J. Dairy Sci. 101:1–10
https://doi.org/10.3168/jds.2017-13611

Reliability of sampling strategies for measuring dairy cattle welfare on commercial farms

Jennifer M. C. Van Os,* Christoph Winckler,† Julia Trieb,† Soraia V. Matarazzo,‡ Terry W. Lehenbauer,§ John D. Champagne,§ and Cassandra B. Tucker*[1]
*Center for Animal Welfare, Department of Animal Science, University of California, Davis 95616
†Department of Sustainable Agricultural Systems, Division of Livestock Sciences, University of Natural Resources and Life Sciences, Vienna, Austria, 1180
‡Department of Agricultural and Environmental Sciences, Santa Cruz State University, Ilhéus, Brazil, 45662
§Veterinary Medicine Teaching and Research Center, University of California-Davis, Tulare 93274

ABSTRACT

Our objective was to evaluate how the proportion of high-producing lactating cows sampled on each farm and the selection method affect prevalence estimates for animal-based measures. We assessed the entire high-producing pen (days in milk <100; range = 81–241 cows) on 10 California farms using measures from the Welfare Quality Protocol for Cattle. Cows were restrained in head locks and visually evaluated for body condition, dirtiness, skin alterations (hair loss, lesions, or swelling), discharge (ocular, nasal, vulvar), diarrhea, and impaired respiration. Lameness was scored upon release. Prevalence was calculated as a percentage of assessed cows. The most common conditions were dirty hindquarters (33.5 ± 10.7%, mean ± standard deviation) and lesions or swelling on the carpal joint (34.4 ± 7.0%) and hock (26.4 ± 16.7%). Diarrhea (8.0 ± 5.8%), lameness (moderate = 7.3 ± 4.7%, severe = 2.2 ± 2.2%), and neck (5.8 ± 12.6%), flank (4.5 ± 5.0%), or hindquarter alterations (5.5 ± 3.9%) were less common. Very fat cows, vulvar discharge, and impaired respiration were rare (≤1%) and were excluded from further analysis. Four sampling strategies were used to generate 20 estimates for each animal-based measure. The strategies were (1) selecting every 10th, 5th, 4th, 3rd, 2nd, 2 of 3, or 3 of 4 cows at the feed bunk (7 estimates/measure); (2) randomly selecting 7 matching proportions of the pen; (3) randomly selecting cows using 3 sample size calculations from the Welfare Quality Protocol; and (4) selecting the first, middle, or final third of cows exiting the milking parlor. Estimates were compared with true values using regression analysis and were considered accurate if they met 3 criteria: the coefficient of determination was ≥0.9 and the slope and intercept did not differ significantly from 1 and 0, respectively. All estimates met the slope and intercept criteria, whereas the coefficient of determination increased when more cows were sampled. All estimates were accurate for neck alterations, ocular discharge (22.2 ± 27.4%), and carpal joint hair loss (14.1 ± 17.4%). Selecting a third of the milking order or using the Welfare Quality sample size calculations failed to accurately estimate all measures simultaneously. However, all estimates were accurate when selecting at least 2 of every 3 cows locked at the feed bunk. Using restraint position at the feed bunk did not differ systematically from computer-selecting the same proportion of cows randomly, and the former may be a simpler approach for welfare assessments.
Key words: health, on-farm assessment, sampling, validation

INTRODUCTION

Welfare describes an animal's state and spans a spectrum from good to poor. Based on this concept, assessments of farm animal welfare increasingly focus on evaluating animal-based measures (e.g., Whay et al., 2003) in addition to aspects of the environment or management. Animal welfare assurance programs not only seek to accurately classify the state of the animals on a given farm, but are also faced with feasibility constraints, namely investment of time and the number of assessors.

A key issue is the appropriate number of animals to sample on a given farm, as this can affect both the accuracy of welfare classification and the costs associated with conducting the assessment. Studies on dairy cattle (Endres et al., 2014) and swine (Mullan et al., 2009) have shown some associations between the true prevalence of an animal-based measure and the sample size needed to accurately estimate it, such that conditions with low prevalence sometimes require sampling

Received July 31, 2017.
Accepted October 17, 2017.
[1]Corresponding author: cbtucker@ucdavis.edu

more animals. This means that accurately estimating low-prevalence conditions may present a feasibility challenge on larger operations, particularly given the continued global trend of farms increasing in size (reviewed by Robbins et al., 2016).

