



## Practical challenges and potential approaches to predicting low-incidence diseases on farm using individual cow data: A clinical mastitis example

D. M. Liebe,<sup>1</sup> N. M. Steele,<sup>2</sup> C. S. Petersson-Wolfe,<sup>3</sup> A. De Vries,<sup>4</sup> and R. R. White<sup>1\*</sup>

<sup>1</sup>Department of Animal and Poultry Sciences, Virginia Tech, Blacksburg 24060

<sup>2</sup>DairyNZ Ltd., Private Bag 3221, Hamilton, 3240, New Zealand

<sup>3</sup>Department of Dairy Science, Virginia Tech, Blacksburg 24060

<sup>4</sup>Department of Animal Sciences, University of Florida, Gainesville 32611

### ABSTRACT

Clinical mastitis (CM) incidence is considerable in terms of cows affected per year, but cases are much less common in terms of detections per cow per milking. From a modeling perspective, where predictions are made every time any cow is milked, low CM incidence per cow day makes training, evaluating, and applying CM prediction models a challenge. The objective of this study was to build models for predicting CM incidence using time-series sensor data and choose models that maximize net return based on a cost matrix. Data collected from 2 university dairy farms, the University of Florida and Virginia Polytechnic Institute and State University, were used to gather representative data, including 110,156 milkings and 333 CM cases. Variables used in the models were milk yield, protein, lactose, fat, electrical conductivity, days in milk, lactation number, and activity as the number of steps, lying time, lying bouts, and lying bout duration. Models that predicted either likelihood of CM caused by gram-negative (GN) or gram-positive (GP) bacteria on each day were derived using extreme gradient boosting with weighting favoring true-positive cases, logistic responses, and log-loss errors. Model accuracies were determined using data randomly held out from the training set on each run. All variables considered were in terms of change (slope) over previous days, including the day CM was visually detected. The GN models had a median sensitivity (Se) of 52.6% and specificity (Sp) of 99.8%, whereas the GP models had a median Se of 37.5% and Sp of 99.9% when tested on the held-out data. In our models optimized to reduce cost from predictions, the Se was much less than Sp, suggesting that CM models might benefit from greater model weighting placed on Sp. Results also highlight the importance of positive

predictive value (true positive cases per predicted positive case) along with Sp and Se, as models built on sparse data tend to predict too many false-positive cases. The calculated partial net return of our GN and GP models were −\$0.15 and −\$0.10 per cow per lactation, respectively, whereas International Organization for Standardization (ISO) standard models with Se of 80% and Sp of 99% would return −\$1.32 per cow per lactation. Models chosen that minimized the cost to the farmer differed markedly from models that met ISO guidelines, showing asymmetry in targets between Sp and Se when the disease incidence rate is low. Because of the unique challenges that low-incidence diseases like CM present, we recommend that future CM predictive models consider the economic and practical implications in addition to the traditional model evaluation metrics.

**Key words:** clinical mastitis, prediction, model

### INTRODUCTION

Costs associated with mastitis in the dairy cow are estimated to be \$2 billion in the United States annually. Despite the high cost, mastitis incidence in the United States is estimated to be between 25 and 41 cases per 100 cows per lactation, equivalent to roughly one clinical mastitis (CM) case per 890 (Pol and Ruegg, 2007) to 1,460 (USDA, 2014) days per cow (0.07% to 0.1%). This small proportion of cases relative to healthy checks poses a challenge to modelers attempting to predict CM. Historically, the model usefulness or appropriateness in predicting mastitis has been evaluated based on 2 outcomes: sensitivity (Se) and specificity (Sp; Sargeant et al., 2001; Pyörälä, 2003; Koskinen et al., 2009; Ganda et al., 2016; Khatun et al., 2017):

$$\text{Se} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}, \text{ and}$$

$$\text{Sp} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}.$$

Received February 15, 2021.

Accepted October 16, 2021.

\*Corresponding author: [rrwhite@vt.edu](mailto:rrwhite@vt.edu)

The International Standard ISO/FDIS 20966 (automatic milking installations-requirements and testing) of the International Organization for Standardization (ISO) includes an annex describing a minimum requirement of 80% Se with a Sp greater than 99% for approved mastitis detection systems (Hogeveen et al., 2010). Authors commonly report Sp and Se in the mastitis detection literature as a gold standard for minimum model performance (Hogeveen et al., 2010). As CM is a disease with low incidence, detection using a model with ISO standard performance in terms of Sp and Se will produce more false alerts than a model that predicts a more prevalent disease. Consider 2 disease models for diseases with 50% and 5% prevalence, both with 80% Se and 99% Sp. The first would produce 80 true positives for every false positive, whereas the other only about 4.2 true positives per false positive. Error rate, or the rate of false positives per positive prediction, is a function of a model's Sp and Se, but also the disease prevalence. Dominiak and Kristensen (2017) show how any arbitrary error rate can be achieved with fixed Se and Sp, simply by lowering the prevalence. This idea that prevalence is an important factor in evaluating disease prediction models, along with Se and Sp, is incongruous with the practice of using only Se and Sp as gold standards. Furthermore, by ignoring prevalence in cases of low prevalence diseases such as CM, we are guaranteeing models with greater error rates than for diseases with greater prevalence.

Additionally, although the ISO suggests a greater Sp than Se, it can be difficult to weigh the tradeoffs between Se and Sp. The false alert rate is critical because false-positive (FP) and false-negative (FN) detection can be an economic and practical burden. For example, one study found that an automated milking system reported 11,156 alerts for CM, of which 159 were true cases, with farmers using their judgment and non-milking-system data to filter these alerts in an attempt to improve efficiency (Steenefeld et al., 2010). Dominiak and Kristensen (2017) commented, "no sensor-based detection model has fulfilled the performance demands needed to generate a satisfyingly low level of false positive alarms" (p. 60). For a sensor-based mastitis detection system to be successful, the benefits must outweigh the costs. As such, evaluating model predictions' net returns may be a more logical way to assess model success.

Previously unweighted mastitis detection algorithms (i.e., no additional penalty for FP or FN) have been reported using a large variety of data sets and methods (de Mol et al., 1997; Maatje et al., 1997; Cavero et al., 2006; Kamphuis et al., 2008). A common way to predict CM cases is to set variable thresholds (e.g., 15% increase in milk electrical conductivity or a count greater

than  $200 \times 10^3$  for SCC), and all cows that exceed these thresholds are considered mastitic (Maatje et al., 1992; Kamphuis et al., 2008). Although models have been built using weighting schemes and imbalanced Se and Sp (Kamphuis et al., 2010; Miekley et al., 2012; Huybrechts et al., 2014), a cost matrix to assess model applicability is novel to our knowledge. Cost matrices attempt to assign an appropriate cost to model prediction errors by considering the actions taken based on the disease predictions. Because calculating potential cost requires only a model's confusion matrix, such an approach could be widely applied to models in this and other similar fields.

