ABSTRACT

Our objective was to evaluate how sampling strategies (i.e., how many cows to sample and which animals to include) used in 4 dairy cattle welfare assessment programs affect the classification of dairy farms relative to thresholds of acceptability for animal-based measures. We predicted that classification performance would improve when more cows were sampled and when selecting from all lactating cows versus when some pens were excluded. On 38 freestall farms, we assessed all 12,375 cows for lameness, injuries on the tarsal (hock) and carpal joints, and body condition score and calculated the farm-level prevalence for each measure. Based on approaches used in the industry, we evaluated 6 sampling strategies generated using formulas with precision ($d$) of 15, 10, or 5% applied to either a single high-producing pen or all lactating cows; an additional sample was included with $d = 10\%$ applied to the entire herd, selecting lactating cows in proportion to their representation in the herd. For each sampling strategy, cow records were selected randomly (in 10,000 replicates) to calculate prevalence. The strategy of assessing all cows in the high-producing pen was also compared. Farms were classified as meeting (below) or failing to meet (above) thresholds of ≤15% moderate lameness; ≤20% moderate carpal or hock injuries; <10, <5, and ≤1% severe lameness; or injuries on the carpus or hock; and <5, <3, <1, or 0% thin cows. For each measure and threshold, we calculated pooled percent agreement, kappa, sensitivity, specificity, and positive and negative predictive value for each sampling strategy using true prevalence as the gold standard for herd classification. Across measures and thresholds, classification performance increased with the number of cows sampled [i.e., when narrower precision values ($d = 5$ vs. $10$ vs. $15\%$) were used in the sample size calculation]. Because narrower precision values can dramatically increase sample size, assessment programs may need to consider both feasibility and the degree of misclassification they will accept. Applying the formula directly to lactating cows performed better than applying it to the entire herd and then selecting lactating cows in proportion to their representation in the herd. Farm classifications were similar whether cows in the hospital pen were included or excluded from the sample. Selecting all cows from the high-producing pen resulted in classifications similar to when including all lactating cows, suggesting that assessing cows from the high-producing pen may serve as an acceptable proxy for all lactating cows on the farm.

Key words: animal welfare, animal-based measure, skin lesion, certification program

INTRODUCTION

On-farm assessments often incorporate direct observations of animals (Whay et al., 2003) in addition to examining records and capturing information about management and the environment. For dairy cattle, common approaches for collecting animal-based measures include gait scoring to categorize lameness, identifying the presence of lesions on the legs or other body parts, and scoring body condition to identify thin cows. Some assessment schemes and labeling programs establish thresholds of acceptability for individual measures (e.g., American Humane Certified in the United States, American Humane Association, 2013; Farmers Assuring Responsible Management in the United States, National Milk Producers Federation, 2016; proAction in Canada, Dairy Farmers of Canada, 2017; Dairy Well in the United States, Walker et al., 2017). Other programs use these measures to calculate a composite animal welfare rating (i.e., Welfare Quality in the Euro-
ean Union, Welfare Quality, 2009) or observe animals for monitoring or reporting purposes without setting minimum standards (e.g., Red Tractor in the United Kingdom, Assured Food Standards, 2017).

For programs that conduct direct animal-based observations, an important issue is determining the appropriate number of animals to assess. In addition to affecting how accurately an assessment reflects the status of the animals on a farm, sample size affects the time and labor required to collect measures. Among animal welfare assessment programs, common sampling strategies range from selecting a standard number of animals (i.e., 20 cows/farm for Red Tractor, Assured Food Standards, 2017) to evaluating all animals in the groups of interest (i.e., American Humane Certified, for select measures, American Humane Association, 2013). An intermediate strategy, given the challenge of balancing accuracy and feasibility when assessing large herds, is to use a sampling calculation to select more animals as herd size increases, but at a diminishing rate. Programs that use this strategy differ in the formulas they use and the subgroups of cattle to which they apply the calculations (i.e., the entire herd, lactating cows, or only high-producing cows), resulting in a wide variation in sampling strategies.

