HHS Public Access

Author manuscript

J Dairy Sci. Author manuscript; available in PMC 2020 January 02.

Published in final edited form as:

J Dairy Sci. 2019 August ; 102(8): 7494–7502. doi:10.3168/jds.2019-16351.

Reliability of observational- and machine-based teat hygiene scoring methodologies

David I. Douphrate[*,1], **Nathan B. Fethke**[†], **Matthew W. Nonnenmann**[†], **Anabel Rodriguez**[*], **David Gimeno Ruiz de Porras**[*]

[*]Southwest Center for Occupational and Environmental Health, Department of Epidemiology, Human Genetics & Environmental Sciences, School of Public Health in San Antonio, The University of Texas Health Science Center at Houston

[†]Department of Occupational and Environmental Health, College of Public Health, University of Iowa

Abstract

Removal of teat-end debris is one of the most critical steps in the premilking process. We aimed to estimate inter- and intra-rater reliability of an observation-based rating scale of dairy parlor worker teat-cleaning performance. A nonrandom sample of 8 experienced raters provided teat swab debris ratings scored on a 4-point ordinal visual scale for 175 teat swab images taken immediately after teat cleaning and before milking unit attachment. To overcome the uncertainty associated with visual inspection and observation-based rating scales, we assessed the reliability of an automated observer-independent method to assess teat-end debris using digital image processing and machine learning techniques to quantify the type and amount of debris material present on each teat swab image. Cohen's kappa coefficient (κ) was used to assess inter-rater score agreement on 175 teat swab images, and the intraclass correlation coefficient was used to assess both intra-rater score agreement and machine reliability. The reliability of debris scoring of teat swabs by raters was low (overall $\kappa = 0.43$), whereas the machine-based rating system demonstrated near-perfect reliability (Pearson $r > 0.99$). Our findings suggest that machine-based rating systems of worker performance are much more reliable than observational-based methods when evaluating premilking teat cleanliness. Teat swab image analysis technology can be further developed for training and quality control purposes to enable more efficient, reliable, and independent feedback on worker milking performance. As automated technologies are becoming more popular on dairy farms, machine-based teat cleanliness scoring could also be incorporated into automated milking systems.

Keywords

milk quality; udder hygiene; worker performance; machine learning

[1]The University of Texas Health Science Center at Houston, School of Public Health in San Antonio, Department of Epidemiology, Human Genetics & Environmental Sciences, Southwest Center for Occupational and Environmental Health, 7411 John Smith Drive, Suite 1100, San Antonio, Texas 78229, USA; phone: 1-210-276-9005; david.i.douphrate@uth.tmc.edu.

INTRODUCTION

Udder health must be protected to optimize both milking efficiency and overall milk yield. Therefore, a key dairy management strategy is to ensure that parlor workers milk clean, dry, stimulated teats using proper pre-and postmilking hygiene procedures. Removal of teat-end debris (e.g., manure, bedding, and dirt) is among the most critical steps in the premilking process. Visual inspection of teat wipes is commonly used to evaluate teat-cleaning effectiveness before milking unit attachment, and several observation-based rating scales have been reported for this purpose (Cook, 2002; Hovinen et al., 2005; Reneau et al., 2005).

A limitation to the use of available observation-based rating scales of post-wipe teat swabs is a lack of methodologically rigorous evaluations of their inter- and intra-rater agreement. One study reported (1) good to excellent intra-rater reliability of observation-based hygiene scores among 3 dairy science academic faculty and 1 graduate student, and (2) moderate to good correlation in hygiene scores between 2 faculty and 14 student observers with varying levels of dairy industry experience (Reneau et al., 2005). Additional studies have reported moderate to good agreement in ratings between 2 research team members (Devries et al., 2012) or among repeat ratings by a single research team member (Schreiner and Ruegg, 2003; Ellis et al., 2007). No published studies describe agreement in rating scores between or within individuals from industry stakeholder groups (e.g., producers, veterinarians, and extension personnel). Having industry stakeholder perspective is necessary to effectively evaluate the utility of such a field-based quality control tool.

One assumption underlying the use of observation-based rating scales is that an increase in subjective rating score is associated with an increase in debris or microbial contamination near the teat end, a reduction in milk quality, or both. Also, while visual inspection and observation-based rating scales focus on ascertaining cow (or teat, or both) hygiene status, failure to sufficiently remove chemical sanitizers (e.g., dip) from the teats during milking preparation may also compromise milk quality (Galton et al., 1984; Castro et al., 2012; French et al., 2016). As a result, there is an industry need to develop objective methods of ascertaining teat cleanliness (i.e., presence of teat debris and chemical residue) before milking unit attachment.

The primary objective of this study was to estimate inter- and intra-rater reliability of a common observation-based rating scale used to evaluate worker teat-cleaning performance. Additionally, we introduce a method for an automated, objective assessment of teat swab debris using digital image processing and machine learning techniques. Results of the observer-based scoring methodology were compared with results of machine-based assessment methodology. The intra-rater reliability of the machine-based assessment method (i.e., machine reliability) was also explored. Findings from this study will lead to the research and development of automated solutions using digital sensor technologies for milking performance monitoring in the dairy industry.

