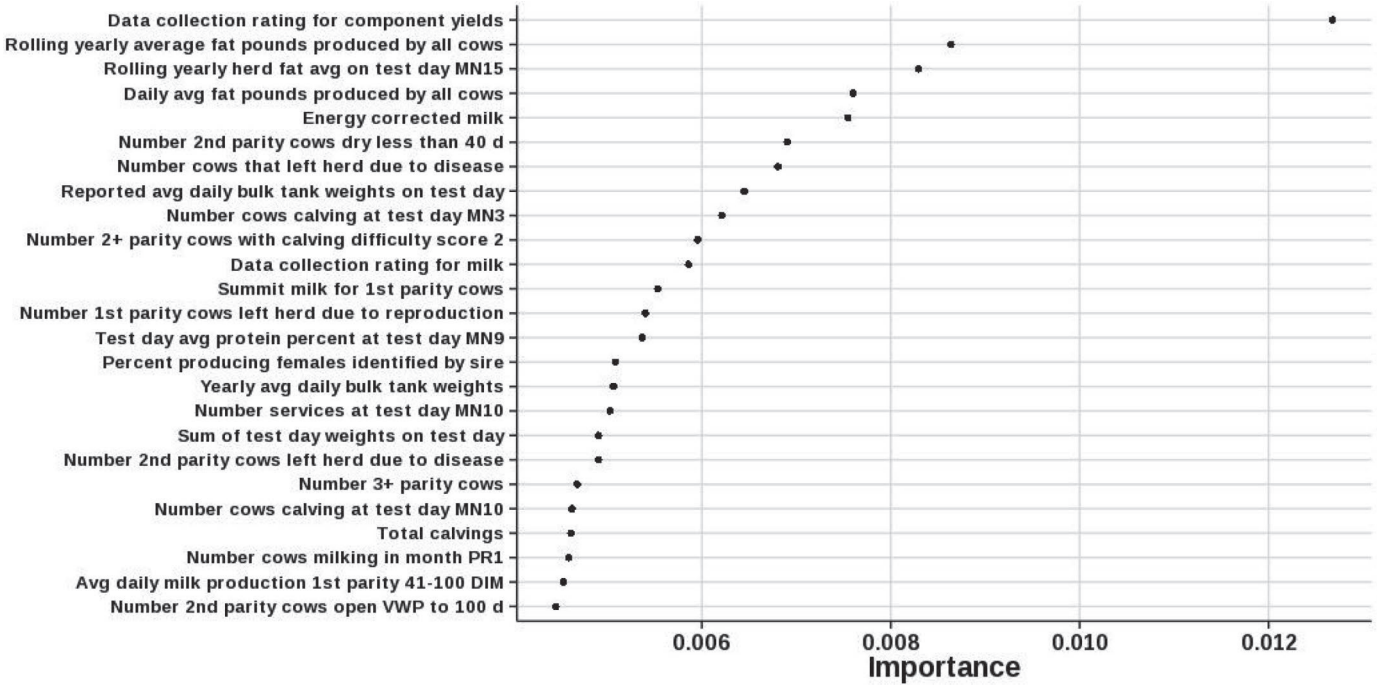

Figure 9. Variable importance plot for the 25 most important variables determined by random forest models of reproductive events at the individual level. Importance averaged over cross-validation folds. Avg = average; MN# = previous test day in DHI-202 Herd Summary (http://www.drms.org/PDF/materials/202Fact.pdf); PR# = month in DHI-202 Herd Summary; VWP = voluntary waiting period.

reasons such as being sold, reproductive problems, or mastitis. Additional key words identified for mastitis at the individual level included milk production and calving. Key words identified frequently for reproductive events also included milk production and calvings. Additional key words identified for metabolic events included fat production, first parity, and milk production.

General Discussion

For all models investigated herein, overall accuracy of prediction improved when fitting the model at the individual level compared with at the herd level. This may be the result of additional processing applied to herd-level data (converting actual incidence to a binary indicator). An alternative to converting the herd-level data to binary data would be to fit regression models. This was not done, however, because the goal was not to predict herd incidence numerically.

