Short communication: Design and validation of a hygiene score for calves

L. M. Kellermann,* A. Rieger, G. Knubben-Schweizer, and M. Metzner
Clinic for Ruminants with Ambulatory and Herd Health Services, Centre for Clinical Veterinary Medicine, Ludwig-Maximilians-Universität Munich, Sonnenstrasse 16, 85764 Oberschleissheim, Germany

ABSTRACT

The objective of this study was the establishment and validation of a scoring system for calf dirtiness. Defined areas of the belly, side, and rear of the calves were scored according to the percentage of dirtiness: score 1 ranges from 0 to 10% of the area (no or little soiling), score 2 from over 10 to 30% (medium soiling), and score 3 applies to soiling of more than 30% of the area (heavy soiling). The scores of the individual regions were summed up to yield the calf hygiene score (CHS) ranging from 3 to 9. The validation of the CHS was performed by 5 veterinarians on 42 calves. It was validated for its inter- and intra-observer-reliability and against a standard method created by image processing of photographs of calves. The agreement between the observers and the standard method was weak to moderate with kappa values from 0.58 to 0.67. Inter- and intra-observer reliability resulted in a moderate to strong agreement with 29 of 36 kappa values between 0.60 and 0.89. The CHS was thus successfully validated as reliable and could be a useful tool for herd health management.

Key words: calf, hygiene, score, reliability, validation

Short Communication

The handling and housing of livestock is of key interest to politicians, farmers, and consumers. Veterinarians aim to improve animal welfare and to prevent disease, while balancing economic, ethical, and food safety interests. Cleanliness and hygiene are important factors in achieving these aims. Different methods have been used to describe the effect of hygienic conditions on calves’ health. Roth et al. (2009) and Calvo-Lorenzo et al. (2016) described the dirtiness caused by feces by observing the amount and consistency of the feces. Jorgensen et al. (2017) additionally differentiated between different regions (perianal region, underside of the tail, and tailhead). Panivivat et al. (2004) used a scoring system where tail head region, thighs, and legs were counted when soiled with manure. This scoring has been applied by Sutherland et al. (2013, 2014, 2017) and Graham et al. (2018). Other studies assessed different body regions and their degree of dirtiness (score or percentage): Terosky et al. (1997) and Klein-Jöbstl et al. (2014) focused on legs, thighs, and the belly; whereas Webster et al. (1985) also included the back and the head. Lundborg et al. (2005) used a very detailed scoring for 8 different body parts of calves, where the percentage of contamination results in a score. To prove that a scoring system produces reliable and valid data, it should be statistically validated, including tests for inter-observer and intra-observer reliability (Meagher, 2009). In dairy science we use various rating systems with a statistically proven functionality, for example, BCS (Edmonson et al., 1989) or lameness scoring (Wells et al., 1993). To the best of our knowledge, no statistically tested scoring system for the hygiene of calves has been published so far. The purpose of this study was therefore to design a calf hygiene score (CHS) that is easy to apply in the field and has a high inter- and intra-observer agreement. Furthermore, the CHS was validated against a standard method (SM) created in an image-based evaluation system. In this short communication, we explain the design of this CHS and present the results of the validation.

For the scoring system, 3 risk zones (RZ) are considered, and each RZ is assigned an individual score. The summation of the RZ results in the CHS. The 3 evaluated RZ are the belly (RZB), the side (RZS), and the rear (RZR; Figure 1). For the evaluation of RZB and RZS, the examiner is on the side of the calf at an angle of 90° and in the middle of the longitudinal axis of the calf, at a distance of 0.5 to 2 m. The examiner should be in a squatting position to keep the eye level at the height of the back line. For the evaluation of RZR, the examiner is at the same distance and posture, located behind the calf in the middle of its longitudinal axis. The RZB is defined as the visible area below a line from the elbow to the knee fold, cranially limited by the foreleg and caudally by the hindleg (Figure 1a). The RZS...
is the visible area caudally of a line orthogonal to the ground through the knee fold. The tail is excluded and only the lateral view is evaluated (Figure 1b). The RZR is the visible area from the tail attachment down to the proximal end of the tarsal joints and the complete tail (including the switch; Figure 1c). The evaluation of the percentage of dirtiness of the surfaces is expressed as a score from 1 to 3 for each individual RZ. Score 1 ranges from 0 to 10% of the area (no or little soiling), score 2 from over 10 to 30% (medium soiling), and score 3 applies to soiling of more than 30% of the area (heavy soiling). All types of soiling must be considered, except loose straw or loose particles in the haircoat, freshly defecated excrements, or a wet coat assumed to be due to rain or drinking water. The RZB and RZS are examined from the right and the left side of the calf and the results of the more soiled RZB and RZS are each documented, such that only the dirtier side of the calf contributes to the score. Schematic drawings were handed out as a scorecard to the examiners for reference (see Figure 1). The scores of the individual regions were summed up to yield the CHS ranging from 3 to 9.