Although more than 130 different etiological agents are known to cause mastitis (Watts, 1988), gram-positive (GP) or gram-negative (GN) bacteria are the most common. Because the protocols for treating mastitis caused by these 2 pathogen types differ, an ideal mastitis detection algorithm would report the probable pathogen type. Recent research has shown differences in sensor variables leading to CM detection in GP and GN cases (Vasquez et al., 2018; Steele et al., 2020). These findings suggest that it might be possible to expand sensor-based mastitic detection algorithms to implicate the infection-causing pathogen. The knowledge that CM presents differently in GP versus GN cases also suggests that model performance could be boosted by separating detection algorithms to be gram specific. This objective of the study was to build separate GN and GP models to predict CM cases using time-series sensor data and then evaluate models based on a partial costs matrix to maximize net return. It was hypothesized that creating models using a partial costs matrix would produce models that would provide greater net returns on farm than those that maximize Se and Sp.

## MATERIALS AND METHODS

### Data

Data were collected from lactating cows on the Virginia Tech (VT) and University of Florida (UF) dairy farms between August 2015 and April 2017. Cows were milked twice daily on both farms at 12-h intervals. A milk meter (AfiMilk MPC, Afimilk Ltd.) measured milk yield and milk electrical conductivity. Additionally, an in-line milk analyzer (AfiLab) collected milk composition data (protein, fat, lactose percent, and cell count). All milk variables were measured twice daily and reported as the average of the 2 values for model building, except yield, which was summed to daily milk yield. Cow activity was collected using Afi Pedometer-Plus at VT and Afi Pedometer at UF, with measured

variables including the number of steps, the number of lying bouts, lying bout duration, and total lying time. All activity variables were summed daily, except for lying bout duration, which was averaged across all measurements within a 24-h period.

A CM case was defined as a cow with abnormal milk with clots or flakes, with or without redness and swelling of the mammary gland in one quarter, as identified by trained milking staff during premilking teat preparation. Detection of CM by farm staff in our study is consistent with the literature (de Mol et al., 1997; Kamphuis et al., 2010; Miekley et al., 2012). A quarter milk sample was aseptically collected from the infected quarter at the time of detection, as described by Steele et al. (2020). Bacteriological procedures were completed per National Mastitis Council guidelines (Middleton et al., 2017). Each CM case was categorized into 1 of 4 groups depending on the type of bacteria isolated: (1) GP pathogens, which included *Staphylococcus aureus*, coagulase-negative *Staphylococcus* spp., and *Streptococcus* spp.; (2) GN pathogens, which included *Klebsiella* spp., *Escherichia coli*, *Citrobacter* spp., *Enterobacter* spp., *Serratia* spp., *Pseudomonas* spp., and *Proteus* spp.; (3) other pathogens, which included *Prototheca*, yeast, and unknown microorganisms; and (4) no growth, in which no pathogens were isolated under aerobic or anaerobic culture (Steele et al., 2020). Cases that were contaminated (i.e., 3 or more dissimilar colony types isolated in culture; Middleton et al., 2017) were excluded, and cases with mixed isolations (i.e., more than one pathogen isolated from a sample regardless of whether the bacterial group was the same) were not available for inclusion in the study.

Based on earlier research, detectable changes in milk or cow behavior could be observed 2 d before CM caused by GN bacteria and 5 d before a GP case of CM (Steele et al., 2020). To leverage the perceived differences between GN and GP cases, only cases assigned to the GP or GN groups were used in the current study. Because of the observed varying lead times between GN and GP cases, in the training data, daily cow records were retrospectively labeled as the cow being clinically mastitic on the day of detection and the preceding 2 or 5 d for GN or GP, respectively. This practice of labeling mastitis differently based on gram type is uncommon with respect to similar research on modeling CM (Steenefeld et al., 2009; Kamphuis et al., 2011). This study chose to label Gram type in an attempt to reduce the complexity of predicting CM since previous research has shown the difference in sensor variable behavior depending on gram type. This data set provided a representative proportion of CM cases to healthy cows (milking = 136,127, CM% = 0.47%).

Supplemental Table S1 ([https://figshare.com/articles/journal\\_contribution/Supplemental\\_Tables\\_docx/14597250/2](https://figshare.com/articles/journal_contribution/Supplemental_Tables_docx/14597250/2), Liebe et al., 2021) shows the summary statistics of the raw data before any data cleansing.

### Data Preparation

Rates of change (specifically slopes) were calculated to detect interest variable changes using linear regression in R (R Core Team, 2018). Slopes were used instead of absolute differences in variables because each cow's baseline values for variables of interest vary (e.g., the peak milk production of cows varies within a herd). For example, an absolute decrease in milk yield of 5 kg/d is less severe the more milk a cow produces each day. Slopes were calculated for lactose, protein, and fat percentages. Additionally, we computed slopes for milk yield, electrical conductivity, SCC, average activity, number of lying bouts, average lying bout duration, and total lying time. We measured slopes for each variable between d -7 and -5, -4, -3, -2, and -1 relative to detected CM, as described in Steele et al. (2020). Missing or otherwise erroneous data prevented slopes from being measured for certain days and variables. Handling these missing and erroneous data cases help to improve the robustness of the complete data set.

On both farms, daily milk yield and composition were determined at morning (AM) and evening (PM) milkings. Because the parlor system used to collect milk data is automated, occasional system failures led to missing data. Such failures included misread animal tags by the in-parlor radio-frequency identification tag reader, milk composition or yield readings that were outside the feasible range and were not reported, and missed events that required manual recording. Computer error or missed milking events are challenging to differentiate simply by examining the raw data. These missing data created a computational challenge during analysis because the maximum number of days used in the slope calculation was 7, meaning each missing day had great potential to affect the resulting slope. Approximately 0.78% of milk yield and composition records were missing at either the a.m. or p.m. measurements. We choose to omit these missing points because cows were milked every 12 h, lending to similar a.m. and p.m. values. After adjusting for these missing values, daily average milk yield and composition were calculated by averaging over paired samples each day by cow. The use of this data-filling technique helped with computing the change in each variable over time.

Data were then standardized to mean = 0 and standard deviation = 1 by subtracting the sample mean from each observation and dividing by the sample