Few studies have evaluated sampling for assessing animal-based measures. For dairy cattle, approaches have included evaluating whether sample estimates fall within a tolerance range relative to the true herd prevalence of lameness (Main et al., 2010; Hoffman et al., 2013) and using linear regression to evaluate how well samples reflect the true prevalence of a wider range of animal-based measures (Endres et al., 2014; Van Os et al., 2018). However, assessment programs are typically less concerned with precise prevalence and more concerned with whether a specific farm meets a predetermined threshold of acceptability. To determine whether additional samples were needed to classify farms accurately relative to thresholds, Heath et al. (2016) tested a multi-step decision-tree approach, similar to techniques used in clinical trials. Rather than determining the true herd prevalence as the gold standard to classify farms, the authors used a sampling calculation from the Welfare Quality (2009) protocol, which is sometimes inaccurate (Van Os et al., 2018). Another approach, applied to pig farms by Mullan et al. (2009), is to generate multiple (i.e., 10,000) sample repetitions per farm and report the resulting proportion of misclassifications relative to thresholds for each animal-based measure. To our knowledge, this approach has not been applied to dairy farms.

In addition to determining how many cattle to sample, another relevant question is which animals to include. In particular, programs vary in whether they consider animals in “hospital” pens; some programs sample from these pens (i.e., FARM, National Milk Producers Federation, 2016), but others exclude them based on the rationale that those cows may already be receiving special care (i.e., proAction, Dairy Farmers of Canada, 2017). Some other programs take intermediate approaches, such as assessing the hospital pen but excluding those cows from the thresholds (i.e., Dairy Well, Walker et al., 2017) or monitoring the pen for only resource-based measures (i.e., Validus, Urbandale, IA, personal communication). Including cows from only certain subgroups has been shown to result in different prevalence estimates compared with selecting from all lactating cows (Endres et al., 2014). To our knowledge, no studies have evaluated the effects of excluding hospital pens on farm classification.

The aim of the current study was to evaluate how different sampling methods, based on those most frequently used in North American dairy cattle welfare certification programs, affect farm classification. The primary objective was to evaluate different sample size calculations; we predicted that accuracy would increase when larger numbers of animals were selected. The second objective was to evaluate how well an assessment of the high-producing pen related to an assessment including all lactating cows; we predicted that classifications would be more accurate when samples were drawn from all lactating cows. The third objective was to evaluate effects of excluding the hospital pen; we again predicted that classifications would be more accurate when samples were drawn from all lactating cows compared with when the hospital pen was excluded.

**MATERIALS AND METHODS**

All lactating cows on 38 freestall dairy operations in the central region of the state of Paraná, Brazil, were assessed between March and October 2016. These farms were part of a convenience sample identified with the help of industry partners working with the C intelona Farmers’ Cooperative (Castro, PR, Brazil). Only freestall operations were included to reduce variability attributable to facility type. All procedures were approved by the Ethics Committees on Research on Humans (protocol #PP1237779, 2015) and Animals (protocol #PP00949, 2014) of the Universidade Federal de Santa Catarina and by the UBC Animal Care Committee (protocol #A15–0082). Details of the on-farm data collection procedures are described in Costa et al. (2018). In brief, as cows exited the milking parlor, their individual identification numbers were recorded and they were evaluated for lameness (5-point integer scale; Flower and Weary, 2006; cows scoring 3 and ≥4 were
considered moderately and severely lame, respectively),
injuries on the tarsal (hock) and carpal joints (3-point
integer scale modified from Cornell Cooperative Exten-
sion, https://ecommons.cornell.edu/bitstream/handle/
1813/36913/hockscore.pdf; scores 2 and 3 were con-
sidered moderate and severe, respectively), and BCS (1 to
5 scale in 0.25 increments; Edmonson et al., 1989; cows
scoring ≤ 2.0 were considered thin). A single, trained
assessor scored all of the lactating cows on all farms in
the study.