A total of 42 Holstein calves were included in the study for evaluation of inter-observer reliability, 41 of them also for the evaluation of intra-observer reliability. One calf was excluded from the intra-observer reliability calculation because it was newborn and changed its appearance from wet to dry between the scoring rounds, so that a comparison within the different rounds of each observer could not be performed. The calves were suckling calves at the age of 0 to approximately 12 wk housed in individual pens or in group igloos, and all belonged to a commercially operating dairy farm in Bavaria. The evaluation took place on 1 d, with 5 veterinarians (observers). Each calf was rated by each observer individually 3 times in a different order within 4 h to create a data set for calculation of inter- and intra-observer reliability. Additionally, within this time window standardized photographs of all calves were taken for each RZ including both sides, from the same observation position as defined above. All 210 pictures were used to establish a SM for the validation. Photoshop CC 2017 (Adobe Systems Inc., San Jose, CA) was used to calculate the actual proportions of soiled surfaces in the photographed RZ of the calves. In brief, the specified area of the respective RZ and the soiled areas on it were marked manually with a polygon-tool and the numbers of pixels were displayed. The sum of pixels of all soiled areas was calculated and divided by the number of pixels of the RZ. The percentage of pixels belonging to the soiled area was assigned to a score and compared with the score awarded for this region by the observers. The verification of the schematic drawings of Figure 1 for a correct percentage of soiled area within the respective RZ was performed with the same graphical method.

A kappa statistic according to the method of Ruddat (Ruddat et al., 2014) using RStudio version 1.1.453 (RStudio Inc., Boston, MA) with R version 3.4.4 (R Core Team, Vienna, Austria) was calculated. This kappa method is suitable for ordinal data and for more than 2 observers. Kappa statistics for the ratings of the individual RZ and the CHS were calculated. The global kappas with 95% confidence interval are displayed in Table 1 for each observer (intra-observer) and in Table 2 for each round (inter-observer). The reference results generated by image editing were included as a sixth observer and represent the SM. Results of the respective exclusion test are also presented in Table 2 as a kappa statistic with 95% confidence interval to identify if the observers 1–5 agree with the SM. Kappa values can range from −1 to 1. Cut-off values for kappa are arbitrary (Landis and Koch, 1977), but for health-related scores the interpretation according to McHugh (2012) can be used: kappa values above 0.80 represent strong observer agreement, values from 0.60 to 0.79 indicate moderate agreement, and values from 0.40 to 0.59 represent weak agreement. Lower values would indicate no or minimal agreement.

The results show that the application of the CHS was highly reliable: 29 of 36 determined kappa values show a moderate to strong inter- or intra-observer agreement. For the CHS, the average intra-observer agreement was a kappa of 0.83 and the average inter-observer agreement was 0.77. Within the RZ, only the RZB showed a noticeably weaker agreement. The validation against the SM showed a weak to moderate agreement. The mean value of all assigned CHS scores for each calf (from each observer for 3 rounds) was calculated and rounded to full digits. In comparison to these 42 CHS mean values, the CHS score assigned by the SM yielded 81% higher and 19% equal values. This was also performed for the scores of each RZ, which resulted in 26% higher and 74% equal scores for RZB and 60% higher and 40% equal scores for RZS and RZR. No SM score was lower than the mean value of the observers’ CHS scores. Further evaluations of the SM scoring results show that 22 of 42 calves would obtain a lower score for the RZB or RZS or even both RZ if they would have been scored on the cleaner side.