Each of the models investigated has benefits and disadvantages aside from their ability to predict health status. Logistic regression cannot easily handle missing values, thus the need for imputing missing values before analysis. Logistic regression also tends to be much more time consuming, even when the number of variables being incorporated is decreased before model fitting. Logistic regression is more easily understood than nonparametric models and it is able to identify influential variables. Support vector machines are a much more flexible class of models compared with logistic regression. Several different kernels can be used when fitting an SVM model. These models do require estimation of tuning parameters, such as penalty or cost. Results can be more difficult to interpret than in logistic regression models as well. Support vector machine models also cannot easily handle missing values, aside from imputation or ignoring missing records. Unlike logistic regression, however, SVM models were able to accept all variables within the data set without having to first perform some sort of feature selection. The most flexible models investigated herein were RF models. These models were able to easily handle missing variables, as well as handle a large number of variables. Random forest models can also identify influential variables. One disadvantage of RF models is that they can be more difficult to interpret than a single decision tree, but tend to have better predictive performance, as they are the result of averaging over many decision trees. As observed from measurements of predictive ability, RF models performed the best overall.

Predictive ability of models investigated herein also depended upon the cross-validation method. In general, true 10-fold cross-validation results resulted in less of a decline in predictive ability when applied to validation data. As previously mentioned, this may be the result of training on more data (90 vs. 75%) multiple times when performing true 10-fold cross-validation. However, several scenarios were tested using data split based on year with a ratio of 90% training and 10%

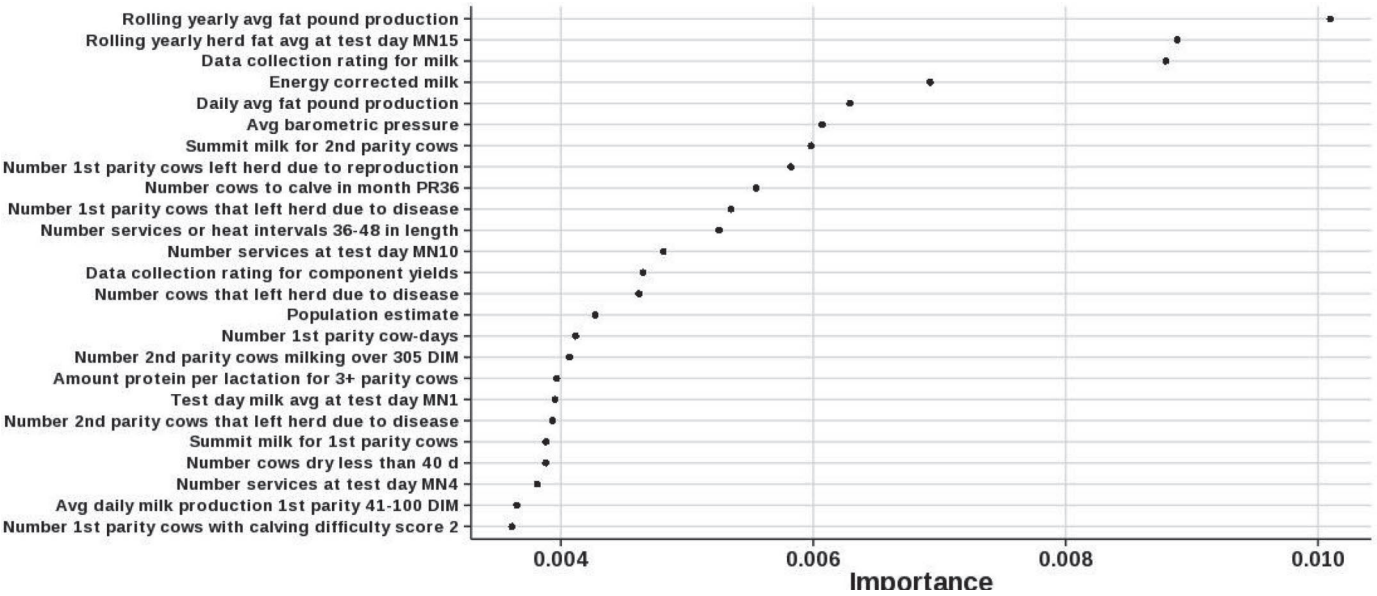

Figure 10. Variable importance plot for the 25 most important variables determined by random forest models of metabolic events at the individual level. Importance averaged over cross-validation folds. Avg = average; MN# = previous test day in DHI-202 Herd Summary (http://www.drms.org/PDF/materials/202Fact.pdf); PR# = month in DHI-202 Herd Summary.