Eight cows with missing values were excluded, result-
ing in a total of 12,375 animals in the data set (mean
± SD: 326 ± 199 cows/farm, range 71 to 901; Table 1;
full data set available at http://dx.doi.org/10.17632/
khwgk5t2zg.2). In addition, we collected farmer-reported
densities of dry cows, pre-weaned heifer calves, and
weaned heifers <1 yr old. The number in the latter 2
groups was doubled to estimate the number of bred
heifers. These categories were summed with the num-
ber of lactating cows to generate an estimate of total
herd size (mean ± SD: 836 ± 470 cattle/farm, range
169 to 2,101; Supplemental Table S1; https://doi.org/

To provide context for the thresholds assessed and the
sampling strategies evaluated, we examined a range of
dairy industry assessment programs that established
thresholds of acceptability for lameness, leg injuries,
and BCS (Supplemental Table S2; https://doi.org/
10.3168/jds.2018-15134) and that selected more cows
as group size increased, but at a diminishing rate
(excerpts of the programs’ minimum sampling recom-
mandations are given in Supplemental Table S3; https:
//doi.org/10.3168/jds.2018-15134). The 4 assessments
that met these criteria were proAction (Dairy Farmers
of Canada, 2017), the FARM Program (National Milk
Producers Federation, 2016 and the National Dairy
FARM Program, Arlington, VA, personal communica-
tion), Dairy Well (Walker et al., 2017), and Validus
(V validus, Urbandale, IA, personal communication).

Sample Size Calculations and Cow Selection

All 4 assessment programs used sample size calcula-
tions derived from Cochran (1977). First, the sample
size for an infinite population \((n_{inf})\) was determined
using the following formula, where \(p = \) the expected
prevalence of the condition (expressed as a proportion,
i.e., 0.50 for 50%, the value that maximizes the result-
ning sample size), \(Z = \) the \(Z\)-value corresponding to
the confidence interval (i.e., 1.96 for 95% CI), and \(d = \)
the desired precision (expressed as a proportion in the
equation, but described throughout as a percentage for
clarity):

\[
\begin{align*}
n_{inf} &= \frac{p(1-p)Z^2}{d^2}. \\
n_{fin} &= \frac{n_{inf}}{1 + \left(\frac{n_{inf} - 1}{N}\right)}.
\end{align*}
\]

Second, the minimum sample size for a finite popula-
tion \((n_{fin})\) was determined using the number of animals
in the population of interest on each farm \((N)\):

All 4 assessment programs use \(P = 0.50\) and \(Z = 1.96,\)
but vary the value for \(d\), using either 15 (proAction,
Dairy Farmers of Canada, 2017), 10 (FARM, National
Milk Producers Federation, 2016; Validus, Urbandale,
IA, personal communication), or 5% (Dairy Well, Walk-
er et al., 2017). Three programs apply the formula for a
finite population directly to the population of lactating
(proAction, Dairy Farmers of Canada, 2017), lactat-
ing or lactating and close-up dry (for lameness and
BCS vs. leg injuries, respectively, Validus, Urbandale,
IA, personal communication), or high-producing cows
(Walker et al., 2017). The FARM program (National
Milk Producers Federation, 2016) likewise focuses on
lactating and dry cows for lameness and leg injuries.
However, the FARM program scores body condition
across all age groups, and thus they have chosen to
apply the sampling formula for a finite population once
to the total herd size, with lactating cows then selected
in proportion to their representation on the farm (Na-
tional Milk Producers Federation, 2016).

To test objective 1, we evaluated 7 sampling strate-
gies (Table 1) based on those used in the existing pro-
grams. For 6 of these, we applied the formula for an
infinite population using \(d = 15, 10,\) and 5% followed
by the formula for a finite population using either a
single pen containing the bulk of the highest producing
and oldest lactating cows (high-pen strategy; mean ±
SD: 107 ± 74 cows/farm, range 30 to 337; Table 1) or all
lactating cows (lactating-cow strategy). For the
seventh strategy, we mimicked the FARM (National
Milk Producers Federation, 2016) method by applying
the formula for an infinite population using \(d = 10\%
followed by the formula for a finite population using
the estimated total herd size, and then selecting lactat-
ing cows in proportion to their representation in the
herd (proportion-of-herd strategy). For comple-
teness, we repeated the latter strategy using \(d = 5\) and 15%.

