ABSTRACT

This paper proposes and discusses a methodology to evaluate the performance of automated mastitis-detection systems with respect to their practical value on farm. The protocols are based on 3 on-farm requirements: (1) to detect cows with clinical mastitis promptly and accurately to enable timely and appropriate treatment, (2) to identify cows with high somatic cell count to manage bulk milk SCC levels, and (3) to report the mastitis infection status of cows at the end of lactation to support decisions on individual cow dry-cow therapy. Separate protocols for each requirement are proposed and discussed, including gold standards, evaluation tests, performance indicators, and performance targets. Aspects that require further research or clarification are identified. Actual field data are used as examples. Further debate is invited, the aim being to achieve international agreement on how to evaluate and report performance of different mastitis-detection technologies. Better performance information will allow farmers to compare different mastitis-detection systems sensibly and fairly before investing. Also, the use of evaluation protocols should help technology providers to refine current, or develop new, automated mastitis-detection systems. Such developments are likely to accelerate adoption of these systems, potentially leading to improved animal health, milk quality, and labor productivity.

Key words: testing protocol, udder health management, mastitis sensing system

INTRODUCTION

In most dairy systems, it is assumed that farmers, informed by the official organization in their country or region, have the responsibility to deliver milk that is of sufficient quality (Hogeveen et al., 2010). Although regulations dealing with milk quality standards differ between dairy-producing countries, general agreement exists that abnormal milk, including milk from diseased [e.g., due to clinical mastitis (CM) or highly elevated SCC levels] or injured udders should be excluded from milk supplied for human consumption. Identifying cows with CM involves visual inspection of the udder and manually checking the foremilk from each quarter at each milking, whereas regular herd tests and various other systems (e.g., the California mastitis test) can be applied to identify cows with elevated SCC. However, increasing herd sizes, reliance on less-skilled labor, and an increasing emphasis on lowering bulk milk SCC (BMSCC) levels (Adkinson et al., 2001; Brightling et al., 2005; Lacy-Hulbert et al., 2010) are all factors contributing to an increased demand for more consistent and less labor-intensive methods to help farmers manage mastitis and BMSCC levels.

Mastitis-detection systems have been developed to automate, or at least contribute to, the processes of managing mastitis and BMSCC. For these automated mastitis-detection systems to be useful on farm, they need to (1) identify cows with CM promptly and accurately to enable timely treatment, maximize cure rate, and reduce the risk of spread of infection; (2) identify cows with high SCC to manage BMSCC levels by withholding cows temporarily or long term; and (3) identify the infection status of cows at the end of the lactation to support decisions on individual cow dry-cow therapy. Most research to date has focused on assessing the sensors’ ability to detect CM (Hogeveen et al., 2010), with little or no evidence of field performance of mastitis-detection systems relative to the other 2 requirements. It is important to define evaluation protocols so that independent and uniform performance information can be provided to farmers. This information will enable farmers to assess if a mastitis-detection system will meet their requirements and they will be able to make more informed investment decisions.

Protocols have been established considering each requirement separately and, where possible, comprise (1) an appropriate and practical gold standard, (2) an appropriate test against the gold standard and (3) minimum performance targets. Data from 2 previous...
field studies are used to illustrate the proposed gold standards and evaluation tests, and to identify areas that require further research.

DEVELOPMENT OF EVALUATION PROTOCOLS

Requirement 1: Detecting Clinical Mastitis Promptly and Accurately

Background Information. The National Mastitis Council defines CM as the presence of flakes, clots, or other gross alterations in milk appearance, irrespective of its SCC level (Smith et al., 2001). Whereas this is a clear definition, the development of a practical and objective assessment of the condition of (quarter) milk for large-scale field evaluation of mastitis-detection systems is less clear. Previous research on CM detection used a variety of gold standards to define CM (Hogeveen et al., 2010) and Claycomb et al. (2009) demonstrated that even minor differences in gold standard definitions had a large effect on the number of true CM cases used for evaluation purposes. In addition, differences in data used for analyses (Kamphuis et al., 2010a) and time windows applied (Kamphuis et al., 2010b) are factors contributing to the variation of reported performance of mastitis-detection systems. Recently, there has been effort to develop meaningful and consistent methods of field evaluation of automatic mastitis-detection systems in relation to an appropriate gold standard, data used for analyses, evaluation tests, and performance targets (ISO, 2007; Mein and Rasmussen, 2008; Claycomb et al., 2009; Hogeveen et al., 2010). The information contained in these reports has formed the basis of the protocol proposed to evaluate the ability of a mastitis-detection system to detect CM.