MATERIALS AND METHODS

Teat Swab Collection and Digital Imaging

To ensure a broad range of the presence of teat-end debris and dip chemical(s), we collected a single teat swab from each of 175 randomly selected cows at a single, large (i.e., milking herd size of about 3,800 cows) dairy farm located in the southwestern United States. At the time of data collection, the dairy farm had not received any rain for at least 2 wk, and pens were dry and well groomed. Immediately following completion of manual premilking routine tasks and before milking unit attachment, a research team member positioned the center of a dry, sterile 12-ply gauze pad (10.2 cm × 10.2 cm square) under the end of the posterior, right teat. The teat periphery was then wiped approximately 3.0 cm inches proximal to the teat end. One gauze pad was used per teat.

Immediately after collection, each swab was secured in a hinged clamshell vise (Figure 1) with a 7.0-cm diameter circular opening. A color digital image of each swab was then captured using a Sony Handycam HDR-SR12 affixed to a frame above the clamshell vise (Figure 1). Images were captured in the RGB (red, green, blue) color space (32 bits per pixel) using the f/4 aperture setting, an exposure time of 350^{-1} s, and digital dimensions of 3,680 × 2,760 pixels (Figure 1). A portable light tent with a 5600K daylight light-emitting diode light source and reflective interior (model PVTT035, Fotodiox) was used to control lighting conditions around the clamshell vise.

Teat Cleanliness Ratings

We recruited a nonrandom sample of 8 experienced raters to independently provide a teat swab score for each image. Rater A was a dairy veterinarian and research scientist, as well as dairy extension specialist with over 12 yr of experience in the industry; rater B was a dairy management specialist and extension specialist with over 15 yr of experience in the industry; rater C was a dairy veterinarian and dairy farm owner/manager with over 15 yr of experience in udder health and teat hygiene management, as well as milker performance training and teat hygiene monitoring; rater D was a dairy researcher and extension scientist with more than 8 yr of experience in the industry; rater E was a dairy extension outreach trainer with 26 yr of experience in the industry including milker training and teat hygiene assessment; rater F was a dairy nutrition scientist and extension specialist with 27 yr of experience in the industry with extensive experience in milker training and teat hygiene assessment; rater G was a dairy producer with 22 yr of experience in the industry; and rater H was a dairy veterinarian and research scientist, and extension specialist with more than 25 yr of extensive experience in milker training and teat hygiene assessment.

Each rater independently rated teat swab images online using Qualtrics survey software (Qualtrics, Provo, UT). Raters were asked to provide a teat swab debris rating of 175 teat swab images using the Teat Cleanliness Scorecard developed by Westfalia Surge (Cook and Reinemann, 2007). This visual scorecard uses a 4-point ordinal scale to assess the degree of manure, dip, and bedding contamination at the teat end after completion of the milking preparation procedure, before milking unit attachment. A score of 1 = clean, with no manure, dirt, or dip; 2 = dip is present, but no manure or dirt; 3 = small amount of dirt and

manure is present; and 4 = larger amount of dirt and manure is present. A representative teat swab image for visual reference was included for each of the 4 rating scores. Teat swab images were presented to each rater in random order, and each rater was unaware of the scores of other raters.

Digital Image Processing

In digital imaging science, image processing includes analysis, manipulation, storage, and display of pixel information. This pixel information can come from sources such as digital photographs or video. Generated outputs of image processing can be an image or set of characteristics or parameters related to the original image that was processed (Hani, 2013). To overcome the uncertainty associated with visual inspection and observation-based rating scales, we used standard digital image processing techniques to objectively quantify the type and amount of debris material present on each teat-end swab.

First, a custom color classifier was created using the LabVIEW Vision Development Module (version 2017, National Instruments Inc., Austin, TX). Briefly, the color classifier is a database containing samples of each of 4 categories (or “classes”) of material that could appear in each swab image (n = 150 samples per class): (1) “dirt,” or a region consisting of manure, bedding, or dirt; (2) “green,” or a region consisting of green-tinted dip solution; (3) “yellow,” or a region consisting of yellow-tinted dip solution; (4) “swab,” or a region consisting of clean swab material. Note, the milking parlor operation from which teat-end swabs were collected (described above) used a yellow iodine-based premilking dip and a green postmilking dip. Residue from the green postmilking dip (from the prior milking) was visible on several of the teat swabs. Samples were identified manually by selecting regions of the swab images consistent with a material class, and then assigning the class labels to the selected regions. Approximately 25 of the full set of 175 swab images were randomly selected as the source data for the class samples. To quantify color information, the LabVIEW color classifier tool first converts each class sample to the **HSL** (hue, saturation, luminance) color space, creates histograms of each channel (i.e., H, S, and L), suppresses the luminance channel (by 12.5%), and then combines the result into a color “feature” (essentially, a vector describing the color information in the sample).

The distribution of color features across the samples in each class is then used with a predictive (i.e., statistical) model as the basis for assigning a class to each pixel (or pixel area) in a teat swab image. We used the *k*-nearest neighbor (**kNN**) supervised machine learning technique for this purpose (Duda et al., 2000). The use of kNN is common in biomedical imaging applications for which regions of tissue must be isolated for diagnostic determinations (Steenwijk et al., 2013). In our application, the HSL-based color feature of each 3 × 3 pixel area within the circular region of a teat swab image is calculated and compared with the color features of all samples in the color classifier database. The 11 samples nearest the 3 × 3 pixel area (i.e., $k = 11$) are identified based on color feature and calculated using the Manhattan distance function (Muralidharan and Chandrasekar, 2011). The value of k was selected as that which maximized accuracy of the classifier through repeated 10-fold cross-validation (i.e., with $k = 1, 2, \ldots, 25$; Arlot and Celisse, 2010). The kNN model then assigns 1 of the 4 classes to the 3 × 3 pixel area based on a majority vote of

these 11 nearest samples. In addition, an “unassigned” class is assigned to 3 × 3 pixel areas for which the kNN model is unable to assign 1 of the other 4 classes (i.e., based on the color feature of the area and the distance function). The number of pixels assigned to each class is then calculated and expressed as a proportion of the total pixel count within the circular region of the swab image (see Figure 2 for example results).