We selected cows randomly using a simple ran-
dom sampling technique (PROC SURVEYSELECT,
METHOD = SRS) in SAS (version 9.4, SAS Institute
Inc., Cary, NC). Aside from the high-pen strategy, cows
were chosen proportionally from all pens within each
farm (STRATA = farm and pen), including the hospital pen. For each sampling strategy, the selection of cows was replicated 10,000 times, without repeating cows within a replicate (all code available at http://dx.doi.org/10.17632/khwgk5t2zg.2). We did this to simulate a single assessor taking different same-day snapshots of the same farm, selecting a new sample each time, to reduce the possibility of drawing erroneous conclusions based on a single sample.

**Farm Classification**

Based on the standards from the 4 assessment schemes, we used threshold levels of ≤15% for moderate lameness (score 3); <10%, <5%, and ≤1% for severe lameness (scores ≥ 4), hock, and carpal injuries (score 3); and <5, <3, <1, and 0% for thin cows (scores ≤ 2.0). None of the assessments currently has standards for moderate hock and carpal injuries (score

<table>
<thead>
<tr>
<th>Lactating cows (no.)</th>
<th>High-producing pen(^1)</th>
<th>Hospital pens(^2) (no. of cows)</th>
<th>Sampling calculation applied to</th>
<th>Lactating cows(^3)</th>
<th>High-producing pen(^4)</th>
<th>Total herd(^5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>(d^6 = 15)</td>
<td>10</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>71</td>
<td>36</td>
<td>—</td>
<td></td>
<td>27</td>
<td>41</td>
<td>60</td>
</tr>
<tr>
<td>109</td>
<td>60</td>
<td>—</td>
<td></td>
<td>31</td>
<td>51</td>
<td>85</td>
</tr>
<tr>
<td>120</td>
<td>60</td>
<td>—</td>
<td></td>
<td>32</td>
<td>54</td>
<td>92</td>
</tr>
<tr>
<td>122</td>
<td>54</td>
<td>Hospital + fresh (9)</td>
<td></td>
<td>32 (31)</td>
<td>54 (52)</td>
<td>93 (88)</td>
</tr>
<tr>
<td>138</td>
<td>40</td>
<td>—</td>
<td></td>
<td>33</td>
<td>57</td>
<td>102</td>
</tr>
<tr>
<td>147</td>
<td>39</td>
<td>Lame (32)</td>
<td></td>
<td>33 (31)</td>
<td>58 (53)</td>
<td>107 (89)</td>
</tr>
<tr>
<td>153</td>
<td>153</td>
<td>—</td>
<td></td>
<td>34</td>
<td>59</td>
<td>110</td>
</tr>
<tr>
<td>154</td>
<td>65</td>
<td>—</td>
<td></td>
<td>34</td>
<td>59</td>
<td>110</td>
</tr>
<tr>
<td>161</td>
<td>79</td>
<td>Hospital + high SCC (22)</td>
<td></td>
<td>34 (33)</td>
<td>60 (57)</td>
<td>114 (102)</td>
</tr>
<tr>
<td>161</td>
<td>31</td>
<td>—</td>
<td></td>
<td>34</td>
<td>60</td>
<td>114</td>
</tr>
<tr>
<td>172</td>
<td>85</td>
<td>Hospital + high SCC (5)</td>
<td></td>
<td>34 (34)</td>
<td>62 (61)</td>
<td>119 (117)</td>
</tr>
<tr>
<td>180</td>
<td>70</td>
<td>—</td>
<td></td>
<td>35</td>
<td>63</td>
<td>123</td>
</tr>
<tr>
<td>185</td>
<td>88</td>
<td>Lame + fresh (9)</td>
<td></td>
<td>35 (35)</td>
<td>63 (62)</td>
<td>125 (121)</td>
</tr>
<tr>
<td>189</td>