Proposed Gold Standard. For a field evaluation, a CM episode is defined as the observation of clots (>2 mm in average diameter) at 2 out of 3, or all 3, consecutive cow milkings (Mein and Rasmussen, 2008). A single CM episode is not necessarily limited to 3 consecutive milkings but can continue as long as clots are found at 2 out of 3 consecutive milkings. Presence of clots should be assessed using in-line filters (Claycomb et al., 2009). Situation A in Figure 1 illustrates how clot observations should be used to identify CM: 5 of 15 consecutive cow milkings had clots present. Using the proposed gold standard, this example shows one CM episode that started at the first observation of clots (milking 6). The second clot observation at milking 8 serves as a confirmation that the first observation was the start of a CM episode. For the purposes here, the CM episode ends at the last milking in which clots are observed where this is followed by 4 consecutive cow milkings without clot observation (milking 9). Situation A (Figure 1) also has 2 stand-alone clot observations at milking 1 and milking 14. These clot observations, however, fall outside the proposed gold standard definition and are, therefore, not considered to be (part of) a CM episode.

Proposed Evaluation Test. Field data should be collected from a minimum of 3 commercial farms and collection should continue until a minimum of 20 CM episodes per farm have been identified (ISO, 2007; Hogeveen et al., 2010). A mastitis alert should be generated by the detection system within 4 milkings (48-h time window) around the first cow milking with clots in a confirmed CM episode (Hogeveen et al., 2010; Table 1). Time-window analysis (Sherlock et al., 2008) should be used to link alerts generated by a mastitis-detection system with CM episodes. This time-window analysis is illustrated in Figure 1, where 4 situations (A, B, C, and D) show 15 consecutive cow milkings, with 5 having clots present; there is 1 CM episode (milking 6 through milking 9) and 2 stand-alone clot observations (milking 1 and milking 14). The time window of 4 cow milkings is centered around the start of the milking with the first clot observation [i.e., a time window of 2 cow milkings before the first cow milking with clots and 1 cow milking after the first cow milking with clots of a confirmed CM episode (milking 4 through milking 7)]. Situation A in Figure 1 illustrates a situation where the mastitis-detection system generated 3 alerts. The first occurred at the start of the time window of 4 cow milkings and is, therefore, considered a true positive (TP) alert. The other 2 alerts occurred outside that time window and are, therefore, false positive (FP). Situation B (Figure 1) represents a situation where 3 alerts are generated but all fall outside the time window of 4 cow milkings in which a mastitis alert should be expected and, therefore, they are all FP. This also implies that the CM episode in situation B remains undetected and, therefore, the time window of 4 cow milkings receives a single false-negative (FN) alert. Situation C in Figure 1 illustrates a more complex situation where consecutive mastitis alerts occurred within 1 CM episode. The first alert at milking 3 in situation C (Figure 1) is an FP alert. The alert at milking 4 is the first alert within the time window in which a mastitis alert should be expected and is, therefore, TP. If that first alert had occurred at milking 5, 6, or 7, then this would have been classed as TP also. In the situation of consecutive mastitis alerts within the time window of 4 cow milkings (alerts at milking 5 and 7 in situation C), these should be considered as confirmative alerts of the first mastitis alert. Therefore, these confirmative alerts are merged with the first alert and they are counted as 1 TP alert. The mastitis alerts at milkings 8, 9, and 10 in situation C are also confirmative alerts, as they appear within 2
cow milkings of the previous alert that was merged with a TP alert. However, if the first mastitis alert had been at milking 9 (situation D; Figure 1), that mastitis alert would not fulfill the requirement of a mastitis-detection system alerting promptly for CM. On the other hand, the alert would fall within a CM episode and because of that it cannot be considered as an FP alert. Instead, it is suggested to describe this alert as a late TP alert and the alert at milking 10 would be a confirmative alert. In other words, late TP alerts are those alerts that fall outside the time window of 4 cow milkings but occurred within the CM episode. Therefore, that alert is considered a late TP alert. The alert at milking 10 is again a confirmative alert. The alerts and number of cow milkings are used to calculate sensitivity (SN) and the false-alert rate per 1,000 cow milkings (FAR1000). In all 4 situations, milking 1 and 14 are ignored, as these involve single-clot observations. The alert at milking 13 is ignored also, as this is generated 1 cow milking before the single-clot observation at milking 14. For situation C, this leaves 1 FP at cow-milking 3, 1 timely TP alert at cow-milking 4, and 13 cow milkings to be included for FAR1000. The hypothetical mastitis-detection system in situation C, therefore, has a timely SN of 100% and a FAR1000 of 77.