In addition to cross-validation, performance of the kNN model was assessed using a confusion matrix to estimate the frequency of predicted classes for each set of samples manually assigned to a particular class. Specifically, of the 150 samples of each class, 100 were randomly selected and used to train the kNN algorithm (with $k = 11$, as described above). The remaining 50 samples of each class were used to generate the confusion matrix. The confusion matrix can then be used to estimate the classifier’s accuracy and predictive value (Table 1). For example, 52 of the 200 samples in the test database (recall, n = 50 samples per class) were predicted to belong the dip/green class. Of these, 49 were correct (i.e., true positives) and 3 were incorrect (i.e., false positives). The predictive value of the algorithm in identifying dip/green is then 49/52 = 0.94. The overall classifier accuracy is computed as the arithmetic mean of class-specific accuracies (0.97, in this case).

Statistical Analysis

Agreement in the observer ratings (i.e., inter-rater agreement) among the full set of 175 teat swab images was estimated using Cohen’s kappa coefficient (κ). To permit estimation of intra-rater agreement, 25 randomly selected images were duplicated and rotated 180° (i.e., observers rated a total of 200 images: 175 original and 25 duplicates). The presentation order of the 200 images within the Qualtrics survey software was randomized once (i.e., all raters were presented all 200 images in the same order). The intraclass correlation coefficient was then used to estimate (intra-rater) agreement in the observer ratings of the 25 original and 25 duplicated images.

Each of the 175 original teat swab images was processed using the trained color classifier to obtain estimates of the proportion of each undesirable material class on each swab (i.e., %dirt, %green, %yellow, and %total, where %total is the sum of %dirt, %green, and %yellow). Distributions of the proportions of each undesirable material class were then summarized (using means and SD) for the full set of 175 original images and by quartiles. Additional inter-rater reliability analyses were then performed to estimate the agreement in observer ratings among images assigned to each quartile of each undesirable material class (using Cohen’s κ).

Reliability of the color classifier (i.e., intra-machine reliability) was also assessed. First, the circular region of each swab image was distorted in Adobe Photoshop (Adobe, San Jose, CA) software using the “twirl” tool (with a setting of 90°), and the distorted images were processed using the trained color classifier to obtain new estimates of the proportion of each undesirable material class on each swab. To estimate intra-machine reliability, the strength of the linear relationship between the processing results from original and distorted images was estimated using the Pearson correlation coefficient. The purpose of introducing distortion was to alter the color information within each 3 × 3 pixel area presented to the color classifier but without altering the relative proportions of the colors within the overall circular

region. To quantify the change in the overall relative proportions of colors introduced by the distortion process, histograms of the red, green, and blue color channels were generated (using LabVIEW) for the circular regions of both the original and distorted images. For each color channel, differences in the pixel counts across the full range of color values in the histograms represented <0.25% of the total number of pixels within the circular region (which we consider negligible).

Stata (v. 15.1, StataCorp LP, College Station, TX) was used for all statistical analyses. The Committee for the Protection of Human Subjects and Animal Welfare Committee at the University of Texas Health Science Center at Houston both approved all study procedures.

RESULTS

Inter- and Intra-Rater Reliability

The number of teat swab images each rater assigned to each of the 4 rating scores varied widely (Table 2). For example, the proportion of the 175 original images classified as 4 (i.e., large amount of dirt and manure) by each rater ranged from 3.4 to 20.5%. Raters provided a unanimous score for only 13.5% of the images, whereas the range of scores (maximum − minimum) was at least 2 for more than 25% of the images (data not shown). Overall, the inter-rater agreement in the rating scores for the 175 original images was marginal to poor (κ = 0.43; Table 2). Furthermore, only moderate intra-rater agreement was observed in the rating scores of the 25 randomly selected and duplicated images (intraclass correlation coefficient = 0.74, 95% CI = 0.62–0.85).

Across the full set of 175 original images, the mean proportion of pixels assigned to the undesirable material classes (using the color classifier) was greatest for the premilking dip solution (%yellow = 8.66 ± 10.81%), followed by manure, bedding, and dirt (%dirt = 2.04 ± 5.21%), and then the postmilking dip solution (%green = 0.39 ± 1.23%; Table 2). Examining the material classification results by quartiles underscores the substantial negative skew of the distributions. Consistent with the overall inter-rater agreement in observer rating scores of the full set of 175 images (i.e., κ = 0.43), estimates of inter-rater agreement in observer rating scores of images assigned to each quartile of each undesirable material class were marginal to poor (i.e., κ ranging from 0.21 to 0.49). For the “total” material class, which reflects the overall amount of undesirable material on the swab, inter-rater agreement was higher for images assigned the 1st (κ = 0.44) and 4th quartiles (κ = 0.46) compared with images assigned to the 2nd (κ = 0.29) and 3rd quartiles (κ = 0.27). A similar pattern was observed across quartiles of the “yellow” material class.