<td>70</td>
<td>—</td>
<td></td>
<td>35</td>
<td>64</td>
<td>127</td>
</tr>
<tr>
<td>216</td>
<td>80</td>
<td>—</td>
<td></td>
<td>36</td>
<td>67</td>
<td>139</td>
</tr>
<tr>
<td>222</td>
<td>60</td>
<td>Hospital + fresh (20)</td>
<td></td>
<td>36 (35)</td>
<td>67 (65)</td>
<td>141 (133)</td>
</tr>
<tr>
<td>254</td>
<td>105</td>
<td>Hospital + fresh (16)</td>
<td></td>
<td>37 (36)</td>
<td>70 (69)</td>
<td>153 (147)</td>
</tr>
<tr>
<td>294</td>
<td>83</td>
<td>Hospital (11)</td>
<td></td>
<td>37 (37)</td>
<td>73 (72)</td>
<td>167 (163)</td>
</tr>
<tr>
<td>300</td>
<td>81</td>
<td>—</td>
<td></td>
<td>38</td>
<td>73</td>
<td>169</td>
</tr>
<tr>
<td>324</td>
<td>99</td>
<td>—</td>
<td></td>
<td>38</td>
<td>74</td>
<td>176</td>
</tr>
<tr>
<td>337</td>
<td>47</td>
<td>Hospital + fresh (39)</td>
<td></td>
<td>38 (37)</td>
<td>74 (72)</td>
<td>177 (165)</td>
</tr>
<tr>
<td>339</td>
<td>89</td>
<td>Hospital + fresh (43)</td>
<td></td>
<td>38 (37)</td>
<td>75 (73)</td>
<td>180 (167)</td>
</tr>
<tr>
<td>348</td>
<td>64</td>
<td>—</td>
<td></td>
<td>38</td>
<td>75</td>
<td>183</td>
</tr>
<tr>
<td>348</td>
<td>86</td>
<td>Hospital (17)</td>
<td></td>
<td>38 (38)</td>
<td>75 (75)</td>
<td>183 (178)</td>
</tr>
<tr>
<td>363</td>
<td>87</td>
<td>Lame (46)</td>
<td></td>
<td>38 (38)</td>
<td>76 (74)</td>
<td>187 (174)</td>
</tr>
<tr>
<td>370</td>
<td>126</td>
<td>Hospital (18)</td>
<td></td>
<td>38 (38)</td>
<td>76 (76)</td>
<td>189 (184)</td>
</tr>
<tr>
<td>381</td>
<td>196</td>
<td>—</td>
<td></td>
<td>39</td>
<td>77</td>
<td>192</td>
</tr>
<tr>
<td>421</td>
<td>193</td>
<td>—</td>
<td></td>
<td>39</td>
<td>78</td>
<td>201</td>
</tr>
<tr>
<td>422</td>
<td>143</td>
<td>Hospital (20)</td>
<td></td>
<td>39 (39)</td>
<td>78 (78)</td>
<td>201 (197)</td>
</tr>
<tr>
<td>432</td>
<td>134</td>
<td>Hospital + fresh (17)</td>
<td></td>
<td>39 (39)</td>
<td>79 (78)</td>
<td>204 (200)</td>
</tr>
<tr>
<td>432</td>
<td>66</td>
<td>—</td>
<td></td>
<td>39</td>
<td>79</td>
<td>204</td>
</tr>
<tr>
<td>516</td>
<td>158</td>
<td>Hospital (12), lame (57)</td>
<td></td>
<td>40 (39)</td>
<td>81 (79)</td>
<td>221 (207)</td>
</tr>
<tr>
<td>598</td>
<td>180</td>
<td>—</td>
<td></td>
<td>40</td>
<td>83</td>
<td>234</td>
</tr>
<tr>
<td>680</td>
<td>337</td>
<td>Hospital (2)</td>
<td></td>
<td>40 (40)</td>
<td>84 (84)</td>
<td>246 (243)</td>
</tr>
<tr>
<td>690</td>
<td>30</td>
<td>Hospital (105), lame (33)</td>
<td></td>
<td>40 (40)</td>
<td>84 (82)</td>
<td>247 (227)</td>
</tr>
<tr>
<td>801</td>
<td>272</td>
<td>Hospital + low (26)</td>
<td></td>
<td>41 (41)</td>
<td>86 (86)</td>
<td>260 (257)</td>
</tr>
<tr>
<td>901</td>
<td>319</td>
<td>Hospital (16)</td>
<td></td>
<td>41 (41)</td>
<td>87 (87)</td>
<td>270 (268)</td>
</tr>
</tbody>
</table>

\(^1\)The pen containing the bulk of the highest producing and oldest lactating cows in the herd.