Figure 1. Applying the proposed gold standard to identify clinical mastitis (CM) episodes and using time-window analyses to link CM episodes with alerts generated by a mastitis-detection system. The 4 situations (A, B, C, and D) all have 15 consecutive cow milkings in which 5 have clots that were scored for density [low (L), medium (M), or high (H)]. There was 1 CM mastitis episode (milking 6 through 9) and the time window of 4 cow milkings in which an alert was expected is indicated by dark gray blocks (milking 4 through 7). Situation A illustrates 3 separate alerts: one occurred at the start of the time window of 4 cow milkings and is, therefore, a true positive (TP). The other 2 were outside that period and are false positive (FP). Situation B had 3 FP alerts; all alerts occurred outside the time window of 4 cow milkings and outside the CM episode. The CM episode in situation B remained undetected and is considered 1 false-negative (FN) alert. Situation C had 2 FP alerts and a TP alert, followed by a series of confirmative alerts (dotted arrows) within the time window of 4 cow milkings. These are combined into 1 TP alert; the alerts on milkings 8, 9, and 10 are confirmative alerts also, as they occurred within 2 cow milkings of an alert that was merged with a TP alert. Situation D again had 2 FP alerts. The alert at milking 9 falls outside the time window of 4 cow milkings but occurred within a CM episode. Therefore, that alert is considered a late TP alert. The alert at milking 10 is again a confirmative alert. The alerts and number of cow milkings are used to calculate sensitivity (SN) and the false-alert rate per 1,000 cow milkings (FAR1000). In all 4 situations, milking 1 and 14 are ignored, as these involve single-clot observations. The alert at milking 13 is ignored also, as this is generated 1 cow milking before the single-clot observation at milking 14. For situation C, this leaves 1 FP at cow-milking 3, 1 timely TP alert at cow-milking 4, and 13 cow milkings to be included for FAR1000. The hypothetical mastitis-detection system in situation C, therefore, has a timely SN of 100% and a FAR1000 of 77.
outside the required time window of 4 cow milkings but do fall within a CM episode. If the first alert had been at milking 10, that alert would fall outside the time window of 4 cow milkings and outside a CM episode and, therefore, should be considered an FP (situation A and B; Figure 1). Every consecutive alert after milking 10 would be a separate FP alert.

The FP, FN, and TP mastitis alerts (Table 1) are used to calculate sensitivity (SN) and the false-alert rate per 1,000 cow milkings (FAR1000). The SN refers to the proportion of CM episodes that are correctly detected by a mastitis-detection system, with overall SN being calculated as SN = (TP/(TP + FN)) × 100% (Table 1). The FAR1000 refers to the number of FP alerts per 1,000 cow milkings, calculated as FAR1000 = (1,000 × FP)/total cow milkings (Hillerton, 2000; Rasmussen, 2005; Sherlock et al., 2008; Hogeveen et al., 2010).

### Results and Discussion Based on Application of the Proposed Protocol to Field Data

The proposed clot-based gold standard was applied to data from a field study conducted at a New Zealand dairy farm where the main diet was grazed pasture and where cows calved seasonally in a block (B. Dela Rue, unpublished data). Data were collected during late lactation from approximately 1,900 cows milked once per day in a 60-stall rotary parlor over a 4-wk study commencing on March 15, 2010. By the end of the study period (April 10, 2010), cow numbers had decreased to approximately 1,250 as cows were dried off because of low feed levels or the interval before expected calving date. During the study period, in-line filters with a removable stainless steel screen and a pore size of 0.9 mm² (Vision Mastitis Detector; Ambic Equipment Ltd., Witney, UK; Figure 2) were fitted to the long milk tubes at every stall. These filter screens were removed and visually inspect-
ed for the presence of clots by trained farm staff at the end of each cow milking following cup removal. When clots on the filter were identified, the filter screen was replaced with a new one and cow identification, stall number, and date were recorded. All filter screens with clots were photographed to assign a clot score relating to density of clots at a later date. Density scores (low, medium, and high), representing the percentage of area of the in-line filter covered, were used to provide an indication of the severity of infection. For consistency, one single skilled person (G. Mein) conducted all the clot scoring based on the scoring protocol provided by the supplier of the in-line filters (Figure 2).

From the 67,919 cow milkings recorded during the study period of 4 wk involving approximately 1,900 cows, clots were observed on in-line filters on 387 (0.6%) occasions. Of the 387 clot observations, 345 (89%) had a proper digital image that allowed the clot to be scored for its density. The majority of these clots (n = 345) had a medium-density score (50%). Another 39% had a high-density score and the remaining clots (11%) were scored as low density. The 387 occasions where clots were observed would be equivalent to 20 clot events per 100 cows per month, which would be a high CM-incidence rate. However, after applying the proposed clot-based gold standard, 61 CM episodes were identified, which is equivalent to 3.2 CM episodes per 100 cows per month. This value is within the range of 4.8 CM episodes per 100 cows per month reported by Claycomb et al. (2009) and the 20 to 40 CM cases per 100 cows per year reported by Hillerton (2000). It is also within the range of the New Zealand national guidelines of having 1 CM episode per 100 cows per month (DairyNZ Ltd., 2012a). Another 239 cow milkings with clot observations were excluded as being (part of) a CM episode, as they appeared during 1 cow milking only. Of these stand-alone clot observations, 209 were scored for their density (14% low, 58% medium, and 28% high density). The clot score was used to provide information on the severity of infection, where high-density clots represented a more severe infection than low-density clots. With this in mind, the high prevalence of medium- and high-density stand-alone clot observations requires careful handling in the proposed evaluation test. These clot observations fall outside a CM episode by the proposed definition, but clots may be of a severity for a farmer to start an intervention treatment immediately without waiting for another confirmatory milking with clots. In addition, Mollenhorst et al. (2012) identified that farmers milking robotically prefer a mastitis-detection system that is accurate in finding severe (high-density clot) cases of CM. If this preference also applies to farmers that milk using conventional milking systems, these cow milkings may very well be of interest to farmers who may want to check these cows visually for other