Intra-Machine (kNN Classifier) Reliability

Comparison of the material classification results from the 175 original images and 175 distorted images suggested near-perfect intra-machine reliability. Specifically, for each undesirable material class (i.e., dirt, green, yellow, and total), the Pearson correlation coefficient describing the strength of the linear relationship between results from the original and distorted images exceeded 0.99 (Figure 3).

DISCUSSION

Premilking teat and udder sanitization is a vital step in reducing bacteria at the teat end during milking as well as the number of bacteria entering milking equipment, which ultimately can be transferred from one cow to another by the milking machine (Bade et al., 2008). Additionally, inadequate teat cleaning and drying can result in elevated bacterial counts or chemical residues in bulk tank milk, which can compromise milk quality (Galton et al., 1984, 1986; Elmoslemany et al., 2010). Observation-based rating scales (such as the one used in this study) are often used as a quality control metric to evaluate worker performance of cleaning teat ends before milking unit attachment. By evaluating teat ends for debris and chemical residue, parlor management can more effectively inform workers of their teat-preparation effectiveness and consistency.

Like all quality control measurement tools used on a farm, reliability of the measurement tool is paramount. For observation-based rating tools, scoring should be consistent both within and between evaluators. Our findings suggest that scoring of teat swabs from experienced raters (both within and between raters) using the Teat Cleanliness Scorecard is not reliable. Although rating score descriptors provided with the Teat Cleanliness Scorecard (and replicated in our Qualtrics survey) include language pertaining to the presence of dip, manure, and bedding, it was not clear if raters would elect to emphasize one particular material class over the others. The observed results suggest that the extent of inter-rater agreement was not strongly dependent on either the type or amount of material visible on the teat swab images. However, further inspection of Table 3 suggests somewhat greater inter-rater agreement when the amount of material on the swab was either less than 1% (i.e., 1st quartile all material classes and 3rd quartile of green) or more than 25% (i.e., 4th quartiles of yellow and total). This finding suggests that observation-based scoring of teat cleanliness could be reduced to a dichotomous scale (i.e., “clean” or “not clean”). A related question is whether the color classification results agree with the raters’ scores. Unfortunately, the wide variation in rating scores (i.e., consensus rating for only 13.5% of the images) and the absence of a true “gold standard” rating for each teat swab image do not permit a robust evaluation of the association between the raters’ scores and the color classification results.

Our findings suggest worker performance can be inconsistently evaluated by the same or different evaluators. Inconsistent evaluation could result in role ambiguity among parlor workers, which has been shown to compromise job performance (Abramis, 1994). Worker productivity is dependent on the existence of clearly defined standards for evaluating worker performance for a given job task (Shikdar and Das, 2003). Worker performance goal-setting and performance feedback can have motivating effects on worker performance as well as affect worker satisfaction. Research studies have shown that specific performance goals or standards result in higher quality performance among workers (Locke, 1968; Awdia et al., 1996; Phillips and Gully, 1997). However, a reliable evaluation tool is needed to assess performance and provide feedback to dairy parlor workers.

Contrary to observation-based evaluation methods, a machine-based rating system using digital image processing methodologies to objectively quantify the type and amount of teat-end debris demonstrated near-perfect reliability in our study. The use of image processing to

classify and quantify teat-end debris appears promising, as it eliminates bias associated with observation-based rating schemes and introduces a new, data-driven approach to optimize dairy quality control systems. Subjective assessment of teat-end cleanliness is replaced with an objective assessment based on engineering measurements with given accuracy and reliability. Additionally, a machine-based system can provide a nonbiased, reliable method to assess worker performance quality and consistency.

While our study established observational- and machine-based teat scoring reliability, we did not ask raters to indicate if a swab image represented an acceptable level of remaining teat end debris after cleaning and before milking unit attachment. The determination of acceptable levels of debris on a swab post teat preparation is a highly subjective decision based on personal experience and preference, and we anticipate considerable variation in rater judgements on what constitutes an unacceptable degree of remaining debris on a teat end before milking. Future comparisons of observational and machine scoring methods could establish unacceptable debris thresholds based on objective image analysis.

It is important to note that measurement reliability does not imply measurement validity. In our study, we did not measure teat swab bacterial loading, which precluded the correlation of swab image color pixel counts with bacterial types or loads. The exterior of the cow's udder and teats is recognized as one of several sources of microorganisms that are naturally associated with the skin of the animal as well as microorganisms derived from the environment in which the cow is housed and milked (Murphy and Boor, 2010). Future studies should validate machine-based teat swab scoring methods using image analysis against actual teat swab bacterial loads. Such validation could provide an additional mechanism of correlating worker teat-cleaning performance with bulk tank milk quality.

Our findings suggest the utility of machine-based rating of teat swab images for the purpose of worker performance quality control is promising. However, to maximize the likelihood of industry acceptance and usage on the farm, there is a need to develop more general and efficient image capture and processing methods. Digital image analysis technology is now being used in other agricultural, field-based applications using smart mobile phone technology to quantify turfgrass color (Karcher and Richardson, 2003), enumerate flower numbers to estimate grapevine yield (Diago et al., 2014), and estimate banana ripeness (Intaravanne et al., 2012). Sensor technologies using smart mobile phones have the potential to enable powerful "lab-on-smartphone platforms" for important applications across industries (Rateni et al., 2017). We envision teat swab image analysis technology can be further developed and deployed on a smart mobile phone platform for parlor worker training and quality control purposes to enable efficient, reliable, and objective feedback on worker milking performance. As robotic milking is becoming more popular in milking facilities, machine-based teat cleanliness scoring could also be incorporated into these automated milking systems.

