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.