![Figure 2](image_url)
signs of CM. If the 28% of high-density clot observations reported in this study are indicative of the true prevalence of cows with high-density clots occurring for 1 milking, this may be evidence that these transient cows should be included in the gold standard definition for CM. Further research is required to confirm the high percentage of transient cows and to identify if farmer practice better matches a density-based gold standard rather than a gold standard based on consecutive clots. In the meantime, we propose to exclude the cow milkings with single observations of clots from evaluation purposes and to exclude the alerts at these milkings or at milkings plus or minus 1 cow milking around these single-clot observations. For situation C (Figure 1), this means that there was 1 correctly detected CM episode (milking 6 through milking 9, first alerted at milking 4), 1 FP at milking 3, and a total of 13 cow milkings to be included in the calculation of FAR1000; the hypothetical mastitis-detection system would, therefore, have an SN of 100% and a FAR1000 of 77.

In addition to transient cows, another group of cows is not considered by the proposed gold standard: cows with watery milk are excluded by definition when using a clot-based gold standard. It is hard to differentiate watery milk from normal milk in a controlled experimental setting (Rasmussen, 2003) and is likely to be even more difficult under the milking time constraints of commercial farms. However, watery milk may be an indicator of an impending severe mastitis infection, with electrical conductivity values comparable to that of milk showing many clots or milk that was clearly altered from normal (Kamphuis et al., 2008a). If future research demonstrates that watery milk should be included in a CM gold standard, a practical and objective way to assess wateriness of milk will be necessary. One option that can be explored to objectively assess watery milk is the use of color sensors. Kamphuis et al., (2008a) reported that quarters with watery milk have higher values for reflection of light-emitting diodes representing the red and blue color frequency, indicating that watery milk is more bluish or purplish than normal milk or milk with flakes and clots. Newer versions of color sensors that measure light-emitting diode light transmittance (Song et al., 2010) may be even better at identifying watery milk, but this should be explored in future research.

The application of the gold standard to field data revealed the significant effort required to assess the performance of mastitis-detection systems in identifying CM episodes. Given that the proposed protocol is recommended to be conducted on 3 commercial farms, the labor intensiveness and the subsequent costs of this protocol may be a limiting factor for the industry. Consequently, a need may exist to develop a proxy measure for CM that is easier to assess. Instead of using a clot-based gold standard, treated cases of CM may be sufficient to retrieve at least an indicative performance of an automated mastitis-detection system. Using the strategy of treated cases of CM, however, requires the performance results to be handled with care because the gold standard (treated cases of CM) is not completely independent from the sensor under study and because decisions to treat cows for CM is subjective. For these reasons, a gold standard based on treated cases of CM is likely to be an underestimation of the true prevalence of CM on a farm (Claycomb et al., 2009). Because of these risks, using treated cases of CM as the gold standard can never replace a clot-based gold standard completely. However, it may be good enough to give a first indicative result, which is much more than most farmers have access to today. Further research is required to answer the question of how good this proxy measure is in providing an indicative performance estimate for CM detection, particularly the risk of overestimating detection performance when using a gold standard that directly or indirectly uses information from the sensor under study.

In summary, based on current knowledge concerning requirement 1, the proposed clot-based gold standard is the most useful for evaluating the ability of an automated mastitis-detection system to find cows with CM. It is recommended that further study should focus on the prevalence and importance of transient cows and watery milk to determine if these categories should be included in a gold standard definition for CM.