validation. No significant difference in performance was found. This may indicate that the risk in previous years is lowly correlated with risk in later years. There could also be an unrelated trend occurring over time that true cross-validation was able to break down but splitting based on year was not. These models seem to have difficulty predicting a trend in data versus predicting the current scenario. Models developed specifically for identifying changing trends may prove advantageous in this case and are an area for future research (e.g., Nikolov, 2012). Another alternative approach would be to use herds with proven high-quality data to train the models. Regardless, we expect that as more data are collected and incorporated, differences between cross-validation methods will be minimized.

An interesting "side effect" of logistic regression and RF models is the ability to identify the variables affecting the trait of interest. It is important to note, however, that speculating on the significance of different variables, albeit interesting, is complicated by the fact that some of these associations may be transient and not causative. With that said, logistic regression and RF models identified several of the same key factors for each health event category. This lends further support to the importance of these factors for their respective health event. Mastitis models identified herd turnover as a key factor. Common factors for metabolic events included leaving the herd and production. Reproductive models identified production and days open as being associated with incidence of reproductive health events. A relationship between reproductive health events and (increased) days open has been identified by other researchers (e.g., Cobo-Abreu et al., 1979; Lee et al., 1989; Fourichon et al., 2000). Future research using longitudinal models should allow for a better understanding of the effect of individual variables. Producers could use the variables identified as influential in determining health status to more closely monitor the herd if they have specific health concerns. Consultants could also use the identified variables to enhance their advice and suggestions given to producers based on field experience and knowledge.

Ultimately, routinely recorded herd data could be incorporated into herd management strategies to alert producers to potential problems. This research suggests that benchmarking of herd health is feasible with routinely collected data; however, further research is needed. As previously mentioned, models developed to identify and predict trends may have improved performance when dealing with dynamic data such as these. Time-series classification models have been successfully applied to predict "trending" topics on social media sites (Nikolov, 2012). Improvement in predictive ability may also be possible by modeling individual health events as opposed to grouping events into categories. The factors that predispose a cow to retained placenta, for example, may not be the same as the factors that increase a cow's risk of cystic ovaries. An alternative to the approach taken here would be to select certain well-recorded and economically important health events to be monitored for unfavorable changes. Further development and incorporation of predictive models into herd management routines will further improve dairy herd health.

CONCLUSIONS

Results of these analyses indicate that machine learning algorithms, specifically random forest, can be used to accurately identify herds and cows likely to experience a health event of interest. Random forest models had higher predictive ability than parametric models typically used to identify characteristics affecting health status. Random forest models also outperformed support vector machine models. Influential variables were identified for each health event with logistic regression and random forest models. Among identified variables, herd turnover, milk production, parity, and weather conditions were selected, regardless of health event category. Based on these data, our results provide evidence for the feasibility of utilizing routinely recorded herd data to predict herd health status.

ACKNOWLEDGMENTS

The authors thank Dairy Records Management Systems (Raleigh, NC) and the Council on Dairy Cattle Breeding (Reynoldsburg, OH) for providing the data used in this study, as well as Cai Li and Ravi Mathur from the Bioinformatics Department at North Carolina State University for their contributions with initial model selection. J. B. Cole was supported by appropriated project 1265-31000-096-00, "Improving Genetic Predictions in Dairy Animals Using Phenotypic and Genomic Information," of the Agricultural Research Service of the United States Department of Agriculture. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. The USDA is an equal opportunity provider and employer.

REFERENCES

Agger, J., and L. Alban. 1996. Welfare in Danish dairy herds 3. Health management and general routines in 1983 and 1994. Acta Vet. Scand. 37:79–97.