Several study limitations should be acknowledged. First, a small sample size of 8 raters to assess observational-based scoring methodology could have resulted in wide variability of rater scoring due to rater biases. Additionally, front-line parlor managers responsible for worker performance supervision were not recruited for participation in the observation-

based rating evaluation. Second, teat swabs were taken from a single dairy, which limited environmental and milking procedure conditions to this farm. Data collection took place when the dairy had clean, dry housing pens. As a result, we were unable to collect teat swabs with higher amounts and variability of debris. Third, we used dry, sterile gauze pads as a sampling medium to wipe teat ends for debris, which may have resulted in incomplete removal and ultimate underestimation of residual debris on teat ends after cleaning. Wiping with gauze pads moistened with sterile, osmotically neutral buffer (e.g., PBS) may be more effective in removing residual debris after teat cleaning and allow for the assessment of bacterial loading.

CONCLUSIONS

Findings from this study suggest that an observational-based method of scoring teat swab debris is not a reliable approach to evaluating teat cleanliness before milking. On the contrary, a machine-based rating system using digital image processing methodologies to objectively quantify type and amount of teat-end debris demonstrated near-perfect reliability in our study. A machine-based teat swab scoring technique may facilitate a more objective comparison of milking performance across workers. Further development of teat swab image analysis, deployed on a mobile phone platform, can potentially provide a mechanism for objective, accurate, and robust machine-based teat cleanliness assessment.

ACKNOWLEDGMENTS

This project represents a collaborative effort by researchers representing two National Institute for Occupational Safety and Health (NIOSH)-funded Agricultural Safety and Health Centers: High Plains and Intermountain Center for Agricultural Health and Safety (HICAHS) at Colorado State University, and the Great Plains Center for Agricultural Health at the University of Iowa.

This work was supported by the Center for Disease Control and Prevention (CDC), NIOSH through the High Plains and Intermountain Center for Agricultural Health and Safety (grant no. U54 OH008085–08), as well as by the Southwest Center for Occupational and Environmental Health (grant no. 5T42OH008421). The contents of this report are solely the responsibility of the authors and do not necessarily represent the official views of the CDC or NIOSH.

REFERENCES

Abramis D, 1994 Work role ambiguity, job satisfaction, and job performance: meta-analyses and review. Psychological Reports 75, 1411–1433.

Arlot S, Celisse A, 2010 A survey of cross-validation procedures for model selection. Statistics Surveys 4, 40–79.

Awdia G, Brown K, Kristof-Brown A, Locke A, 1996 Relationship of goals and microlevel work processes to performance on a multi-path manual task. Journal of Applied Psychology 81, 483–497.

Bade R, Reinemann D, Thompson P, 2008 Method for assessing teat and udder hygiene, ASABE Annual International Meeting, Providence, RI.

Breen J, Bradley A, Green M, 2009 Quarter and cow risk factors associated with a somatic cell count greater than 199,000 cells per milliliter in United Kingdom dairy cows. Journal of Dairy Science 92, 3106–3115. [PubMed: 19528588]

Castro SB, Berthiaume R, Robichaud A, Lacasse P, 2012 Effects of iodine intake and teat-dipping practices on milk iodine concentrations in dairy cows. Journal of dairy science 95, 213–220. [PubMed: 22192200]

Cook N, 2002 The influence of barn design on dairy cow hygiene, lameness, and udder health, 35th Ann. Conv. Amer. Assoc. Bov. Pract. American Assoc of Bovine Practitioners, Madison, WI, pp. 97–103.

Cook N, Reinemann D, 2007 A tool box for assessing cow, udder and teat hygiene, 46th Annual Meeting of the National Mastitis Council, San Antonio, TX, pp. 31–43.

de Pinho Manzi M, Nóbrega D, Faccioli P, Troncarelli M, Menozzi B, Langoni H, 2012 Relationship between teat-end condition, udder cleanliness and bovine subclinical mastitis. Research in Veterinary Science 93, 430–434. [PubMed: 21669449]

Devries T, Aarnoudse M, Barkema H, Leslie K, von Keyserlingk M, 2012 Associations of dairy cow behavior, barn hygiene, cow hygiene, and risk of elevated somatic cell count. Journal of Dairy Science 95, 5730–5739. [PubMed: 22884345]

Diago M, Sanz-Garcia A, Millan B, Blasco J, Tardaguila J, 2014 Assessment of flower number per inflorescence in grapevine by image analysis under field conditions. Journal of the Science of Food and Agriculture 94, 1981–1987. [PubMed: 24302287]

Duda R, Hart P, Stork D, 2000 Pattern classification. Wiley-Interscience.