**Requirement 2: Finding Cows With High SCC Levels for Managing BMSCC**

**Background Information.** One strategy to manage BMSCC is to withhold cows (or quarters) from supply temporarily, or permanently by early drying off or, in some cases, by culling cows. Reasons for withholding cows from supply may include the BMSCC trending toward penalty levels or toward a farmer’s own maximum BMSCC level target. Whatever the reason, farmers will want to reduce BMSCC by withholding the milk of as few cows as possible. To achieve that outcome, farmers need information about which cows contribute the most to BMSCC levels. A cow’s total SCC contribution can be calculated from her milk yield multiplied by her SCC. For those farmers who herd test regularly, this information can be found on herd test records. It is proposed that this same herd test information can be used to evaluate mastitis-detection systems for their ability to identify cows with high SCC for BMSCC management.
Proposed Gold Standard. A herd test should be conducted on at least 3 commercial farms to calculate the total SCC contribution for each cow. Using these data, cows are ranked where the cow with the highest total SCC contribution is ranked 1. The same data can be used to calculate the estimated BMSCC by summing all individual cows’ SCC contribution and then dividing that by the total volume produced by all cows in the herd (Holdaway et al., 1996). The theoretical cumulative decrease in estimated BMSCC can be calculated by excluding cows one by one based on their ranking and by recalculating the estimated BMSCC. By plotting this decrease, a BMSCC ranking curve can be created. The herd test information used to produce this ranking curve is proposed as the gold standard for this requirement.

Proposed Evaluation Test. Cows should be ranked according to the output of the mastitis-detection system to create a second ranking curve. The percentage of cows that need to be excluded to decrease BMSCC by 25%, for example, should be calculated using the gold standard and compared with the percentage of cows to be excluded as indicated by the mastitis-detection system. It is recommended to use large herds in the evaluation of mastitis-detection systems, as in small herds there is the risk of small changes in the absolute number of cows to be excluded having a large effect when expressed as percentages.

Proposed Minimum Performance Target. The proposed tentative performance target for this requirement is that a mastitis-detection system performs satisfactorily if it excludes no more than twice the percentage of cows relative to the percentage excluded by the gold standard.

Results and Discussion Based on Application of the Proposed Protocol to Field Data. The proposed gold standard, evaluation test, and performance target were applied retrospectively to data from cows milked robotically by 2 automatic milking systems (AMS; Fullwood Ltd., Shropshire, UK) as described by Kamphuis et al. (2008b). Average milk production was 3,700 kg/cow, with an average BMSCC of 189,000 cells/mL. At each milking, sensors on the AMS measured milk yield, estimated SCC at the cow level (system 1), and measured electrical conductivity, both at the cow level (system 2) and at the quarter level (system 3). The quarter with the highest electrical conductivity within a cow milking was chosen to represent the output value of system 3. Herd tests were conducted every 14 d to measure SCC at the cow level. Kamphuis et al. (2008b) provides a detailed description of how composite udder milk samples were collected for system 1 and for herd tests. Briefly: 2 composite milk samples were collected consecutively from the receiver at the end of milking for each cow. The first sample was used to estimate SCC at the cow level (system 1); the second sample was used for herd testing to determine SCC at the cow level. Data from early and late lactation were excluded from analyses due to insufficient sensor information available. Data included for analyses were collected from September 2006 through the end of February 2007. During this time period, 11 herd tests were conducted and a total of 940 cow milkings had all information (herd test data and sensor data from systems 1, 2, and 3) available.

Figure 3 illustrates the ranking curves of the gold standard and of system 1, 2, and 3. To decrease BMSCC by 25%, the gold standard suggests excluding the top 1.5% of the ranked cows. Applying the performance target means that a mastitis-detection system should achieve the 25% decrease in BMSCC by excluding a maximum of 3% of the cows when ranked according to the output of that system. Figure 3 (n = 940 records) demonstrates that system 1 would meet the proposed minimum performance target, but systems 2 and 3 would require approximately 10% of the ranked cows to be removed before a 25% reduction in BMSCC would be achieved. Using threshold values other than a 25% reduction in BMSCC showed similar results (Table 2 and Figure 3): System 1 consistently meets the proposed target compared with the gold standard, whereas systems 2 and 3 do not. However, Table 2 highlights that using thresholds lower than a 25% reduction in BMSCC results in very few cows to be excluded when ranked according to the gold standard. This, in turn, can cause the results for mastitis-detection systems to appear to be more variable due to large changes in percentages when small numbers of cows are affected. Therefore, a 25% reduction in BMSCC seems to be a sensible threshold to compare mastitis-detection systems against. Looking in more detail at this 25% threshold revealed that the performance of the gold standard and the 3 mastitis-detection systems varied between herd tests (Table 3). Table 3 demonstrates that system 1 was consistent in the percentage of cows that it excluded from the bulk tank to achieve the proposed target (excluding approximately 3%, except for herd tests 2 and 11). In contrast, systems 2 and 3 varied considerably in the percentage of cows they excluded from the bulk tank and they consistently suggested excluding a higher percentage from the herd than indicated by system 1. Had an evaluation been done using herd test 2 only, all 3 mastitis-detection systems would have failed to meet the target, although system 1 did meet the target for all other herd tests (Table 3). Similarly, had the evaluation been done using herd test 3 only, all systems would meet the target, even though systems 2 and 3 failed the target for most other herd
tests (Table 3). These results suggest that multiple herd tests are recommended for evaluation purposes. The gold standard consistently excluded approximately 2% of the ranked cows to decrease BMSCC levels by 25% (or the nearest % above). The only exception was herd test 11, where the gold standard identified to exclude approximately 12% of cows to decrease BMSCC by 25%. Whether this unexpectedly high percentage is due to the low BMSCC level at that particular herd test or whether other reasons caused the data on that particular herd test to be different from the other herd tests is unknown. However, knowledge about the causes of inherent variation between herd tests is irrelevant because performances of mastitis-detection systems are relative to that of the gold standard; if the gold standard requires more cows to be excluded from the bulk tank, the systems are also allowed to exclude more cows to get to the same 25% decrease in BMSCC. Further research is in progress to determine the minimum number of herd tests, the minimum number of cows to be sampled per herd test, and the required distribution of SCC within this sampled group to get a reliable performance indication from mastitis-detection systems; the number of consecutive herd tests required to get a representative gold standard SCC value for each cow; and if a difference exists in performance when the test is conducted using data from herds with high BMSCC (≥250,000 cells/mL) or low BMSCC (<150,000 cells/mL).