\(^2\)The farmers' designations for their hospital pens on farms with this pen type.

\(^3\)The sampling formula was applied to the number of lactating cows in the herd. Values in parentheses indicate the number of cows sampled when cows in hospital pens were excluded from the total count of lactating cows in herds with this pen type.

\(^4\)The sampling formula was applied to a single pen containing the bulk of the highest producing and oldest lactating cows in the herd.

\(^5\)The sampling formula was applied to the total herd size, with lactating cows selected in proportion to their representation in the herd.

\(^6\)The desired precision \((d = 15, 10, or 5\%)\) used in the sampling formula.
Among recent studies on freestall operations, some reported average prevalence of hair loss on the hock and carpus between 10 to 15% (Zaffino Heyerhoff et al., 2014; Adams et al., 2017), but others report much higher values (38 to 76%; von Keyserlingk et al., 2012; Endres et al., 2014; Cook et al., 2016). Given this variation, we used a threshold of ≤20% for moderate hock and carpal injuries.

For each measure, true (Table 2; Supplemental Figure S4, https://doi.org/10.3168/jds.2018-15134) and estimated prevalence were calculated as a percentage of all lactating cows or of each sampling replicate, respectively. For consistency across all 38 farms, we included cows from the hospital pen on farms that had this pen type because the farms without a hospital pen would not have the option of removing special-needs cows from their evaluations. Farms were classified as meeting or failing to meet the acceptability thresholds. Within each sampling replicate, a $2 \times 2$ matrix was constructed for the number of true negatives (TN; farms passing based on both true and estimated prevalence), true positives (TP; farms failing to meet standards based on both true and estimated prevalence), false positives (FP; farms that should meet standards based on true prevalence but failed based on the sample estimate), and false negatives (FN; farms that should fail based on true prevalence but passed based on the sample estimate).

The number of TN, TP, FP, and FN were pooled among the 10,000 replicates (similar to meta-analysis of diagnostic accuracy using a consistent test method; Riley et al., 2015) to calculate 6 classification metrics. Accuracy (percentage agreement) and Cohen’s kappa coefficient of concordance (which accounts for agreements by chance and is thus more conservative than percentage agreement; Hollenbeck, 1978; Martin and Bateson, 2007) are reported in the results. The level of agreement for kappa values are classified based on Landis and Koch (1977): $\kappa < 0 =$ poor, $\kappa \geq 0$ and $\leq 0.2 =$ slight, $\kappa > 0.2$ and $\leq 0.4 =$ fair, $\kappa > 0.4$ and $\leq 0.6 =$ moderate, $\kappa > 0.6$ and $\leq 0.8 =$ substantial, and $\kappa > 0.8 =$ almost perfect agreement. As well, sensitivity ($TP/(TP + FN)$, the proportion of farms that should fail classified as failing by the sample estimates], specificity [$TN/(TN + FP)$, the proportion of farms that should pass classified as passing by the sample estimates], positive predictive value [$PPV; TP/(TP + FP)$, the proportion of farms classified as failing by the sample estimates that should according to true prevalence], and negative predictive value [$NPV; TN/(TN + FN)$, the proportion of farms classified as passing by the sample estimates that should according to true prevalence] are shown in Supplemental Materials.

### High Pen

The Dairy Well assessment program (Walker et al., 2017) trains its evaluators to assess the entire high-producing pen of cows (“ideal”) or to apply the sample size calculation if the former strategy is infeasible. To evaluate how well the strategy of assessing all cows in the high pen reflects the larger population of lactating cows, we compared prevalence within the high pen to that of all lactating cows on each farm. For each mea-