Ellis K, Innocent G, Mihm M, Cripps P, McLean G, Howard V, Grove-White D, 2007 Dairy cow cleanliness and milk quality on organic and conventional farms in the UK. Journal of Dairy Research 74, 302–310. [PubMed: 17451622]

Elmoslemany A, Keefe G, Dohoo I, Wichtel J, Stryhn H, Dingwell R, 2010 The association between bulk tank milk analysis for raw milk quality and on-farm management practices. Preventive Veterinary Medicine 95, 32–40. [PubMed: 20381889]

French EA, Mukai M, Zurakowski M, Rauch B, Gioia G, Hillebrandt JR, Henderson M, Schukken YH, Hemling TC, 2016 Iodide Residues in Milk Vary between Iodine-Based Teat Disinfectants. Journal of food science 81, T1864-T1870.

Galton D, Petersson L, Merrill W, 1986 Effects of premilking udder preparation practices on bacterial counts in milk and on teats. Journal of Dairy Science 69.

Galton D, Petersson L, Merrill W, Bandler D, Shuster D, 1984 Effects of premilking udder preparation on bacterial population, sediment, and iodine residue in milk. Journal of dairy science 67, 2580–2589. [PubMed: 6520268]

Hani A, 2013 From pixels to medical imaging, IEEE International Conference on Signal and Image Processing Applications, Melaka, Malaysia.

Hovinen M, Aisla AM, Pyörälä S, 2005 Visual detection of technical success and effectiveness of teat cleaning in two automatic milking systems. Journal of Dairy Science 88, 3354. [PubMed: 16107426]

Intaravanne Y, Sumriddetchkajorn S, Nukeaw J, 2012 Cell phone-based two-dimensional spectral analysis for banana ripeness estimation. Sensors and Actuators B: Chemical 168, 390–394.

Karcher D, Richardson M, 2003 Quantifying turfgrass color using digital image analysis. Crop Science 43, 943–951.

Locke E, 1968 Toward a theory of task motivation and incentives. Organizational Behavior and Human Performance 5, 484–500.

Munoz M, Bennett G, Ahlström C, Griffiths H, Schukken Y, Zadoks R, 2008 Cleanliness scores as indicator of Klebsiella exposure in dairy cows. Journal of Dairy Science 91, 3908–3916. [PubMed: 18832213]

Muralidharan R, Chandrasekar C.J.I.J.o.C.T., Technology, 2011 Object recognition using SVM-KNN based on geometric moment invariant. 1, 215–220.

Murphy S, Boor K, 2010 Sources and causes of high bacteria counts in raw milk: an abbreviated review. eXtension.org. Retrieved from https://articles.extension.org/pages/11811/sources-and-causes-of-high-bacteria-counts-in-raw-milk:-an-abbreviated-review Accessed 10th Nov 2018

Neja W, Mariusz B, Małgorzata J, Sawa A, 2016 Effect of cow cleanliness in different housing systems on somatic cell count in milk Acta Veterinaria Brno 85, 55–61.

Phillips J, Gully S, 1997 Role of goal orientation, ability, need for achievement and focus of control in the self-efficacy and goal setting process. Journal of Applied Psychology 82, 792–802.

Rateni G, Dario P, Cavallo F, 2017 Smartphone-based food diagnostic technologies: a review. 17.

Reneau J, Seykora A, Heins B, Endres M, Farnsworth R, Bey R, 2005 Association between hygiene scores and somatic cell scores in dairy cattle. JAVMA 227, 1297–1301. [PubMed: 16266020]

Sant'anna A, Paranhos da Costa M, 2011 The relationship between dairy cow hygiene and somatic cell count in milk. Journal of Dairy Science 94, 3835–3844. [PubMed: 21787920]

Schreiner D, Ruegg P, 2003 Relationship between udder and leg hygiene scores and subclinical mastitis. Journal of Dairy Science 86, 3460–3465. [PubMed: 14672175]

Shikdar A, Das B, 2003 The relationship between worker satisfaction and productivity in a repetitive industrial task. Applied Ergonomics 34, 603–610. [PubMed: 14559421]

Steenwijk M, Pouwels P, Daams M, van Dalen J, Caan M, Richard E, Barkhof F, Vrenken H, 2013 Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). NeuroImage: Clinical 3, 462–469. [PubMed: 24273728]

Figure 1.
Left: hinged clamshell vise used to secure teat swab; middle: photo box with camera frame setup; right: teat swab image.