In summary, the application of the proposed gold standard, evaluation test, and performance targets revealed that this was straightforward to conduct and simple to interpret. This evaluation could be conducted by farmers, if desired, to verify that their system is working adequately or by farmers wishing to invest in automated mastitis-detection systems.

**Requirement 3: Identifying the Infection Status of Cows at the End of Lactation to Support Individual Cow Dry-Cow Therapy Decisions**

Generally, guidelines differ between countries for determining which cows should be treated with antibiotics at drying off. For example, blanket antibiotic dry-cow therapy is recommended in the Netherlands (UGCN, 2012), whereas selective dry-cow therapy is recommended in Ireland for herds with low-risk mastitis profiles and with sufficient herd testing data and bacteriology results (Animal Health Ireland, 2012). In New Zealand, national guidelines are designed to ensure that all cows are protected by some form of treatment during the dry period, whether this is an internal teat...
sealant or antibiotics (DairyNZ Ltd., 2012b). The New Zealand guideline advises to first assess the mastitis risk at the herd level. If considered high, then all cows should be treated with antibiotic dry-cow therapy. If the mastitis risk at the herd level is considered low, the CM and SCC histories for each cow are required to identify cows that should receive long-acting antibiotic dry-cow therapy and those that can be treated with an internal teat sealant. The recommendation is to apply antibiotic dry-cow therapy to cows with high SCC levels (>150,000 cells/mL for cows; >120,000 cells/mL for heifers) at any one of the last 3 herd tests and for cows and heifers that had CM at any time during the current lactation. An internal teat sealant can be applied to all other cows in the herd (DairyNZ Ltd., 2012b).

To get a reasonable number of truly infected cows, larger herds with relatively high BMSCC levels are preferred for collecting data to evaluate mastitis-detection systems for their ability to identify udder health status. Collecting data from these herds increases the likelihood of finding infected cows, as approximately 40% of a herd or 15% of all quarters could be infected with a major pathogen with a BMSCC of 400,000 cells/mL (Eberhart et al., 1982; Holdaway et al., 1996). To determine the true infection status of the cows in these herds, the most robust method would be to collect milk samples for bacteriological culturing (Sargeant et al., 2001). However, bacteriological culturing of milk samples from every single cow in these large herds will be expensive. To decrease costs, a subsample of the herd could be selected for bacteriological culturing but this increases the risk that the number of infected cows useful for evaluation purposes will be too low. Furthermore, when sampling cows for bacteriological culturing, a risk exists of samples being assessed as no growth or as contaminated. Contaminated samples are common when sampling is done by farm staff or in herds in which cows are not accustomed to having their teats handled by people. A recent field study in Victoria (Australia) revealed that 23 and 16% of milk samples collected by farmers from cows with CM (n = 3,073) were assessed as no growth and contaminated, respectively. Figures for milk samples collected by farmers from cows with subclinical mastitis (n = 1,079) were 17

Table 2. Absolute number and the percentage (in parentheses) of the top-ranked cows that, when excluded, decreased the bulk milk SCC (BMSCC) level by 5, 15, 25, 35, and 50% (or the nearest percentage above), when ranking was based on the gold standard (GS) or on the outcome of system 1 (sensor estimating SCC at the cow level), system 2 (sensor measuring electrical conductivity at the cow level), or system 3 (sensor measuring electrical conductivity at the quarter level)1

<table>
<thead>
<tr>
<th>Ranking cows based on</th>
<th>Absolute number (and percentage) of cows excluded to decrease BMSCC by</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
</tr>
<tr>
<td>GS</td>
<td>2 (0.2)</td>
</tr>
<tr>
<td>System 1</td>
<td>3 (0.3)</td>
</tr>
<tr>
<td>System 2</td>
<td>20 (2.1)</td>
</tr>
<tr>
<td>System 3</td>
<td>9 (1)</td>
</tr>
</tbody>
</table>

1The analyses were conducted for herd test records combined (n = 940).