Breen, J. 2012. zipcode: U.S. ZIP Code database for geocoding. R package version 1.0. http://CRAN.R-project.org/package=zipcode.

Breiman, L. 2001. Random forests. Mach. Learn. 45:5–32.
Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. 1984. Classification and regression trees. CRC Press, Boca Raton, FL.
Bucklin, R. A., L. W. Turner, D. K. Beede, D. R. Bray, and R. W. Hemken. 1991. Methods to relieve heat stress for dairy cows in hot, humid climates. Appl. Eng. Agric. 7:241–247.
Caraviello, D. Z., K. A. Weigel, M. Craven, D. Gianola, N. B. Cook, K. V. Nordlund, P. M. Fricke, and M. C. Wiltbank. 2006. Analysis of reproductive performance of lactating cows on large dairy farms using machine learning algorithms. J. Dairy Sci. 89:4703–4722. http://dx.doi.org/10.3168/jds.S0022-0302(06)72521-8.
Cobo-Abreu, R., S. W. W. Martin, J. B. Stone, and R. A. Willoughby. 1979. The rates and patterns of survivorship and disease in a university dairy herd. Can. Vet. J. 20:177–183.
Coppa, M., A. Ferlay, C. Chassaing, C. Agabriel, F. Glasser, Y. Chilliard, G. Borreani, R. Barcarolo, T. Baars, D. Kusche, O. M. Harstad, J. Verbič, J. Golecký, and B. Martin. 2013. Prediction of bulk milk fatty acid composition based on farming practices collected through on-farm surveys. J. Dairy Sci. 96:4197–4211. http://dx.doi.org/10.3168/jds.2012-6379.
Correa, M. T., C. R. Curtis, H. N. Erb, J. M. Scarlett, and R. D. Smith. 1990. An ecological analysis of risk factors for postpartum disorders of Holstein-Friesian cows from thirty-two New York farms. J. Dairy Sci. 73:1515–1524. http://dx.doi.org/10.3168/jds.S0022-0302(90)78819-4.
Dairy Records Management Systems. 2014. DHI-202 Herd Summary. http://www.drms.org/PDF/materials/202Fact.pdf.
De Vries, A., J. D. Olson, and P. J. Pinedo. 2010. Reproductive risk factors for culling and productive life in large dairy herds in the eastern United States between 2001 and 2006. J. Dairy Sci. 93:613–623. http://dx.doi.org/10.3168/jds.2009-2573.
Dechow, C. D., and R. C. Goodling. 2008. Mortality, culling by sixty days in milk, and production profiles in high- and low-survival Pennsylvania herds. J. Dairy Sci. 91:4630–4639. http://dx.doi.org/10.3168/jds.2008-1337.
Dhakal, K., F. Tiezzi, J. S. Clay, and C. Maltecca. 2015. Inferring causal relationships between reproductive and metabolic health disorders and production traits in first-lactation US Holsteins using recursive models. J. Dairy Sci. 98:2713–2726.
Diamond, H. J., T. R. Karl, M. A. Palecki, C. B. Baker, J. E. Bell, R. D. Leeper, D. R. Easterling, J. H. Lawrimore, T. P. Meyers, M. R. Helfert, G. Goodge, and P. W. Thorne. 2013. U.S. Climate Reference Network after one decade of operations: Status and assessment. Bull. Am. Meteorol. Soc. 94:489–498. http://dx.doi.org/10.1175/BAMS-D-12-00170.1.
Enevoldsen, C., J. Hindhede, and T. Kristensen. 1996. Dairy herd management types assessed from indicators of health, reproduction, replacement, and milk production. J. Dairy Sci. 79:1221–1236.
Esposito, G., P. C. Irons, E. C. Webb, and A. Chapwanya. 2014. Interactions between negative energy balance, metabolic diseases, uterine health and immune response in transition dairy cows. Anim. Reprod. Sci. 144:60–71. http://dx.doi.org/10.1016/j.anireprosci.2013.11.007.
Fawcett, T. 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27:861–874. http://dx.doi.org/10.1016/j.patrec.2005.10.010.
Faye, B. 1992. Interrelationships between health status and farm management system in French dairy herds. Prev. Vet. Med. 12:133–152. http://dx.doi.org/10.1016/0167-5877(92)90076-r.
Fellows, I. 2014. wordcloud: Word Clouds. R package version 2.5. http://CRAN.R-project.org/package=wordcloud.
Fourichon, C., H. Seegers, and X. Malher. 2000. Effect of disease on reproduction in the dairy cow: A meta-analysis. Theriogenology 53:1729–1759. http://dx.doi.org/10.1016/S0093-691X(00)00311-3.
González-Recio, O., and S. Forni. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genet. Sel. Evol. 43:7 http://dx.doi.org/10.1186/1297-9686-43-7.