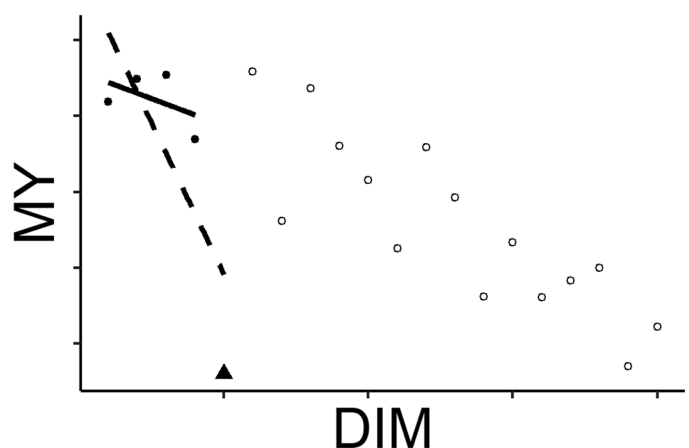
standard deviation, and values greater than 5 standard deviations from the mean were considered outliers and were removed. Although these labeled outliers could be indicative of acute changes, the few data points were removed in an attempt to train robust models. Outlier data that were truly erroneous could be detrimental to predicting a disease with such a low prevalence in the data already. This approach eliminated approximately 5% of values in each variable measured, which is consistent with other research (Kamphuis et al., 2010). Conductivity, daily lying bout, lying bout duration, and step activity measurements had a higher proportion of outlier data than milk yield and composition variables. Outliers were also identified based on the average of previous values to deal with variables with significant variance. With milk yield, an outlier at one point in lactation may be typical later in lactation. An illustration of this issue is shown in Figure 1, where the first 5 records form the dashed linear regression best fit line. However, the rate of change over the previous 4 d, shown with the solid best fit line, is dissimilar to that on d 5. The top 5% of points in terms of deviation from the 4-d average change on d 5 were identified as outliers. Although some of the points identified as outliers could be true physiological responses to acute illness, highly deviating slope values would decrease the robustness of the models because patterns would be more difficult to detect. Visual inspection of these outlier points supported the decision to remove them. Many of the most dissimilar points were either clearly machine error or values rebounded to more reasonable values in the next milking. By considering relative change instead of absolute deviation of single-day records for a particular cow and replacing points that deviate considerably from the slope of the previous 4 d, we were able to detect more outliers. In this example, detected outliers on d 5 were replaced with an average of that variable measured over the previous 4 d.

Data from each farm were cleaned and standardized separately. When combined standardization was used, data were found to be too dissimilar between farms and made slopes less indicative of changes. The differences between farms were likely due to VT farm being a much newer facility, with less static management practices that sometimes lead to much more movement of cows. Standardization allows for comparison across variables in terms of change, is common in multivariate model building, and has been used to standardize conductivity scores in past CM sensor models (de Mol and Woldt, 2001). The combined data set with standardization and outliers removed is summarized in Supplemental Table S2 ([https://figshare.com/articles/journal\\_contribution/Supplemental\\_Tables\\_docx/](https://figshare.com/articles/journal_contribution/Supplemental_Tables_docx/)

14597250/2, Liebe et al., 2021;  $n = 110,156$ ,  $CM\% = 0.30\%$ ). Although the CM prevalence in this data set was greater than the estimated 0.1% of the average US farm (Pol and Ruegg, 2007; USDA, 2014), the data set labeled CM cases on the days before the recorded case. Because multiple days were labeled as positive for each CM case, the on-farm average and any herd data set prevalence of positive cases should be necessarily different values. Also, it is important to consider that individual farm CM prevalence can be very close to zero but cannot be negative, making the distribution unlikely to be symmetrical. Farms with greater CM prevalence skew the distribution to the right. Accounting for the preceding day labeling and a small number of farms' disproportionate effects, the total CM incidence in the data set is similar to other research in the field of sensor-based detection modeling, with de Mol et al. (1997) reporting a CM incidence of 0.14% per cow day in 75,000 milkings or Miekley et al. (2012) reporting 0.5% CM incidence per day in 46,000 cow days. These data cleaning techniques are useful in preparing training data sets and with existing farm monitoring systems because missing data, outliers, and skewed values are commonplace.

### Boosting

Gradient boosting tree algorithms (sometimes referred to as just “boosting”) have recently become a popular machine learning technique because of their accuracy and speed. Boosting is a method of learning a data set by deriving iteratively more accurate “weak



**Figure 1.** Example plot of milk yield (MY) by DIM with an outlier data point on d 5 (▲). Non-outliers are shown as open circles (o), with points included in the regression as solid circles. The linear regression line for d 1 to 4 is shown with a solid black line, whereas the same rate of change for d 1 to 5 is shown as a dashed line.



learners.” Weak learners in a boosting tree algorithm are decision trees that are simple (few nodes) and have predictable biases that can be corrected with additional decision trees. By iteratively correcting these trees by adding another tree, improved accuracy can be achieved. This method is also useful for learning features’ relative importances by analyzing the trees’ Gini indices. Gini indices measure a tree’s ability to produce classifications proportional to the label’s true proportion in the data and are commonly used to identify feature importance of decision-tree-based algorithms (Ushikubo et al., 2017).

Extreme gradient boosting (**XGBoost**) was the algorithm used for modeling CM because of its previous success in other sparse data problems (e.g., store sales prediction, high energy event classification, online course dropout rate; Chen and He, 2014; Chen and Guestrin, 2016). The XGBoost version 0.82.1 was used in R (R Core Team, 2018) for all model building.

### Model Building

The response variable used in all algorithms was the presence of CM, caused by either GN or GP bacteria, depending on the model. To better account for low incidence, GN cases were predicted using all data, including healthy cows and GP cases as negative examples. Conversely, GP cases were predicted using data including healthy cows and GN cases as negative examples. Using cases with the opposite pathogen type as a negative example in the training data allowed for a larger data set and should theoretically improve the algorithm’s ability to differentiate between CM types. The algorithms’ potential explanatory variables included slopes for daily lying bout, lying bout duration, lying time, activity, conductivity, fat, lactose, protein, and milk yield. Additional variables were included for lactation number (either first, second, or third and greater) and DIM discretized into 7 bins (each 50 d, with the final bin including all cases where DIM >350). Seven categories were chosen for DIM to describe the lactation curve because this best balanced the need to separate differently sloping areas of the curve while maintaining large enough sample sizes within each category.

The XGBoost model predicted CM on a continuous scale (probability between 0 and 1). The threshold for positive predictions was 0.5. The error weighting related to positive predictions was scaled 1,000,000 times that of a negative in the XGBoost framework using the “scale\_pos\_weight = 1000000” command. The increase in weight makes predictions on the few positive cases in the data more important. The value of 1,000,000 was chosen because accuracy did not improve beyond this value.

### Model Evaluation

Evaluation of models was completed using held-out data, with 75% of the data used to train models and 25% to test. Random splits of this proportion were repeated 100 times and Se and Sp were recorded. Change in potential net return ( $\Delta_{\text{cost}}$ ) based on the model decision was used as a metric for the cost associated with each model. The  $\Delta_{\text{cost}}$  is a comparison of actions brought on by the model’s prediction compared with the lack of any action. Therefore, in the case that the model predicts no action,  $\Delta_{\text{cost}} = \$0$ . This may seem counter-intuitive because clearly farmers would eventually step in to treat the CM case, even if the model disagreed. The net-zero cost on nonaction is appropriate for comparison to other CM models because reported performance metrics for most CM models do not adjust for human intervention improvements. For a given metric of a CM model, this means that given the input data only, said model alone would perform as predicted.