Assessment schemes have attempted to balance accuracy and feasibility in their guidelines for the number of animals to score. Among the Welfare Quality protocols, these sample size recommendations vary. Within the swine protocol (Welfare Quality, 2009b), certain measures call for different sample sizes, but the number of pigs sampled does not increase directly with farm size. For dairy cattle (Welfare Quality, 2009a), a single sample size, based on Cochran (1977), is used for all animal-based health measures on a given farm, and the number of cows sampled grows at a diminishing rate as herd size increases. To our knowledge, no work has directly evaluated the reliability of using these sample size calculations to estimate prevalence across a large selection of the animal-based measures in the Welfare Quality (2009a) protocol for dairy cattle.

A few studies have focused primarily on lameness, an important animal-based measure. Strategies have included comparing sample estimates against a tolerance range relative to the true prevalence (Main et al., 2010; Hoffman et al., 2013) or using sequential sampling based on a technique used in clinical trials (Heath et al., 2016), although the latter study assumed that the sample size calculations in the Welfare Quality (2009a) protocol represented the true prevalence. One study, however, used a more stringent linear regression approach to evaluate the sample size needed to estimate lameness and other select animal-based measures (Endres et al., 2014). Estimates were deemed accurate only if their relationship to the true prevalence had a coefficient of determination ($\mathbf{R^2}$) $\geq$0.9 and the slope and intercept did not differ from 1 and 0, respectively. Our research group has used the same linear regression criteria to identify schedules for sampling cattle behavior through direct observation (Stackhouse-Lawson et al., 2015; Chen et al., 2016; Tresoldi et al., 2016) or using data loggers (Ledgerwood et al., 2010).

In addition to determining how many cows to sample, questions remain about which animals to select. The Welfare Quality (2009a) protocol for dairy cattle provides recommendations such as selecting every *n*th cow either in the milking order or when locked at the feed bunk, or by choosing cattle distributed throughout the home pen and engaged in various activities (e.g., standing, lying, or feeding) to approximate random sampling. In studies that benchmark measures of dairy cattle welfare across farms, a common approach is to score cows during milking (von Keyserlingk et al., 2012), but a single assessor would be unable to observe the same cow from the front, back, and while walking to score all of the animal-based measures in the Welfare Quality (2009a) assessment.

Our objective was to evaluate how prevalence estimates for a wide range of animal-based measures are affected by the proportion of high-producing cows sampled and the method used to select them. We focused on this population because considerable interest exists in studying confined, high-producing cattle (e.g., Espejo et al., 2006; von Keyserlingk et al., 2012; Cook et al., 2016), which are particularly at risk for welfare concerns such as lameness (Barkema et al., 1994). We evaluated the accuracy of the estimates generated by sampling various proportions of cattle. In addition, we compared the strategies of choosing cows based on their position while restrained at the feed bunk, selecting ear tag numbers randomly, or assessing different cows at separate times during milking.

MATERIALS AND METHODS

Animals, Housing, and Measures

Between October and December 2013, 2 assessors visited 10 California dairy farms (n = 10 farms, 1,805 cows total), all of which milked cows twice daily. These farms were a convenience sample among clients of the Veterinary Medicine Teaching and Research Center of the University of California-Davis (UC Davis) School of Veterinary Medicine. On each farm, all cows in 1 pen of high-producing cows (DIM <100) were assessed. The housing type, breed managed, herd size, and size of the high-producing pen from each farm are shown in Table 1. Each cow was identified by ear tag number and was evaluated using select measures (Table 2) from the dairy cow section of the Welfare Quality (2009a) protocol. The measures selected were all nonbehavioral ones assessed at the individual animal level. The evaluation was conducted while cows were restrained in head locks in the pen with the exception of lameness, which was scored when cows were released. One assessor recorded ear tag numbers and nasal and ocular discharge from the front of the cow. The second assessor recorded all other measures from inside of the pen, releasing 1 cow at a time from the head locks to score lameness and to gain a full side view of the next cow. Intraobserver reliability was determined using a combination of photos, video clips, and live observations ($\geq$20 cows per measure), and percentage agreement was $\geq$80% for all measures. Cow position along the feed bunk during restraint and the order in which they were milked were also manually recorded.