We identified several factors that could contribute to the weaker agreements detected for the RZB and for the differences between the SM and observers. Especially for the RZB, the specified body position is very important. We cannot exclude the possibility that the observers changed the exact height of their eye position slightly during the assessment. Additionally, changes to the body position of the calf alter the observable sur-
The imprecision and inaccuracy of human perception is a strong but inevitable source of error in optical scoring. The selected categories of percent were chosen because they correspond to what we encountered in the field considering calves on farms. The advantage of a method using percent for assessments is that the number itself has an unambiguous value. Methods using the size of the hand amplify...
the inaccuracy that both the hand and the evaluated animal can be of very different sizes. Descriptions such as “inconspicuous,” “little,” or simply “soiled” are not strictly determined and are highly subjective. To facilitate the visual assessment, charts have been used as an additional tool for reference (Edmonson et al., 1989; Cook, 2002; Reneau et al., 2005). The most common classification of burn wounds in human medicine is described by the Lund and Browder chart (Lund and Browder, 1944), where they divide the body surface in different parts of different importance. Likewise, we defined the areas to be assessed for the CHS precisely on the scorecard and with words, since the mere description of “hind” or “side” could lead to mistakes. The area of RZS and RZB overlap to a small extent. Due to the shape of the calf, we did not obtain clearer boundaries. The overlap was not taken into account in our calculations. Our experience has been that the scoring of newborn calves leads to difficulties (e.g. when they are still wet). We recommend excluding calves born in the last 24 h from this hygiene scoring. The colorization of calves involves the problem that dirt could be less visible on black parts. In order to keep the scoring simple, there were no measurements of the color proportions of the coat. To detect the condition of the coat, sufficient lighting is necessary. We recommend performing the evaluation only if one could read a newspaper text at the calf’s location without any problems. The use of light sources such as flashlights or the evaluation in a dark environment have not been tested and may lead to different scoring results. A convincing scoring system should assess the hygienic condition of the calf as a whole. Mud, urine, milk residues, dust, and other materials should not be ignored because the impact on health is not clear. The definition of dirt for our scoring system is therefore not limited to soiling by feces and was clearly described. To obtain reliable results, it is recommended that the observer try to wipe loose parts off the coat once. Loose straw often hides excrement to which the straw adheres. Nevertheless, the selection of the relevant parts is sometimes challenging (e.g. when only the tips of the hairs are dirty or only slight color changes are visible). Those difficulties were the same for the evaluation of the pictures as for the observer on the farm. However, when creating the SM, we evaluated photographs and were able to zoom in and out in various ways, whereas the observer assessed moving calves in a short period. As a result, it is likely that in the analysis of photographs, even small dirt spots were easier to recognize and therefore a higher percentage of

Table 1. Global kappa values (95% CI) of the intra-observer validation of a calf hygiene score (CHS) by 5 observers in 3 rounds considering 3 risk zones: belly (RZB), side (RZS), and rear (RZR) of n = 41 calves

<table>
<thead>
<tr>
<th>Assessed zone</th>
<th>Observer</th>
<th>Observer</th>
<th>Observer</th>
<th>Observer</th>
<th>Observer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>RZB</td>
<td>0.68</td>
<td>0.51</td>
<td>0.62</td>
<td>0.78</td>
<td>0.43</td>
</tr>
<tr>
<td>(0.46; 0.91)</td>
<td>(0.23; 0.78)</td>
<td>(0.40; 0.83)</td>
<td>(0.63; 0.93)</td>
<td>(0.13; 0.73)</td>
<td></td>
</tr>
<tr>
<td>RZS</td>
<td>0.79</td>
<td>0.81</td>
<td>0.75</td>
<td>0.78</td>
<td>0.56</td>
</tr>
<tr>
<td>(0.67; 0.91)</td>
<td>(0.70; 0.92)</td>
<td>(0.56; 0.93)</td>
<td>(0.68; 0.89)</td>
<td>(0.37; 0.74)</td>
<td></td>
</tr>
<tr>
<td>RZR</td>
<td>0.85</td>
<td>0.84</td>
<td>0.82</td>
<td>0.82</td>
<td>0.68</td>
</tr>
<tr>
<td>(0.76; 0.94)</td>
<td>(0.73; 0.94)</td>
<td>(0.72; 0.94)</td>
<td>(0.69; 0.94)</td>
<td>(0.55; 0.80)</td>
<td></td>
</tr>
<tr>
<td>CHS</td>
<td>0.88</td>
<td>0.87</td>
<td>0.82</td>
<td>0.89</td>
<td>0.70</td>
</tr>
<tr>
<td>(0.80; 0.95)</td>
<td>(0.83; 0.91)</td>
<td>(0.70; 0.95)</td>
<td>(0.83; 0.95)</td>
<td>(0.58; 0.83)</td>
<td></td>
</tr>
</tbody>
</table>

CHS = sum of rating from RZB, RZS, and RZR.