Change in potential net cost ( $\Delta_{\text{cost}}$ ) was calculated as the difference in returns between the model-recommended action and inaction on the part of the farmer. Returns for GN and GP cases were considered the milk production return from a healthy cow, minus the cost associated with milk loss, treatment, and increased mortality. Using the estimates from Cha et al. (2011), a cow without CM was considered to return \$426 per year. The cost of a case of GN CM was estimated at \$211.04, consisting of \$152.76 due to milk loss, \$25.54 due to decreased fertility, and \$32.74 due to treatment-associated costs (Cha et al., 2011). Importantly, none of the cost of a GN CM case was associated with discarded milk, as antibiotic treatment is uncommon for GN cases. A case of GP CM was estimated to cost \$133.73, consisting of \$49.64 due to milk loss, \$15.20 due to decreased fertility, and \$68.89 due to treatment-associated costs (Cha et al., 2011). Here, \$20 was estimated to be the value of discarded milk from antibiotic treatment (Cha et al., 2011). Using these values, the return of a treated cow that was not infected (FP) would be the base return (\$426) less the costs of treatment and only discarded milk loss. Comparing the FP case return to the return of inaction in the case of an uninfected cow,  $\Delta_{\text{cost}} = -\$32.74$  for GN and  $\Delta_{\text{cost}} = -\$88.89$  for GP. The return for treating a truly infected cow (true positive; **TP**) would be the normal return \$426 less the full cost of a GN or GP case. Comparing the TP case return to the return of inaction in the case of an infected cow,  $\Delta_{\text{cost}} = +\$211.04$  for GN and  $\Delta_{\text{cost}} = +\$133.73$  for GP. Importantly, the net costs compare the returns of the same actual cases, with either no action or action based on model predictions. Figure 2 shows how the costs of different actions compare with

**Table 1.** Partial costs (\$) matrix or change in the net return of an animal classified as either clinical mastitis (CM) positive or negative, given that the animal was either CM positive or negative<sup>1</sup>

Item	CM negative	CM positive
Predicted negative	0/0	0/0
Predicted positive	-32.74/-88.89	211.04/133.73

<sup>1</sup>Values are reported for gram-negative/gram-positive cases, respectively.

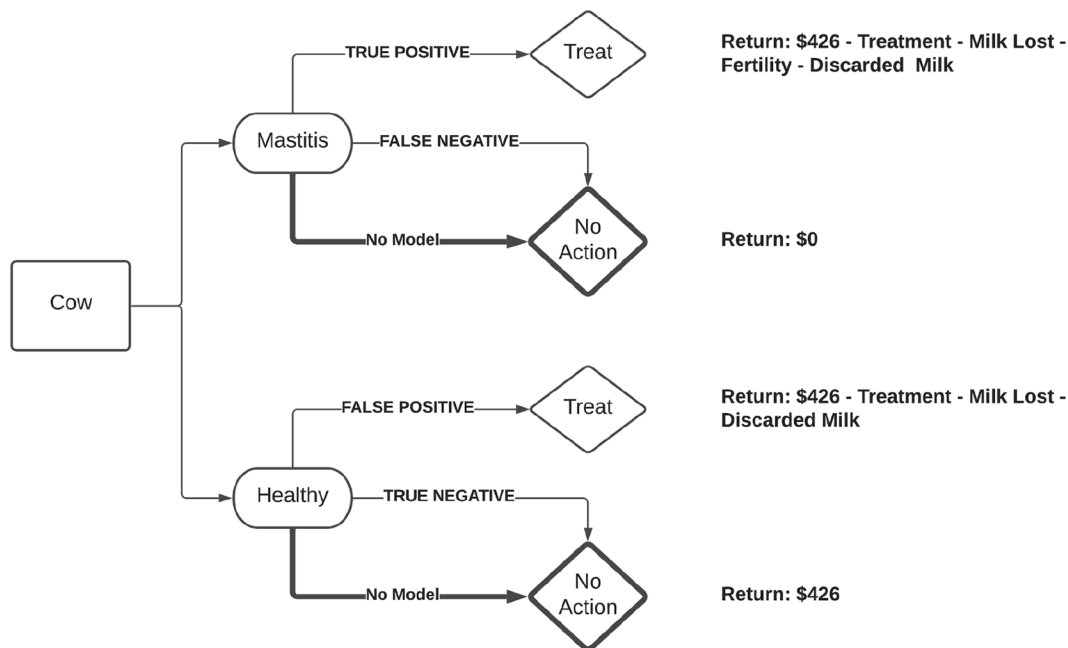
results given no model is employed. The change in cost in both the true- and false-negative cases is zero ( $\Delta_{\text{cost}} = \$0$ ) because these cases reflect the nonaction of using no model. Table 1 illustrates the costs associated with each outcome of the confusion matrix.

It should be noted that another potential cost to inaction is the cost of increased mortality in untreated GP CM cases. A GP CM case would normally be treated with antibiotics. For a model, a CM case can be predicted days before clinical signs are shown. However, prior research on treating GP CM using clinical signs to indicate the start of the treatment process would not provide an accurate estimation of the mortality costs before clinical signs. Once clinical signs of CM have emerged, the mortality costs associated with each additional day of inaction are also difficult to estimate. Because a robust estimate of increased mortality is

challenging, we have omitted additional mortality costs of inaction in the case of GP CM.

The threshold for prediction in the training data was always 0.5 or 50%. Using receiver operator characteristic curves and the cost matrix of missed CM cases and treated CM cases, we determined each resulting model's optimal threshold. The optimal threshold minimized the value of  $[\text{TP} \times \text{TP cost}] - [\text{FP} \times \text{FP cost}]$ , using the net costs associated with GP and GN CM as shown in Table 1 at a given Sp and Se. To prevent the threshold from being near the extremes (i.e., 0% or 100%), thresholds were chosen to maximize the net return and were between 10% and 90%. Excluding thresholds near the extremes prevented models that never predicted CM cases but still produced good net returns per prediction. However, there was no clear improvement by changing the threshold away from the assumed 0.5 used in XGBoost when making predictions. The lack of a consistent threshold that improved net returns, combined with the scale of improvement in returns, led us to choose 0.5 as the threshold for evaluating all proceeding models.

Although a consistent threshold was not found here, in general, the use of a cost matrix to evaluate the threshold for prediction can be useful. Because the cost matrix is not explicitly involved in the penalization of the model, only the weighting scheme for positive



**Figure 2.** Diagram of potential outcomes when predicting clinical mastitis. In both the healthy and mastitic cases, the darker line represents the outcome if no action is taken. The 4 outcomes of a disease prediction model—true and false positives and true and false negatives—are added to show how these suggested actions compare with results under nonaction. True-negative and false-negative predictions would lead to the same outcome as nonaction in the healthy and mastitic cases. Return values for a healthy cow and the 3 major costs associated with mastitis (milk loss, treatment, and fertility) are based on estimates from Cha et al. (2011).

and negative cases, evaluating various thresholds for the outputs of the model, can allow the modeler more control over generation of an applicable model. Where evaluation of the Se and Sp post hoc will provide more insight into the applicability of the model, threshold selection varies the true resulting Se and Sp.

## RESULTS AND DISCUSSION

The GN models in the current work had a median Se of 52.6% and Sp of 99.8% when tested on held-out data. The GP models had a median Se of 37.5% and Sp of 99.9% using the test data (Table 2). These values represent an improvement in mean Sp compared with the values derived on data from the same 2 farms by Steele et al. (2020) using 3-d linear models. The Steele et al. (2020) work focused on a smaller subset of the mastitis data from the VT and UF farms. The 3-d linear models generated a greater Se but lesser Sp; GN models resulted in Se of 71.3% and Sp of 73.8%, whereas for GP models, Se was 47.4% and Sp was 89.4%. In part, these differences are because of the different algorithms used and because of the approach for the selection of prediction thresholds. Reducing the FP rate by using models with greater Sp, as for our models, may result in more suitable models for on-farm use. The conclusion that reducing FP rate at the cost of lowering model Se disagrees with ISO standards requiring relatively consistent values for Se and Sp, and subsequently greater numbers of false positives on sparse data sets.