Green, M. J., A. J. Bradley, G. F. Medley, and W. J. Browne. 2007. Cow, farm, and management factors during the dry period that determine the rate of clinical mastitis after calving. J. Dairy Sci. 90:3764–3776. http://dx.doi.org/10.3168/jds.2007-0107.
Grolemund, G., and H. Wickham. 2011. Dates and times made easy with lubridate. J. Stat. Softw. 40:1–25.
Hijmans, R.J., E. Williams, and C. Vennes. 2012. geosphere: Spherical Trigonometry. R package version 1.2-28. http://CRAN.R-project.org/package=geosphere.
Hill, A. E., D. A. Dargatz, B. A. Wagner, and A. L. Green. 2009. Relationship between herd size and annual prevalence of and primary antimicrobial treatments for common diseases on dairy operations in the United States. Prev. Vet. Med. 88:264–277. http://dx.doi.org/10.1016/j.prevetmed.2008.12.001 http://hdl.handle.net/10113/38499.
Husson, F., and J. Josse. 2012. missMDA: Handling missing values with/in multivariate data analysis (principal component methods). R package version 1.8.2. http://CRAN.R-project.org/package=missMDA.
Joachims, T. 2006. Training linear SVMs in linear time. Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD '06). ACM Press, New York, NY.
Kelton, D. F., K. D. Lissemore, and R. E. Martin. 1998. Recommendations for recording and calculating the incidence of selected clinical diseases of dairy cattle. J. Dairy Sci. 81:2502–2509. http://dx.doi.org/10.3168/jds.S0022-0302(98)70142-0.
Kuhn, M. 2013. caret: Classification and Regression Training. R package version 5.17-7. http://CRAN.R-project.org/package=caret.
Kuhn, M., and K. Johnson. 2013. Applied Predictive Modeling. Springer, New York, NY.
Lee, L. A., J. D. Ferguson, and D. T. Galligan. 1989. Effect of disease on days open assessed by survival analysis. J. Dairy Sci. 72:1020–1026. http://dx.doi.org/10.3168/jds.S0022-0302(89)79197-9.
Lim, A., L. Breiman, and A. Cutler. 2014. bigrf: Big Random Forests: Classification and Regression Forests for Large Data Sets. R package version 0.1-11. http://CRAN.R-project.org/package=bigrf.
Löf, E., H. Gustafsson, and U. Emanuelson. 2007. Associations between herd characteristics and reproductive efficiency in dairy herds. J. Dairy Sci. 90:4897–4907. http://dx.doi.org/10.3168/jds.2006-819.
Nikolov, S. 2012. Trend or No Trend: A Novel Nonparametric Method for Classifying Time Series. Massachusetts Institute of Technology, Cambridge, MA.
Nyman, A.-K., T. Ekman, U. Emanuelson, A. H. Gustafsson, K. Holtenius, K. P. Waller, and C. H. Sandgren. 2007. Risk factors associated with the incidence of veterinary-treated clinical mastitis in Swedish dairy herds with a high milk yield and a low prevalence of subclinical mastitis. Prev. Vet. Med. 78:142–160. http://dx.doi.org/10.1016/j.prevetmed.2006.10.002.
Parker Gaddis, K. L., J. B. Cole, J. S. Clay, and C. Maltecca. 2012. Incidence validation and relationship analysis of producer-recorded health event data from on-farm computer systems in the United States. J. Dairy Sci. 95:5422–5435. http://dx.doi.org/10.3168/jds.2012-5572.
Parker Gaddis, K. L., J. B. Cole, J. S. Clay, and C. Maltecca. 2014. Genomic selection for producer-recorded health event data in US dairy cattle. J. Dairy Sci. 97:3190–3199. http://dx.doi.org/10.3168/jds.2013-7543.
R Core Team. 2014. R: A Language and Environment for Statistical Computing. R Core Team, Vienna, Austria.
R Development Core Team. 2012. R: A Language and Environment for Statistical Computing. R Core Team, Vienna, Austria.
Rasmussen, L. K., B. L. Nielsen, J. E. Pryce, T. T. Mottram, and R. F. Veerkamp. 1999. Risk factors associated with the incidence of ketosis in dairy cows. Anim. Sci. 68:379–386.
Ravagnolo, O., and I. Misztal. 2002. Effect of heat stress on nonreturn rate in Holsteins: Fixed model analyses. J. Dairy Sci. 85:3101–3106.
Sato, K., P. C. Bartlett, L. Alban, J. F. Agger, and H. Houe. 2008. Managerial and environmental determinants of clinical mastitis in Danish dairy herds. Acta Vet. Scand. 50:4 http://dx.doi.org/10.1186/1751-0147-50-4.
Schefers, J. M., K. A Weigel, C. L. Rawson, N. R. Zwald, and N. B. Cook. 2010. Management practices associated with conception rate and service rate of lactating Holstein cows in large, com-