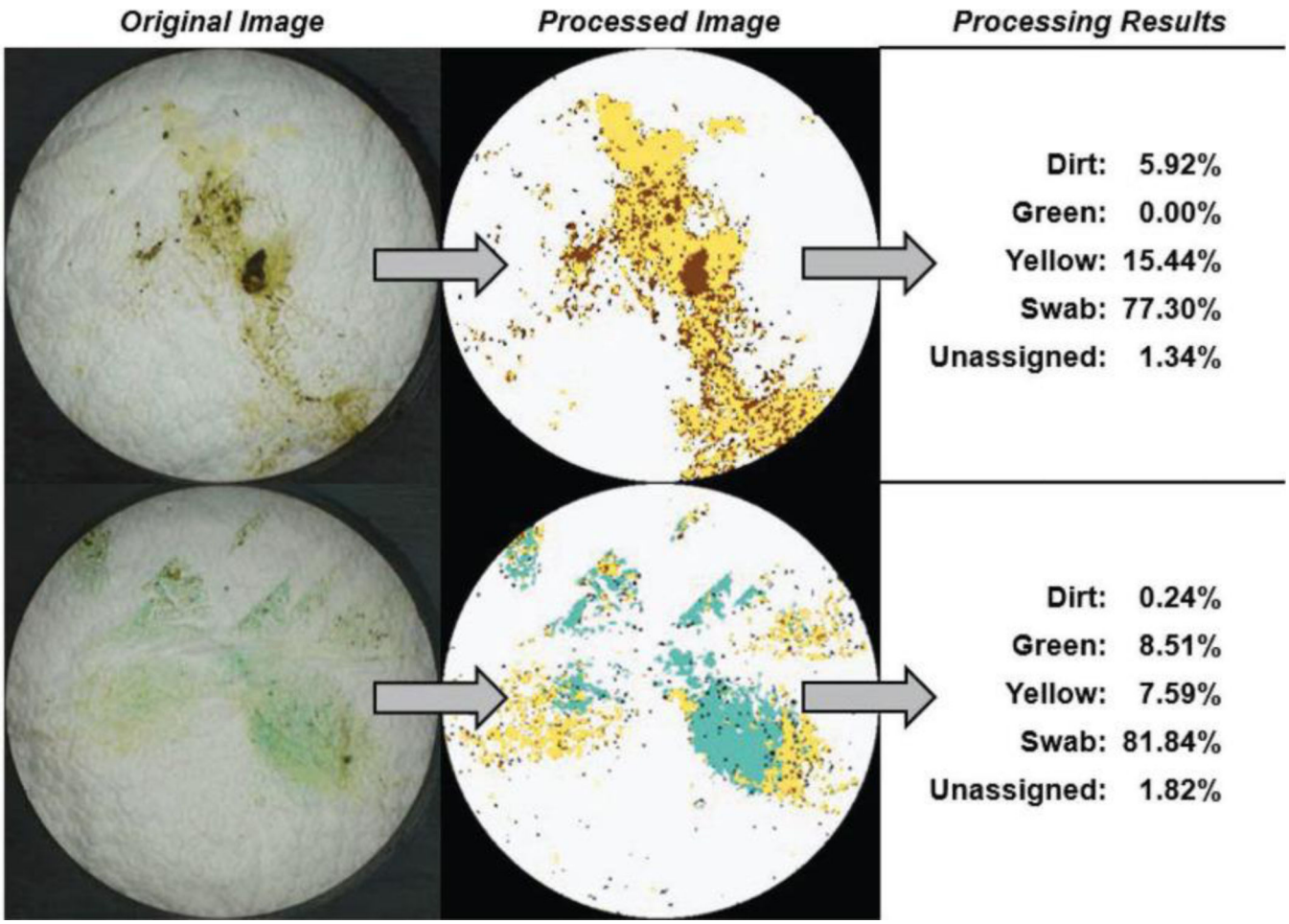

Figure 2.
Example of teat swab original and processed images with processing results.