In terms of Se and Sp, model performance was evaluated for the net cost of each prediction using the following system of equations:

$$TP = IR \times Se,$$

$$FP = (1 - IR) - [(1 - IR) \times Sp], \text{ and}$$

$$\text{cost} = [TP \text{ cost} \times TP] + [FP \text{ cost} \times FP],$$

where TP = true positive rate, IR = incidence rate, and FP = false positive rate.

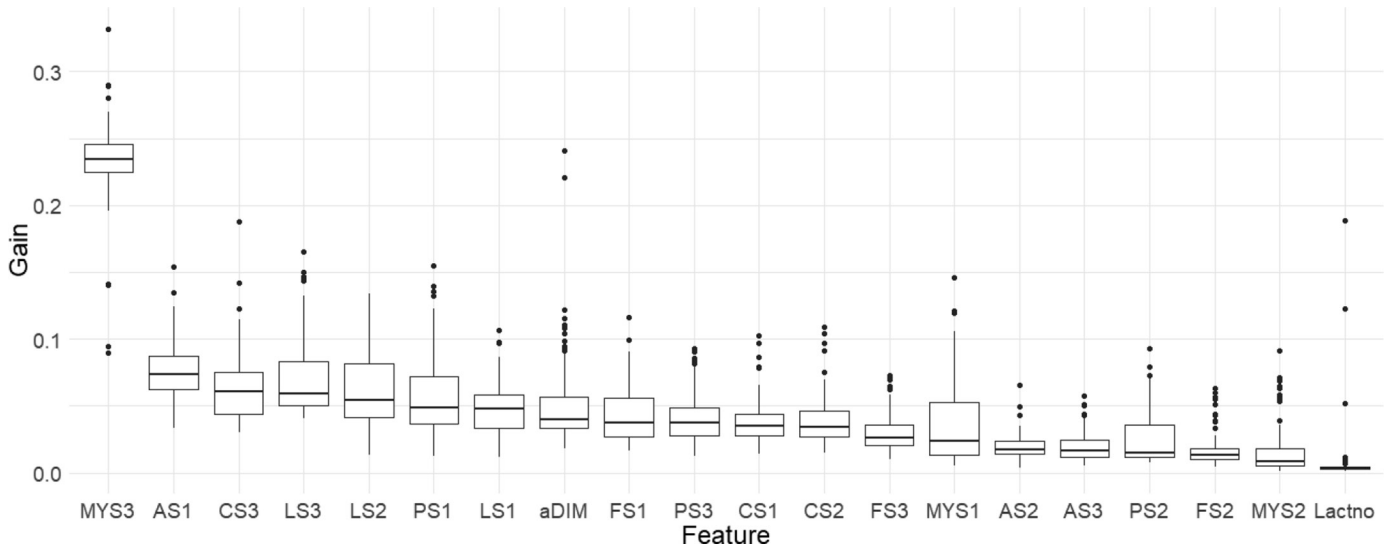
Using the costs associated with each CM infection gram type found in Table 1, the GN prediction model with Se of 52.6% and Sp of 99.8% yielded a net return

of +\$0.27 per cow per lactation, assuming 0.30% CM cases in the data, as in the cleaned data set used here. While 0.30% was used in all calculations in this work, the above equations show how any incidence rate value can be used to evaluate a given model. The incidence rate, along with the Sp and Se values for a model, are all you need to predict the percentages of TP, FP, TN, and FN cases a model will generate. Dominiak and Kristensen (2017) have shown that with a greater disease incidence rate, irrespective of Sp and Se, relative FP and FN predictions decrease. Using the unweighted linear model (Steele et al., 2020) under the same cost matrix produced a net return per cow per lactation of −\$8.10. The optimization parameters of the model being built clearly have importance in the CM prediction problem. For the GP CM prediction model, a Se of 37.5% and Sp of 99.9% yielded an estimated net return per prediction of +\$0.06 per cow per lactation. The net returns per prediction are close to zero compared with the previous unweighted linear model (Steele et al., 2020), which produced a net loss per cow per lactation of −\$9.20. However, both these comparisons assume no producer input into the decisions, because producers provide intuition and insight into treatment decisions (Steenefeld et al., 2010).

Plots of all variables and distributions of their respective importances are shown in Figure 3 for GN models and Figure 4 for GP models. Milk yield slope over the previous 3 d was consistently the most important variable in the GN models, whereas lactose and protein % slopes had the greatest median gains in the GP model accuracy. Lactation number was consistently the least influential variable in the models. In the GN models, the median gain in accuracy attributed to milk yield was more than twice that of the next most important variable, activity, measured as steps per day. Clinical mastitis cases caused by GN bacteria are associated with greater losses in milk yield during first-case occurrences when compared with GP cases (Gröhn et al., 2004). Gram-negative CM is also associated with a faster onset of infection and subsequently faster decline in milk yield compared with GP CM (Smith et al., 1985; Pyörälä et al., 1994; Bannerman et al., 2004); these easily measurable variations attributable to GN CM suggest that classification methods by gram type

**Table 2.** Median sensitivity and specificity values (interquartile range) for 100 models built on 75% of data and tested on 25% of held-out data for gram-positive and gram-negative prediction models

Item	Sensitivity	Specificity
Gram-negative	0.5263 (0.4615–0.6087)	0.9978 (0.9976–0.9979)
Gram-positive	0.3750 (0.3367–0.4152)	0.9985 (0.9983–0.9987)



**Figure 3.** Model importance for all variables in the gram-negative clinical mastitis models, arranged by median importance over the 100 model iterations tested on held-out data. All variables are the change (slope) of the variable over time. MYS = milk yield; AS = activity (steps); CS = conductivity; LS = lactose; PS = protein; FS = fat; aDIM = adjusted DIM (7 bins); Lactno = lactation number (first, second, 3+). The number following the abbreviation indicates the number of days before the day of prediction that the slope is estimated. Each box contains all points between the first and third quartiles, with the median shown as the line within each box. Each whisker extends no further than 1.5 times the distance between the first and third quartiles. Any additional points are shown individually.

are viable. The model results highlight differing variables as most important in the days leading up to CM detection, supporting the supposition that GP and GN cases present in different manners. Pathogen-specific treatment of CM has been shown to be cost-effective in treating GN CM and can reduce the use of antibiotics (Schukken et al., 2011; Fuenzalida and Ruegg, 2019). The variation in CM presentation coupled with the fact that targeted treatment has shown promise suggests that retroactive classification models built on a cost matrix framework could produce tools for recommending CM treatment protocol using on-farm data, but deriving such retroactive models was outside the scope of this study.