<table>
<thead>
<tr>
<th>Measure</th>
<th>Score threshold</th>
<th>All lactating cows on all 38 farms</th>
<th>High-producing pen on all 38 farms</th>
<th>All lactating cows on 20 farms with hospital pens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lameness$^1$</td>
<td>Moderate 3</td>
<td>21.0 ± 4.8</td>
<td>9.2–31.4</td>
<td>24.9 ± 7.8</td>
</tr>
<tr>
<td></td>
<td>Severe ≥4</td>
<td>23.1 ± 8.6</td>
<td>2.8–39.5</td>
<td>25.9 ± 11.1</td>
</tr>
<tr>
<td>Carpal injury$^2$</td>
<td>Moderate 2</td>
<td>37.1 ± 13.8</td>
<td>8.1–69.3</td>
<td>40.3 ± 15.4</td>
</tr>
<tr>
<td></td>
<td>Severe 3</td>
<td>8.7 ± 7.3</td>
<td>0.5–32.3</td>
<td>10.1 ± 9.7</td>
</tr>
<tr>
<td>Hock injury$^2$</td>
<td>Moderate 2</td>
<td>30.6 ± 15.9</td>
<td>0–55.1</td>
<td>30.0 ± 17.8</td>
</tr>
<tr>
<td></td>
<td>Severe 3</td>
<td>8.1 ± 6.7</td>
<td>0–26.0</td>
<td>8.3 ± 9.1</td>
</tr>
<tr>
<td>Thin$^3$</td>
<td>≤2.0</td>
<td>0.7 ± 2.1</td>
<td>0–12.7</td>
<td>0.5 ± 1.5</td>
</tr>
</tbody>
</table>

$^1$Locomotion was scored on a 1 to 5 integer scale (Flower and Weary, 2006), with scores 1 and 2 = sound, 3 = moderately lame, and 4 and 5 = severely lame.

$^2$One carpal joint and one hock joint were each scored for injuries using a 1 to 3 integer scale (modified from Cornell Cooperative Extension, https://ecommons.cornell.edu/bitstream/handle/1813/36913/hockscore.pdf), with score 1 = healthy, 2 = moderate injury, and 3 = severe injury.

$^3$Body condition was scored on a 1 to 5 scale in 0.25 increments (Edmonson et al., 1989). Scores ≤2.0 were considered thin; none of the cows in our sample had scores of 1.0.
sure, prevalence was calculated as a percentage of either the high pen or of all lactating cows. We examined the relationship between these prevalence calculations using linear regressions (PROC REG) and evaluated the $R^2$ value and whether the intercept and slope differed ($P < 0.05$) from 0 and 1, respectively. In addition, for each measure and threshold, farm classifications were compared for the high pen versus all lactating cows (objective 2) by calculating accuracy and kappa.

Effect of Excluding Hospital Pens

To evaluate the effects of excluding the hospital pen (objective 3), we used the subset of 20 farms in our data set with farmer-designated hospital pens (7,746 lactating cows total; 7,153 excluding the hospital pens). We compared farm classifications when cows from the hospital pen were either included or excluded from the sample, with sample size calculated using $d = 15, 10,$ and 5%, selecting lactating cows in 10,000 sample replicates with the same methods as in objective 1.

RESULTS

Objectives 1 and 2: All Farms

Lameness and Leg Injuries. All sampling strategies resulted in relatively high accuracy (79.0 to 99.9%) across all thresholds for lameness (Figure 1; Supplemental Table S5; https://doi.org/10.3168/jds.2018-15134) and carpal (Figure 2) or hock injuries (Figure 3). However, when accounting for chance agreements, kappa was more variable (range: 0.30 to 0.93, fair to almost perfect agreement, except $\kappa = 0$ with all sampling strategies for severe lameness using a threshold of $\leq 1\%$). Sensitivity and PPV were likewise generally high across all sampling strategies and thresholds (ranges: 77.4 to 100% and 72.4 to 100%, respectively; Supplemental Figures and Tables S6–S13; https://doi.org/10.3168/jds.2018-15134). However, specificity was more variable (range: 43.3 to 100%), as was NPV (range: 21.3 to 88.1%, except NPV = 0 with all sampling strategies for severe lameness using a threshold of $\leq 1\%$).