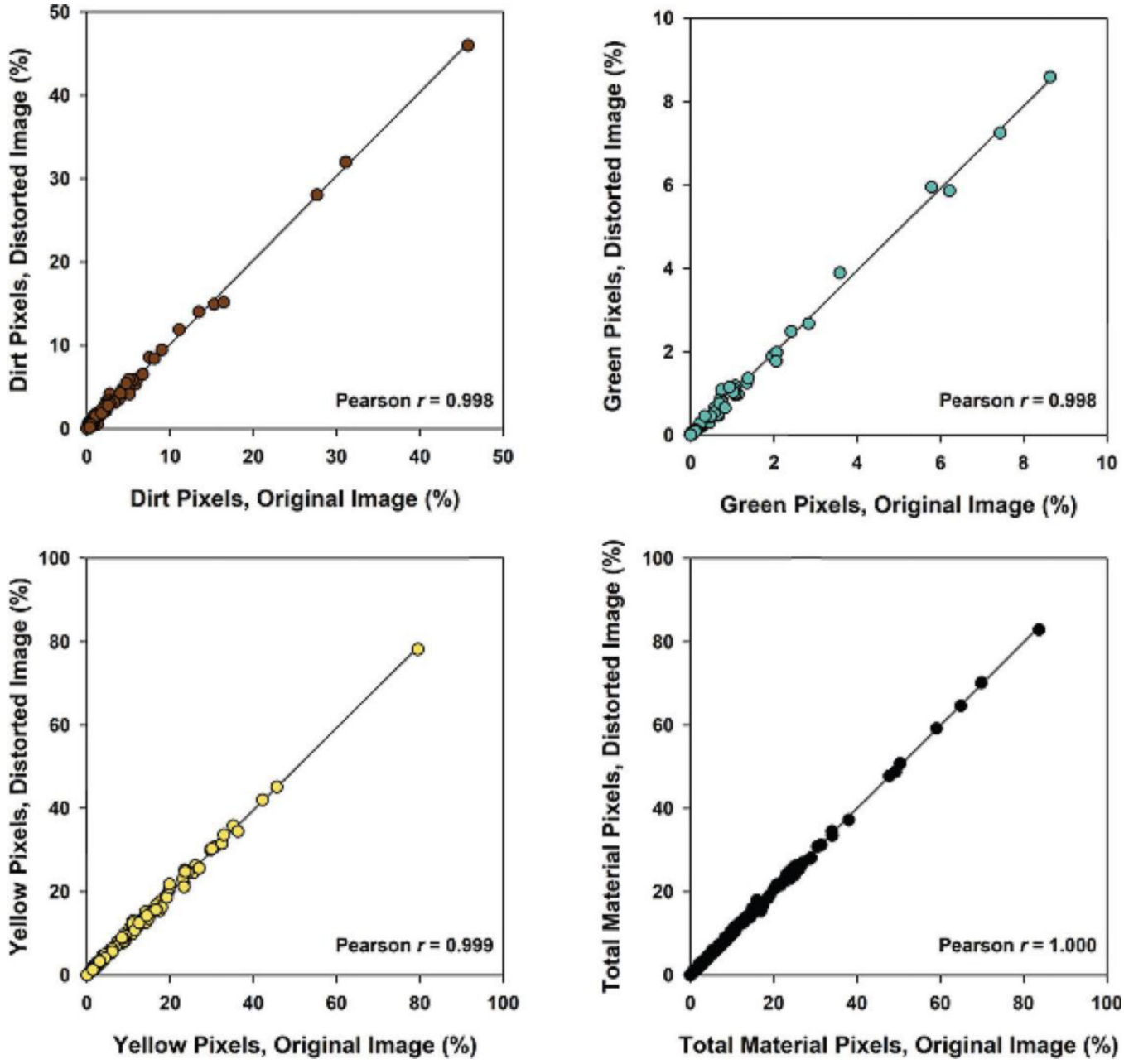

Figure 3.
Scatterplots and Pearson correlation coefficients of material classifications of original versus distorted images (n = 175).

Table 1.

Performance of the *k*-nearest neighbor classifier

	Predicted Class					
Assigned Class	**Dip/Green**	**Dip/Yellow**	**Dirt**	**Wipe**	**Total**	**Accuracy**
Green	49	0	0	1	50	49/50 = 0.98
Yellow	3	45	2	0	50	45/50 = 0.90
Dirt	0	0	50	0	50	50/50 = 1.00
Wipe	0	0	1	49	50	49/50 = 0.98
Total	52	45	53	50	200	
Predictive Value	49/52 = 0.94	45/45 = 1.00	50/53 = 0.94	49/50 = 0.98		

Table 2.

Cross-tabulation of raters' scores and overall inter-rater agreement of 175 teat-swab images[1]

	Rating Scale Score				
	1	**2**	**3**	**4**	**Total**
Rater	**n (%)**	**n (%)**	**n (%)**	**n (%)**	**N**
1	13 (7.4%)	59 (33.7%)	67 (38.2%)	36 (20.5%)	175
2	18 (10.2%)	56 (32.0%)	81 (46.2%)	17 (9.7%)	172
3	28 (16.0%)	64 (36.5%)	77 (44.0%)	6 (3.4%)	175
4	22 (12.5%)	77 (44.0%)	69 (39.4%)	7 (4.0%)	175
5	45 (25.7%)	36 (20.5%)	52 (29.7%)	42 (24.0%)	175
6	49 (28.0%)	86 (49.1%)	28 (16.0%)	10 (5.7%)	173
7	31 (17.7%)	67 (38.2%)	66 (37.7%)	11 (6.2%)	175
8	12 (6.8%)	47 (26.8%)	83 (47.4%)	33 (18.8%)	175
Total	218 (15.6%)	492 (35.2%)	523 (37.4%)	162 (11.6%)	1395

[1]Overall inter-rater agreement: Cohen's kappa = 0.43

Table 3.

Mean (SD) of percent material class with rater score agreement within quartiles (n = 175 nonrepeated teat swab images)

Material Class[1]	N (%)	Mean (sd), in %[2]	Cohen's kappa[3]
Dirt			
All images	175 (100.0)	2.04 (5.21)	
1st quartile	44 (25.1)	0.03 (0.03)	0.42
2nd quartile	44 (25.1)	0.23 (0.11)	0.34
3rd quartile	44 (25.1)	0.92 (0.40)	0.25
4th quartile	43 (24.6)	7.10 (8.78)	0.31
Green			
All images	175 (100.0)	0.39 (1.23)	
1st quartile	103 (58.4)	0.00 (0.00)	0.43
2nd quartile[4]	--	--	--
3rd quartile	30 (17.0)	0.03 (0.03)	0.46
4th quartile	42 (24.0)	1.60 (2.10)	0.33
Yellow			
All images	175 (100.0)	8.66 (10.81)	
1st quartile	44 (25.1)	0.39 (0.39)	0.49
2nd quartile	44 (25.1)	2.42 (0.92)	0.33
3rd quartile	44 (25.1)	8.75 (2.46)	0.21
4th quartile	43 (24.6)	23.43 (11.89)	0.49
Total[5] (non-swab)			
All images	175 (100.0)	11.10 (13.68)	
1st quartile	46 (26.3)	0.71 (0.61)	0.44
2nd quartile	42 (24.0)	3.68 (1.51)	0.29
3rd quartile	44 (25.1)	11.11 (2.60)	0.27
4th quartile	43 (24.6)	29.44 (15.74)	0.46

[1]Quartile of distribution of material class estimated using kNN classifier.

[2]Mean(sd) of proportion of pixels assigned to material class using kNN classifier.

[3]Cohen's kappa (e.g. inter-rater reliability) computed separately among images in each quartile of each material class.

[4]No 2nd quartile because more than 50% of the images contained no Green material (i.e., mean (sd) of %Green = 0.00 (0.00) for 103 of 175 images) and as a result fall into the first quartile

[5]Total = sum of Dirt, Green, and Yellow material identified in the image