To select more applicable models for use on farm, we must be able to contextualize the models created. The cost matrix can be used in the evaluation of existing models and considered when training new models. By plugging the Se, Sp, and prevalence values into the above system of equations, an estimation of cost can be obtained. Let us reconsider the 2 hypothetical dairy farm disease models from the Introduction, both with 95/95 Se/Sp, one for a disease with 50% incidence and the other for a disease with 5% incidence. The ISO guidelines would suggest the models are equal in value, since they both meet the accuracy criteria. For argument's sake, we will use the GN CM cost matrix and apply it to these models. On a hypothetical farm with 50% disease incidence, the model would save the

producer \$99.42 per cow per lactation. The farm with 5% disease incidence would only achieve gains of \$2.13 per cow per lactation. Shifting the incidence by a factor of 10 decreased expected gain by almost 98%. This example echoes the importance of using a context that is relevant to the model you are assessing. In predicting CM cases, the incidence rate is critical, along with the Se and Sp.

This assessment of cost is not perfect because farmers are likely to integrate their own judgment into any treatment decision (Steenefeld et al., 2010), meaning any model would most likely be employed in addition to human observation, which would surely reduce costs in fewer cases ignored when predicted to be negative by a model. However, it allows for a more realistic and objective comparison among models and highlights major challenges with precision dairy technologies' existing ISO standards (International Standard Organization, 2007).

In dairy farming, as in other sectors, profitability is of primary importance to producers. By analyzing the results of the model's predictions and using reasonable costs, the savings from a correctly identified mastitis case is approximately equal to the loss of treating an uninfected cow (Cha et al., 2011). With this cost assumption, the hypothetical model trained on 5% positive cases would have a negative net return per cow treated. To reach net positive returns on predictions, we needed to reach a ratio of



$$\frac{\text{Benefit of True Positives} \cdot \text{True Positives}}{\text{Cost of False Positives} \cdot \text{False Positives}} > 1.$$

Because the costs of FP and benefits of TP are similar, this suggests that measuring the number of TP for each predicted positive instance, called positive predictive value (**PPV**), adds insights that Se and Sp lack on their own. The formula for PPV is

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

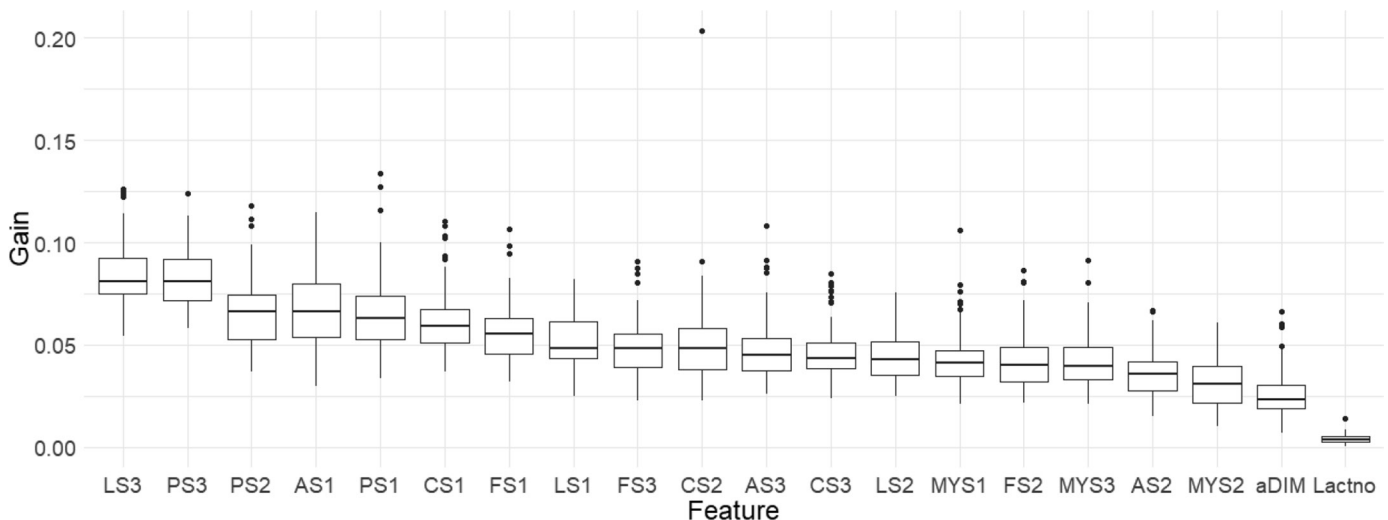
The implications of models with high Sp and Se but low PPV would be high false alert rates and possibly unnecessary treatment. As shown previously, the TP and FP values can be calculated directly from the Sp, Se, and incidence rate of disease in a model. In this work, the GN model (Se = 52.6; Sp = 99.8) and the GP model (Se = 37.5; Sp = 99.9) would have PPV = 44% and PPV = 53%, respectively. The 3-d linear models from Steele et al. (2020) would have PPV = 0.8% for GN (Se = 71.3; Sp = 73.8) and PPV = 1.3% for GP (Se = 47.4; Sp = 89.4). These differences in the chance that a positive prediction is a true case of CM are stark between the 2 approaches and illustrate the misleading nature of using only Se and Sp as measures of performance, especially in diseases with low incidence. In the context of mastitis prediction, farmers would be checking most cows and not finding clinical signs. Farmers will become insensitive to the alerts of a

system like this. “Alarm fatigue” is a well-documented problem in critical care medicine (Graham and Cvach, 2010; Borowski et al., 2011; Emergency Care Research Institute, 2013). Attempting to predict the greatest proportion of true cases possible leads to high rates of FP alerts, increased workload, decreased Se, and increased missed critical events (Graham and Cvach, 2010). Therefore, an appropriate threshold PPV value (or similar statistic) should be established.

Our work predicted net returns per cow per lactation were +\$0.27 for GN and +\$0.06 for GP models using a cost matrix to evaluate models. The value of using a cost matrix-based evaluation to choose models over simultaneously maximizing Sp and Se can be established by comparing the net returns of each approach. Consider that in a data set with 0.3% CM incidence, GN models with Se/Sp = 80/95, 95/95, and 95/98 would have net returns per lactation of −\$1.13, −\$1.03, and −\$0.05, respectively, under our partial costs matrix. The best of these models, Se/Sp = 95/98, would still produce 3 times more false positives per true positives (0.13 vs. 0.44 PPV) when compared with the GN model described in this work with Se/Sp = 52.6/99.8.

## CONCLUSIONS

Although guidelines for CM prediction models tend to focus on Se and Sp values, other measures, including PPV, may be equally crucial due to the low incidence and high false-positive rate in prediction. Using data



**Figure 4.** Model importance for all variables in the gram-positive clinical mastitis models, arranged by median importance over the 100 model iterations tested on held-out data. All variables are the change (slope) of the variable over time. MYS = milk yield; AS = activity (steps); CS = conductivity; LS = lactose; PS = protein; FS = fat; aDIM = adjusted DIM (7 bins); Lactno = lactation number (first, second, 3+). The number following the abbreviation indicates the number of days before the day of prediction that the slope is estimated. Each box contains all points between the first and third quartiles, with the median shown as the line within each box. Each whisker extends no further than 1.5 times the distance between the first and third quartiles. Any additional points are shown individually.

collected from 2 dairy farms, we illustrate some of the challenges associated with the implementation of CM models on farm. Erroneous data, missing data, sparse cases for training, and varying costs for FP and FN predictions all affect the efficacy of models in a workflow involving model and farmer insights. Because the number of predicted and actual true- and false-positive cases can be derived using only the model Se, model Sp, and disease incidence rate, we provide a clearer picture of the relationship between these 3 variables in practice. Based on the cost matrix used in this study, relatively simple calculations show the models' ineffectiveness with even the best Se and Sp values, given a low incidence rate. A current on-farm CM prediction system will likely need to involve the farmer to produce economical results.