mercial dairy herds. J. Dairy Sci. 93:1459–1467. http://dx.doi.org/10.3168/jds.2009-2015.
Simensen, E., O. Østerås, K. E. Bøe, C. Kielland, L. E. Ruud, and G. Naess. 2010. Housing system and herd size interactions in Norwegian dairy herds; associations with performance and disease incidence. Acta Vet. Scand. 52:14 http://dx.doi.org/10.1186/1751-0147-52-14.
Stengärde, L., J. Hultgren, M. Tråvén, K. Holtenius, and U. Emanuelson. 2012. Risk factors for displaced abomasum or ketosis in Swedish dairy herds. Prev. Vet. Med. 103:280–286. http://dx.doi.org/10.1016/j.prevetmed.2011.09.005.
Sullivan, R. 2012. Introduction to Data Mining for the Life Sciences. Springer, New York, NY.
Svensson, C., A.-K. Nyman, K. Persson Waller, and U. Emanuelson. 2006. Effects of housing, management, and health of dairy heifers on first-lactation udder health in southwest Sweden. J. Dairy Sci. 89:1990–1999.
United States Census Bureau. 2012. Methodology for the Intercensal Population and Housing Unit Estimates: 2000 to 2010. Accessed Feb. 10, 2014. https://www.census.gov/popest/methodology/2000-2010_Intercensal_Estimates_Methodology.pdf.
Wilson, S. J., R. S. Marion, J. N. Spain, D. E. Spiers, D. H. Keisler, and M. C. Lucy. 1998. Effects of controlled heat stress on ovarian function of dairy cattle. 1. Lactating cows. J. Dairy Sci. 81:2124–2131.
Windig, J. J., M. P. L. Calus, B. Beerda, and R. F. Veerkamp. 2006. Genetic correlations between milk production and health and fertility depending on herd environment. J. Dairy Sci. 89:1765–1775. http://dx.doi.org/10.3168/jds.S0022-0302(06)72245-7.
Windig, J. J., M. P. L. Calus, and R. F. Veerkamp. 2005. Influence of herd environment on health and fertility and their relationship with milk production. J. Dairy Sci. 88:335–347. http://dx.doi.org/10.3168/jds.S0022-0302(05)72693-X.
Zwald, N. R., K. A. Weigel, Y. M. Chang, R. D. Welper, and J. S. Clay. 2004. Genetic selection for health traits using producer-recorded data. I. Incidence rates, heritability estimates, and sire breeding values. J. Dairy Sci. 87:4287–4294.