J. Dairy Sci. 100:1–5
https://doi.org/10.3168/jds.2016-11872

Short communication: An observational study investigating inter-observer agreement for variation over time of body condition score in dairy cows

P.-A. Morin,*[1] Y. Chorfi,† J. Dubuc,* J.-P. Roy,* D. Santschi,‡ and S. Dufour§
*Faculté de Médecine Vétérinaire, Département des Sciences Cliniques, and
†Faculté de Médecine Vétérinaire, Département de Biomédecine Vétérinaire, Université de Montréal, 3200 rue Sicotte, CP 5000, Saint-Hyacinthe, QC, Canada, J2S 7C6
‡Valacta, Ste-Anne-de-Bellevue, QC, Canada, H9X 3R4
§Faculté de Médecine Vétérinaire, Département de Pathologie et de Microbiologie, Université de Montréal, 3200 rue Sicotte, CP 5000, Saint-Hyacinthe, QC, Canada, J2S 7C6

ABSTRACT

Body condition score (BCS) is strongly correlated with energy reserves. The ease, rapidity of scoring, and high intra- and inter-observer repeatability make it a widely used herd management tool in bovine practice and in scientific studies. Loss or gain of BCS, rather than a single BCS measurement, is frequently used to monitor energy balance in dairy cows. It is unknown if the difference between 2 BCS measures taken at different moments (ΔBCS) would demonstrate inter-observer agreement similar to that of a single BCS measurement. The objective of this study was to compare inter-observer agreement of BCS and ΔBCS in dairy cows when multiple observers perform data collection. An observational study was conducted between April and September 2015; 3 observers independently assessed BCS of 73 Holstein cows from 1 commercial dairy herd. Body condition score assessments of the animals were performed between 1 and 20 d in milk (early lactation; exam 1) and again between 41 and 60 d in milk (peak of milk production; exam 2). Quadratic weighted kappa (κ_w) was computed to quantify agreement between observers for single BCS measurements and ΔBCS. For single BCS measurements, κ_w of 0.79 (95% CI: 0.69, 0.85) and 0.84 (95% CI: 0.77, 0.89) were obtained for exam 1 and exam 2, respectively. Such values would be interpreted as strong agreement and are consistent with the available literature on BCS repeatability. When computing agreement for ΔBCS, a κ_w value of 0.49 (95% CI: 0.32, 0.63) was obtained, suggesting moderate agreement between observers. These findings suggest that studies investigating single BCS measures could use many observers with a high degree of accuracy in the results. When ΔBCS is the parameter of interest, more reliable results would be obtained if one observer conducts all assessments.
Key words: dairy cow, body condition score, inter-observer agreement, kappa

Short Communication

In the early 1980s, Wildman (1982) developed a scale from 1 to 5 to evaluate body condition of dairy cattle at any stage of lactation regardless of BW and size. The strong association between BCS of dairy cows and body energy reserves (Wright and Russel, 1984) and its ease of implementation as a herd management tool, led to its adoption in dairy herd management, as well as in scientific studies (Roche et al., 2009). To optimize data quality, most professionals prefer to have only one evaluator conducting all the BCS measurements (Kristensen et al., 2006). In large field trials conducted over a long period of time, it is not always possible to have only one observer assessing all BCS. The relatively high inter-observer agreement reported among experienced observers in previous studies could suggest that the use of more than one observer would have little effect on the accuracy of BCS measures (Ferguson et al., 1994; Kristensen et al., 2006). Measurement of BCS is usually performed during the peri-partum transition period to qualify loss or gain of energy reserves. For that purpose, difference of BCS (**ΔBCS**) is used and calculated by subtracting the most recent BCS value from its previous measurement for the same cow. Despite the wide use of ΔBCS for monitoring energy balance, no studies have reported the inter-observer repeatability of this measurement. The hypothesis is that, if disagreement between observers is systematic when conducting single BCS measurement, then ΔBCS could yield higher agreement than single BCS. On the other hand, if the difference between observers is random instead of systematic, then the ΔBCS, requiring 2 BCS measure-

Received August 15, 2016.
Accepted December 21, 2016.
[1]Corresponding author: a.morin@umontreal.ca

ments for computation, could potentially yield lower agreement. The objective of the current study is to evaluate inter-observer agreement of ΔBCS computed by subtracting a BCS measure observed at the peak of milk production from another BCS measure observed during the early lactation, and to compare it with inter-observer agreement of single BCS measures.