## ACKNOWLEDGMENTS

We acknowledge the farm staff at the Virginia Tech Dairy Center (Blacksburg, VA) and the University of Florida Dairy Unit (Gainesville, FL) for assistance with data collection and Afimilk Ltd. (Kibbutz Afikim, Israel) for data extraction. Funding to support aspects of this work was received from the Virginia Agricultural Council (Richmond), the Virginia Tech Pratt Fellowship fund, and the United States Department of Agriculture (awards: 2018-02492 and 2017-05943). Nicole Steele was partially supported by New Zealand dairy farmers through DairyNZ Inc. (Hamilton, New Zealand). The authors have not stated any conflicts of interest.

## REFERENCES

- Bannerman, D. D., M. J. Paape, J.-W. Lee, X. Zhao, J. C. Hope, and P. Rainard. 2004. *Escherichia coli* and *Staphylococcus aureus* elicit differential innate immune responses following intramammary infection. *Clin. Vaccine Immunol.* 11:463–472. <https://doi.org/10.1128/CDLI.11.3.463-472.2004>.
- Borowski, M., M. Görges, R. Fried, O. Such, C. Wrede, and M. Imhoff. 2011. Medical device alarms. *Biomed. Tech. (Berl.)* 56:73–83. <https://doi.org/10.1515/bmt.2011.005>.
- Cavero, D., K.-H. Tölle, C. Buxadé, and J. Krieter. 2006. Mastitis detection in dairy cows by application of fuzzy logic. *Livest. Sci.* 105:207–213. <https://doi.org/10.1016/j.livsci.2006.06.006>.
- Cha, E., D. Bar, J. A. Hertl, L. W. Tauer, G. Bennett, R. N. González, Y. H. Schukken, F. L. Welcome, and Y. T. Gröhn. 2011. The cost and management of different types of clinical mastitis in dairy cows estimated by dynamic programming. *J. Dairy Sci.* 94:4476–4487. <https://doi.org/10.3168/jds.2010-4123>.
- Chen, T., and C. Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. *CoRR abs/1603.02754*. <https://doi.org/10.1145/2939672.2939785>.
- Chen, T., and T. He. 2014. Higgs Boson discovery with boosted trees. Pages 69–80 in *Proceedings of the 2014 International Conference on High-Energy Physics and Machine Learning - Volume 42*. JMLR.org.
- de Mol, R. M., G. H. Kroeze, J. M. F. H. Achten, K. Maatje, and W. Rossing. 1997. Results of a multivariate approach to automated oestrus and mastitis detection. *Livest. Prod. Sci.* 48:219–227. [https://doi.org/10.1016/S0301-6226\(97\)00028-6](https://doi.org/10.1016/S0301-6226(97)00028-6).
- de Mol, R. M., and W. E. Woldt. 2001. Application of fuzzy logic in automated cow status monitoring. *J. Dairy Sci.* 84:400–410. [https://doi.org/10.3168/jds.S0022-0302\(01\)74490-6](https://doi.org/10.3168/jds.S0022-0302(01)74490-6).
- Dominiak, K. N., and A. R. Kristensen. 2017. Prioritizing alarms from sensor-based detection models in livestock production - A review on model performance and alarm reducing methods. *Comput. Electron. Agric.* 133:46–67. <https://doi.org/10.1016/j.compag.2016.12.008>.
- Emergency Care Research Institute. 2013. Top 10 Health Technology Hazards Report For 2014. Accessed May 21, 2018. [https://www.ecri.org/Resources/Whitepapers\\_and\\_reports/2014\\_Top\\_10\\_Hazards\\_Executive\\_Brief.pdf](https://www.ecri.org/Resources/Whitepapers_and_reports/2014_Top_10_Hazards_Executive_Brief.pdf).
- Fuenzalida, M. J., and P. L. Ruegg. 2019. Negatively controlled, randomized clinical trial to evaluate intramammary treatment of nonsevere, gram-negative clinical mastitis. *J. Dairy Sci.* 102:5438–5457. <https://doi.org/10.3168/jds.2018-16156>.
- Ganda, E. K., R. S. Bisinotto, D. H. Decker, and R. C. Bicalho. 2016. Evaluation of an on-farm culture system (Accumast) for fast identification of milk pathogens associated with clinical mastitis in dairy cows. *PLoS One* 11:e0155314. <https://doi.org/10.1371/journal.pone.0155314>.
- Graham, K. C., and M. Cvach. 2010. Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms. *Am. J. Crit. Care* 19:28–34, quiz 35. <https://doi.org/10.4037/ajcc.2010651>.
- Gröhn, Y. T., D. J. Wilson, R. N. González, J. A. Hertl, H. Schulte, G. Bennett, and Y. H. Schukken. 2004. Effect of pathogen-specific clinical mastitis on milk yield in dairy cows. *J. Dairy Sci.* 87:3358–3374. [https://doi.org/10.3168/jds.S0022-0302\(04\)73472-4](https://doi.org/10.3168/jds.S0022-0302(04)73472-4).
- Hogeveen, H., C. Kamphuis, W. Steeneveld, and H. Mollenhorst. 2010. Sensors and clinical mastitis—The quest for the perfect alert. *Sensors (Basel)* 10:7991–8009. <https://doi.org/10.3390/s100907991>.
- Huybrechts, T., K. Mertens, J. De Baerdemaeker, B. De Ketelaere, and W. Saey. 2014. Early warnings from automatic milk yield monitoring with online synergistic control. *J. Dairy Sci.* 97:3371–3381. <https://doi.org/10.3168/jds.2013-6913>.
- ISO (International Organization for Standardization). 2007. Automatic milking installations—Requirements and testing. International Organization for Standardization.
- Kamphuis, C., H. Mollenhorst, A. Feelders, D. Pietersma, and H. Hogeveen. 2010. Decision-tree induction to detect clinical mastitis with automatic milking. *Comput. Electron. Agric.* 70:60–68. <https://doi.org/10.1016/j.compag.2009.08.012>.
- Kamphuis, C., H. Mollenhorst, and H. Hogeveen. 2011. Sensor measurements revealed: Predicting the Gram-status of clinical mastitis causal pathogens. *Comput. Electron. Agric.* 77:86–94. <https://doi.org/10.1016/j.compag.2011.03.012>.
- Kamphuis, C., R. Sherlock, J. Jago, G. Mein, and H. Hogeveen. 2008. Automatic detection of clinical mastitis is improved by in-line monitoring of somatic cell count. *J. Dairy Sci.* 91:4560–4570. <https://doi.org/10.3168/jds.2008-1160>.
- Khatun, M., C. E. F. Clark, N. A. Lyons, P. C. Thomson, K. L. Kerrisk, and S. C. García. 2017. Early detection of clinical mastitis from electrical conductivity data in an automatic milking system. *Anim. Prod. Sci.* 57:1226–1232. <https://doi.org/10.1071/AN16707>.
- Koskinen, M. T., J. Holopainen, S. Pyörälä, P. Bredbacka, A. Pitkälä, H. W. Barkema, R. Bexiga, J. Roberson, L. Sølverød, R. Piccinini, D. Kelton, H. Lehmusto, S. Niskala, and L. Salmikivi. 2009. Analytical specificity and sensitivity of a real-time polymerase chain reaction assay for identification of bovine mastitis pathogens. *J. Dairy Sci.* 92:952–959. <https://doi.org/10.3168/jds.2008-1549>.
- Liebe, D., N. Steele, C. Petersson-Wolfe, A. De Vries, and R. White. 2021. Supplemental Tables. Figshare. Journal contribution. <https://doi.org/10.6084/m9.figshare.14597250.v2>.
- Maatje, K., R. M. de Mol, and W. Rossing. 1997. Cow status monitoring (health and oestrus) using detection sensors. *Comput. Electron. Agric.* 16:245–254. [https://doi.org/10.1016/S0168-1699\(96\)00052-X](https://doi.org/10.1016/S0168-1699(96)00052-X).