Table 2. Global kappa values (95% CI) of the inter-observer validation of a calf hygiene score (CHS) by 5 observers in 3 rounds considering 3 risk zones: belly (RZB), side (RZS), and rear (RZR) of n = 42 calves, and a comparison with the standard method (SM)

<table>
<thead>
<tr>
<th>Evaluation unit</th>
<th>RZB</th>
<th>RZS</th>
<th>RZR</th>
<th>CHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round 1 (observer 1–5)</td>
<td>0.57</td>
<td>0.67</td>
<td>0.81</td>
<td>0.82</td>
</tr>
<tr>
<td>(0.41; 0.73)</td>
<td>(0.52; 0.81)</td>
<td>(0.72; 0.90)</td>
<td>(0.76; 0.90)</td>
<td></td>
</tr>
<tr>
<td>Round 2 (observer 1–5)</td>
<td>0.59</td>
<td>0.60</td>
<td>0.63</td>
<td>0.74</td>
</tr>
<tr>
<td>(0.41; 0.77)</td>
<td>(0.56; 0.81)</td>
<td>(0.56; 0.75)</td>
<td>(0.63; 0.85)</td>
<td></td>
</tr>
<tr>
<td>Round 3 (observer 1–5)</td>
<td>0.56</td>
<td>0.70</td>
<td>0.68</td>
<td>0.76</td>
</tr>
<tr>
<td>(0.43; 0.70)</td>
<td>(0.58; 0.81)</td>
<td>(0.57; 0.79)</td>
<td>(0.69; 0.84)</td>
<td></td>
</tr>
<tr>
<td>Total2 (observer 1–5)</td>
<td>0.58</td>
<td>0.67</td>
<td>0.70</td>
<td>0.77</td>
</tr>
<tr>
<td>(0.49; 0.66)</td>
<td>(0.59; 0.76)</td>
<td>(0.64; 0.77)</td>
<td>(0.71; 0.84)</td>
<td></td>
</tr>
<tr>
<td>Total (observer 1–5) plus SM3</td>
<td>0.58</td>
<td>0.60</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td>(0.49; 0.67)</td>
<td>(0.52; 0.68)</td>
<td>(0.53; 0.67)</td>
<td>(0.60; 0.73)</td>
<td></td>
</tr>
</tbody>
</table>

1CHS = sum of rating from RZB, RZS, and RZR.
2Total = values for rounds 1, 2, and 3 combined.
3The SM was created by image processing of photographs of calves.
dirty areas was calculated. The kappa values of total (observer 1–5) plus SM demonstrate that the scoring still provides a sufficiently realistic image. Whilst the SM gives a more accurate estimate of absolute soiling, hygiene scoring is normally applied in order to assess animals relative to one another. Thus, the observer-based CHS is a useful tool for assessing hygiene in the field and is reliable as shown by the observer agreement values. Graham et al (2018) received a weighted kappa of only 0.22 for agreement between the scoring for diarrhea and the scoring for hide cleanliness. The CHS should therefore not be used to predict diarrhea in calves. Under field conditions, calves would often only be seen from one side. This side can be the cleaner or the dirtier side, and that for RZB and RZS. For reliable scoring it is necessary to examine both sides of the calves. Doing otherwise could change the scoring results to a lower CHS, as described in the results.

Goldberg et al. (2014) compared a mobile app with the chart previously mentioned by Lund and Browder (1944). Users marked the area by hand on a 3-dimensional model in the app, which then calculated the surface area. The results obtained were more accurate and the app was reported to be easier to apply than the conventional method. The development of an app could also be useful in the field of calf hygiene. The definitions of the RZ and other rules set up for the CHS could facilitate the creation of such an app in the future.

In conclusion, we have created an easy-to-use and accurate score for the hygiene of calves. By defining dirtiness, the regions that are assessed, and the method of how to score them, we obtained good observer agreement. Through this validation, the CHS is qualified as a reliable method to assess the hygienic condition of calves in monitoring herd health. The extent of the relationship between the score and calf health should be evaluated in further studies.

ACKNOWLEDGMENTS

Inga Ruddat (Hannover, Germany) provided us with valuable information about her methods of statistical analysis. We are grateful to Helen Fox (University of Cambridge alumnus) for her proofreading and advice. We also thank the observers of our validation for their time and high degree of concentration. The authors have not stated any conflicts of interest.

REFERENCES


pvem.2004.11.014.


Sutherland, M. A., G. M. Worth, and M. Stewart. 2014. The effect of rearing substrate and space allowance on the behavior and physiol-

ORCIDs
L. M. Kellermann @ https://orcid.org/0000-0002-3423-8717
A. Rieger @ https://orcid.org/0000-0002-6826-4497
G. Knubben-Schweizer @ https://orcid.org/0000-0002-0928-5933
M. Metzner @ https://orcid.org/0000-0003-0752-6059