The study protocol was presented to the Research Ethical Committee of the Université de Montréal (Saint-Hyacinthe, QC, Canada). Because animals were only observed, the committee approved the current research without the need for a certificate. An observational repeatability study was conducted from April to September 2015 in a single commercial dairy herd of 240 lactating Holstein cows conveniently selected and located in the vicinity of the Bovine Ambulatory Clinic of the Faculté de Médecine Vétérinaire of the Université de Montréal (St-Hyacinthe, QC, Canada). Cows in this herd were housed in a free stall barn, fed a TMR, and milked by milking robots. Mean annual herd milk production was 11,000 kg per cow per year.

Sample size calculation was based on Rotondi and Donner (2012) using the package kappaSize developed for the R software 3.2.3 (The R Project for Statistical Computing, Vienna, Austria). For this analysis, the following parameters were used: an expected quadratic weighted kappa statistic (κ_w) value of 0.86 (Kristensen et al., 2006); a lower bound of 0.7 corresponding to the median of κ_w in the strong agreement category (Landis and Koch, 1977b); a higher bound set to "Not Available," allowing the procedure to generate the number of required subjects for a 1-sided confidence interval; 3 observers; a desired type 1 error rate of 0.05; a normal distribution of BCS and ΔBCS in the population; and finally, 5 possible categories of BCS (Wildman, 1982) with respective prevalence of 0.01, 0.3, 0.38, 0.3, and 0.01. With those settings, a total of 38 cows were estimated. The R package used for the current sample size estimation can only accommodate measurements in less than or equal to 5 categories. Authors were aware that the BCS scale they used included 9 categories (Ferguson et al., 1994). The power of a study generally increases as the number of categories for the variable evaluated increases, thus requiring a smaller sample size (Cohen, 1983). The sample size estimated for the current study was higher than the sample size truly needed. Nevertheless, the estimate of 38 cows was considered a minimum sample size due to the ease of collecting BCS measurements.

Before farm sampling began, the observers, one veterinarian (observer 3) and 2 animal health technicians (observers 1 and 2), reviewed the BCS chart of Elanco (Greenfield, IN) based on the works of Wildman (1982) and Ferguson et al. (1994). Observers 1, 2, and 3 had, respectively, 15, 2, and 9 yr of experience at BCS scoring of dairy cows. Furthermore, observer 3 initially trained observer 2 on BCS scoring. Cows were systematically enrolled as they calved and examined simultaneously and independently by the 3 observers at 2 different moments. A first evaluation (exam 1) was conducted in the first 3 wk of lactation, and a second evaluation (exam 2) was performed between 6 and 8 wk of lactation. Observers were blinded to BCS values reported at first observation, and to the current and previous values from other observers. A minimum interval of 28 d between evaluations was enforced. Sampling cows in early lactation was chosen because this period is critical in term of energy deficiency, lipolysis, and weight loss (Smith and McNamara, 1990; Renaville et al., 2002; Lucy et al., 2009). Measures of BCS, and consequently subcutaneous fat, are generally at their nadir between 40 and 100 DIM (McNamara, 1991; Pedron et al., 1993; Gillund et al., 2001), which generally reflects equality in energy inputs and outputs.

For each cow, the difference between BCS measurements obtained at exam 1 and exam 2 (ΔBCS) was computed for each observer using the 2 collected measurements. Descriptive statistics (mode, minimum, maximum, median, lower, and upper quartiles) were computed for BCS measures obtained at exam 1, at exam 2, and for ΔBCS using the MEANS procedure of SAS (version 9.4, SAS Institute Inc., Cary, NC). Descriptive statistics were also computed for each observer. Scatter plots (SGPLOT procedure in SAS) comparing results across observers were built for single BCS measures observed at exam 1 and exam 2, as well as for ΔBCS, to visually compare agreement between all possible pairs of observers and with the equality line corresponding to perfect agreement (Dohoo et al., 2003).

Body condition score is a qualitative ordinal measurement (Wildman, 1982; Ferguson et al., 1994). The κ_w statistic (Sim and Wright, 2005) was chosen to report agreement beyond chance so that more weight is attributed to large measurement differences than to small ones. For each pair of observers, agreements at exam 1, at exam 2, and for ΔBCS were estimated using κ_w from the FREQ procedure of the SAS software (version 9.4, SAS Institute Inc., Cary, NC). The Bowker's test of symmetry (Bowker, 1948), testing for equal κ_w coefficients for multiple strata with multiple categories, was used to assess heterogeneity between pairs. When this test was statistically nonsignificant, an overall κ_w was produced (Kristensen et al., 2006). Based on Fleiss (1971) and Landis and Koch (1977a), a κ_w value comparing the 3 observers together is equivalent to the weighted average of the individual pairs. These overall κ_w values (Barnhart et al., 2002) were calculated using