Table 3. Percentage of the top-ranked cows that, when excluded, decreased the bulk milk SCC level by 25% (or the nearest percentage above), when ranking was based on the gold standard (GS) or on the outcome of system 1 (sensor estimating SCC at the cow level), system 2 (sensor measuring electrical conductivity at the cow level), or system 3 (sensor measuring electrical conductivity at the quarter level)1

<table>
<thead>
<tr>
<th>Herd test Records, no.</th>
<th>SCC ×10^3 in bulk milk</th>
<th>GS</th>
<th>System 1</th>
<th>System 2</th>
<th>System 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>161</td>
<td>472</td>
<td>3.1</td>
<td>3.1</td>
<td>14.9</td>
</tr>
<tr>
<td>2</td>
<td>59</td>
<td>486</td>
<td>3.4</td>
<td>13.6</td>
<td>22.0</td>
</tr>
<tr>
<td>3</td>
<td>43</td>
<td>404</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>4</td>
<td>53</td>
<td>372</td>
<td>3.8</td>
<td>3.8</td>
<td>9.4</td>
</tr>
<tr>
<td>5</td>
<td>81</td>
<td>370</td>
<td>1.2</td>
<td>1.2</td>
<td>3.7</td>
</tr>
<tr>
<td>6</td>
<td>87</td>
<td>318</td>
<td>2.3</td>
<td>2.3</td>
<td>4.6</td>
</tr>
<tr>
<td>7</td>
<td>69</td>
<td>356</td>
<td>1.4</td>
<td>1.4</td>
<td>5.8</td>
</tr>
<tr>
<td>8</td>
<td>66</td>
<td>367</td>
<td>1.5</td>
<td>1.5</td>
<td>13.6</td>
</tr>
<tr>
<td>9</td>
<td>171</td>
<td>516</td>
<td>1.2</td>
<td>1.2</td>
<td>7.6</td>
</tr>
<tr>
<td>10</td>
<td>82</td>
<td>244</td>
<td>3.7</td>
<td>4.9</td>
<td>19.5</td>
</tr>
<tr>
<td>11</td>
<td>68</td>
<td>216</td>
<td>11.7</td>
<td>17.6</td>
<td>35.3</td>
</tr>
<tr>
<td>Combined</td>
<td>940</td>
<td>395</td>
<td>1.5 (25.3)</td>
<td>1.8 (25.4)</td>
<td>10.6 (25.9)</td>
</tr>
</tbody>
</table>

1The analyses were conducted for each herd test separately (n = 11) and for all herd test records combined (n = 940). Figures in parentheses indicate the percentage decrease in bulk milk SCC level after excluding the reported percentage of ranked cows.
and 40% for no growth and contaminated, respectively (Charman et al., 2012). These culture-negative and contaminated samples are useless for evaluation purposes. So, the high costs involved with bacteriological culturing of many cows and the risk of ending up with too few milk samples with bacteriological results useful for evaluation, either due to subsampling or contamination of samples, make the bacteriological culturing of milk samples a less-than-ideal practical gold standard for this third requirement.

Polymerase chain reaction assays are diagnostic tests that allow identification of mastitis pathogens within a few hours (Koskinen et al., 2009). Koskinen et al. (2010) demonstrated promising advantages of a commercially available PCR assay (PathoProof Mastitis PCR Assay; Finnzymes Oy, Vantaa, Finland) over conventional bacteriology, including speed, automated interpretation of results, and increased SN. That same PCR assay also proved the ability to identify major and minor mastitis pathogens in culture-negative milk samples from cows with CM or with high SCC (Taponen et al., 2009; Wellenberg et al., 2010; Bexiga et al., 2011). However, 4 key issues need to be addressed before PCR can be used as a gold standard to identify the true infection status of cows and to replace or complement culture-based diagnostic tests. The first issue is the higher costs due to added reagent and instrument costs of PCR assays compared with bacterial culture-based methods (Koskinen et al., 2009). Second, PCR detects only the predetermined bacterial species (currently 12 for the PathoProof PCR Assay; Penny et al., 2012). Third, from these 12 pathogens currently targeted, only 2 (Streptococcus agalactiae and Mycoplasma bovis) are obligate udder pathogens. The other 9 pathogens can replicate outside the udder and, therefore, can originate from teat skin or the environment. Accordingly, these organisms can be detected by the PCR assay even when they are not causing udder health infections (Penny et al., 2012). Fourthly, research is required to assess the clinical relevance of finding nonviable pathogens in milk samples (Schukken et al., 2010; Wellenberg et al., 2010) before PCR assay can complement or replace culture-based diagnostic tests (Schukken et al., 2010).