- Maatje, K., P. J. M. Huijsmans, W. Rossing, and P. H. Hogewerf. 1992. The efficacy of in-line measurement of quarter milk electrical conductivity, milk yield and milk temperature for the detection of clinical and subclinical mastitis. *Livest. Prod. Sci.* 30:239–249. [https://doi.org/10.1016/S0301-6226\(06\)80013-8](https://doi.org/10.1016/S0301-6226(06)80013-8).
- Middleton, J., L. Fox, G. Pighetti, and C. Petersson-Wolfe. 2017. *Laboratory Handbook on Bovine Mastitis*. National Mastitis Council.
- Miekley, B., I. Traulsen, and J. Krieter. 2012. Detection of mastitis and lameness in dairy cows using wavelet analysis. *Livest. Sci.* 148:227–236. <https://doi.org/10.1016/j.livsci.2012.06.010>.
- Pol, M., and P. L. Ruegg. 2007. Treatment practices and quantification of antimicrobial drug usage in conventional and organic dairy farms in Wisconsin. *J. Dairy Sci.* 90:249–261. [https://doi.org/10.3168/jds.S0022-0302\(07\)72626-7](https://doi.org/10.3168/jds.S0022-0302(07)72626-7).
- Pyörälä, S. 2003. Indicators of inflammation in the diagnosis of mastitis. *Vet. Res.* 34:565–578. <https://doi.org/10.1051/vetres:2003026>.
- Pyörälä, S., L. Kaartinen, H. Käck, and V. Rainio. 1994. Efficacy of two therapy regimens for treatment of experimentally induced *Escherichia coli* mastitis in cows. *J. Dairy Sci.* 77:453–461. [https://doi.org/10.3168/jds.S0022-0302\(94\)76973-3](https://doi.org/10.3168/jds.S0022-0302(94)76973-3).
- R Core Team. 2018. *R: A Language and Environment for Statistical Computing*.
- Sargeant, J. M., K. E. Leslie, J. E. Shirley, B. J. Pulkraabek, and G. H. Lim. 2001. Sensitivity and specificity of somatic cell count and California Mastitis Test for identifying intramammary infection in early lactation. *J. Dairy Sci.* 84:2018–2024. [https://doi.org/10.3168/jds.S0022-0302\(01\)74645-0](https://doi.org/10.3168/jds.S0022-0302(01)74645-0).
- Schukken, Y. H., G. J. Bennett, M. J. Zurakowski, H. L. Sharkey, B. J. Rauch, M. J. Thomas, B. Ceglowski, R. L. Saltman, N. Belomestnykh, and R. N. Zadoks. 2011. Randomized clinical trial to evaluate the efficacy of a 5-day ceftiofur hydrochloride intramammary treatment on nonsevere gram-negative clinical mastitis. *J. Dairy Sci.* 94:6203–6215. <https://doi.org/10.3168/jds.2011-4290>.
- Smith, K. L., D. A. Todhunter, and P. S. Schoenberger. 1985. Environmental pathogens and intramammary infection during the dry period. *J. Dairy Sci.* 68:402–417. [https://doi.org/10.3168/jds.S0022-0302\(85\)80838-9](https://doi.org/10.3168/jds.S0022-0302(85)80838-9).
- Steele, N. M., A. Dicke, A. De Vries, S. J. Lacy-Hulbert, D. M. Liebe, R. R. White, and C. S. Petersson-Wolfe. 2020. Identifying gram-negative and gram-positive clinical mastitis using daily milk component and behavioral sensor data. *J. Dairy Sci.* 103:2602–2614. <https://doi.org/10.3168/jds.2019-16742>.
- Steenefeld, W., L. C. van der Gaag, H. W. Barkema, and H. Hogeveen. 2009. Providing probability distributions for the causal pathogen of clinical mastitis using naive Bayesian networks. *J. Dairy Sci.* 92:2598–2609. <https://doi.org/10.3168/jds.2008-1694>.
- Steenefeld, W., L. C. van der Gaag, W. Ouweltjes, H. Mollenhorst, and H. Hogeveen. 2010. Discriminating between true-positive and false-positive clinical mastitis alerts from automatic milking systems. *J. Dairy Sci.* 93:2559–2568. <https://doi.org/10.3168/jds.2009-3020>.
- USDA. 2014. *Health and Management Practices on U.S. Dairy Operations*. USDA.
- Ushikubo, S., C. Kubota, and H. Ohwada. 2017. The Early Detection of Subclinical Ketosis in Dairy Cows Using Machine Learning Methods. Pages 38–42 in *Proceedings of the 9th International Conference on Machine Learning and Computing*. ACM.
- Vasquez, A. K., D. V. Nydam, C. Foditsch, M. Wieland, R. Lynch, S. Eicker, and P. D. Virkler. 2018. Use of a culture-independent on-farm algorithm to guide the use of selective dry-cow antibiotic therapy. *J. Dairy Sci.* 101:5345–5361. <https://doi.org/10.3168/jds.2017-13807>.
- Watts, J. L. 1988. Etiological agents of bovine mastitis. *Vet. Microbiol.* 16:41–66. [https://doi.org/10.1016/0378-1135\(88\)90126-5](https://doi.org/10.1016/0378-1135(88)90126-5).

## ORCIDS

- D. M. Liebe  <https://orcid.org/0000-0003-4447-4120>  
 N. M. Steele  <https://orcid.org/0000-0001-9915-0954>  
 C. S. Petersson-Wolfe  <https://orcid.org/0000-0002-2766-1306>  
 A. De Vries  <https://orcid.org/0000-0003-4511-0388>  
 R. R. White  <https://orcid.org/0000-0001-5713-012X>