Determination of SCC has been widely implemented as an alternative to bacteriological culturing as a screening test to identify udder infections in lactating cows (Sargeant et al., 2001). However, research also identified SCC to have low SN and specificity when identifying udder infection status in early lactation (Sargeant et al., 2001) and at drying off (VMRG, 1992). At the recommended SCC threshold to make individual dry-cow therapy decisions in New Zealand (150,000 cells/mL; DairyNZ Ltd., 2012b), the SN and specificity of SCC to identify cows infected with a major pathogen was 91 and 56%, respectively (VMRG, 1992). From a mastitis management perspective, and perhaps from a national mastitis management program point of view, these figures may be good enough for using SCC values to make dry-cow therapy decisions. It would ensure that almost all cows that do require antibiotic treatment (91%) would receive it and that the use of antimicrobials in the dairy industry would be reduced: approximately 56% of the cows that do not require antibiotics would not be treated with dry-cow therapy but would receive teat sealant. However, using this SCC threshold as the gold standard to identify an infected udder would result in labeling too many uninfected cows as infected. Evaluating a mastitis-detection system using these mislabeled cows will be misleading. A second disadvantage of using SCC as the gold standard, given its relatively poor relationship with udder infection status, is the risk of disadvantaging mastitis-detection systems that do not measure SCC directly. It would be unfair to exclude mastitis-detection systems simply because it is not possible to set an alert threshold that represents a certain SCC level.

This third requirement of mastitis-detection systems is of lesser importance to the day-to-day management of mastitis. A mastitis-detection system that performs well with respect to requirement 1 and 2 is likely to provide useful CM and SCC information that farmers can use to identify cows with a very low risk of being infected and that may require teat sealant only. However, if mastitis-detection systems do not perform well enough and the mastitis profile of the whole herd would then be classed as high (based on BMSCC values), blanket antibiotic therapy would be the recommended practice. On the other hand, it is likely that this third requirement will become more important for a mastitis-detection system in the future. With the increasing pressure, rightly or wrongly, from human health and animal rights organizations to decrease or even remove pharmaceutical use from the primary food-production chain (Hillerton, 2011), it will become increasingly important to apply antibiotic treatments only to those cows that truly require them. Further discussion or research is required to identify a robust and practical gold standard that enables evaluation of mastitis-detection systems for their ability to identify infected cows.

**CONCLUSIONS**

To be an effective on-farm tool for managing mastitis, an automated mastitis-detection system should be able to help farmers in the detection of cows with CM and in identifying cows with high SCC to manage BMSCC levels. The identification of the infection
status of cows to support decisions on individual cow dry-cow therapy is currently of lesser importance for the day-to-day mastitis management but is likely to become a more important requirement in the future. The proposed protocols and minimum performance targets for the first 2 requirements, however, will provide independent and uniform information, which enables farmers to make more informed investment decisions. These proposed evaluation protocols have been developed from the point of view of the end user. We acknowledge that the proposed protocols are in parts not specific enough to be implemented in practice yet, but they are presented to generate discussion and further development. Field testing is required so that progress is made toward industry-accepted performance assessment. It is envisaged that these protocols can serve as a framework for the development of future and improved automated mastitis-detection systems, which, in turn, potentially will contribute to an improvement in milk quality, animal welfare, and labor productivity.

Data used to illustrate the proposed evaluation protocols were collected at 2 New Zealand farms and the proposed protocol for the first requirement is essentially based on systems milking twice daily. In this twice-per-day milking situation, the time window of 4 cow milkings in which an alert for CM is expected from a mastitis-detection system fits the requirement of an alert falling within a 48-h time window, as suggested by Hogeveen et al. (2010). However, from a treatment perspective, the time window of 4 cow milkings may be too wide when milking once daily. The time window of 4 cow milkings seems suitable for robotic milking systems, but further research is needed regarding technical requirements of collecting clot information on these systems and how the evaluation protocol should deal with incomplete or failed milkings. The protocols and proposed minimum performance targets are easily adapted to requirements of other countries or to requirements of other milking systems. For example, the proposed minimum performance targets regarding the FAR1000 for CM detection may still be too low for farmers using AMS. With these systems, farmers are not present during the milking process and their code of practice usually involves visually checking cows that are listed on a mastitis-alert list. It is known that farmers using robotic milking systems are eager to minimize the additional labor of checking large numbers of FP alerts (Claycomb et al., 2009) and that they prefer a system that generates a low number of FP alerts per day (Mollenhorst et al., 2012). However, although differences may occur between countries and milking systems, the proposed guidelines can be adapted to fit each new requirement and are, therefore, applicable to any automated mastitis-detection system.

ACKNOWLEDGMENTS

This study was funded by DairyNZ Inc. (Hamilton, New Zealand) on behalf of New Zealand Dairy Farmers (project SY1006).

REFERENCES