![Figure 1. Pooled accuracy (circles) and Cohen’s kappa (triangles) for farm classifications when using thresholds of (A) $\leq 15\%$ for moderately or (B) $< 10\%$, (C) $< 5\%$, and (D) $\leq 1\%$ for severely lame cows. Classifications were determined by comparing the true prevalence of moderate and severe lameness among all lactating cows on each farm against estimates (10,000 sample repetitions each) generated by sampling. Six sample sizes were calculated using $d = 15, 10,$ and 5%, selecting lactating cows in 10,000 sample replicates with the same methods as in objective 1. The filled symbols indicate when the formula using $d = 10\%$ was applied to the total herd size, with lactating cows selected in proportion to their representation in the herd. The dashed lines indicate levels of agreement based on kappa: $\kappa \leq 0.2 = \text{slight}$, $\kappa > 0.2$ and $\leq 0.4 = \text{fair}$, $\kappa > 0.4$ and $\leq 0.6 = \text{moderate}$, $\kappa > 0.6$ and $\leq 0.8 = \text{substantial}$, and $\kappa > 0.8 = \text{almost perfect}$.]
**Thin Cows.** Out of the 38 farms in our data set, only 14 farms (of which 8 had hospital pens) had any thin cows (Supplemental Figure S4G; https://doi.org/10.3168/jds.2018-15134). Because there were no thin cows on the remaining 24 farms, they were classified correctly for all thresholds regardless of the sampling method. The results for thin cows are shown in Supplemental Tables and Figures S5, S9, and S13–S16 (https://doi.org/10.3168/jds.2018-15134), but are not discussed further.

**Desired Precision.** As predicted, using a narrower desired precision ($d = 5$ vs. $10$ vs. $15\%$) generally increased the accuracy (percentage agreement) of farm classifications. This pattern was observed whether sampling from among all lactating cows or only those in the high-producing pen and was apparent across almost all thresholds and measures of lameness (Figure 1) and leg injuries (Figures 2 and 3); the exceptions to this pattern differed by only 0.2 to 0.4\% (Figures 1C, 2C, 3B). Likewise, using narrower desired precision almost always resulted in higher kappa, sensitivity, and specificity (Supplemental Figures and Tables S5–S9, https://doi.org/10.3168/jds.2018-15134), and PPV and NPV (Supplemental Figures and Table S10–S13, https://doi.org/10.3168/jds.2018-15134).

**High-Producing Pen.** When comparing true prevalence among cows in the high-producing pen versus among all lactating cows, the association was strong for moderate hock injuries ($R^2 = 0.86$; Figure 4) as well as for severe carpal injuries and thin cows ($R^2 = 0.86$ and 0.97, respectively), although the slope differed from 1 for the latter 2 measures ($P \leq 0.006$); the slope and intercept did not differ from 0 and 1, respectively, for any of the other measures ($P \geq 0.16$; all relationships shown in Supplemental Figure S4, https://doi.org/10.3168/jds.2018-15134). The association between prevalence among high-producing and all lactating cows was also relatively strong for moderate carpal and severe hock injuries ($R^2 = 0.76$ and 0.69, respectively), but less so for moderate and severe lameness ($R^2 = 0.43$ and 0.47, respectively). Nonetheless, for the purposes of farm classification relative to thresholds, sampling

![Figure 2](image-url) **Figure 2.** Pooled accuracy (circles) and Cohen’s kappa (triangles) for farm classifications when using thresholds of (A) $\leq 20\%$ for moderate or (B) $<10\%$, (C) $<5\%$, and (D) $\leq 1\%$ for severe injuries on the carpal joint. Classifications were determined by comparing the true prevalence of moderate and severe carpal injuries among all lactating cows on each farm against estimates (10,000 sample repetitions each) generated by sampling. Six sample sizes were calculated using a formula with precision ($d = 15$, 10, or 5\%) applied to either all lactating cows or a single pen of high-producing cows. The filled symbols indicate when the formula using $d = 10\%$ was applied to the total herd size, with lactating cows selected in proportion to their representation in the herd. The dashed lines indicate levels of agreement based on kappa: $\kappa \leq 0.2 =$ slight, $\kappa > 0.2$ and $\leq 0.4 =$ fair, $\kappa > 0.4$ and $\leq 0.6 =$ moderate, $\kappa > 0.6$ and $\leq 0.8 =$ substantial, and $\kappa > 0.8 =$ almost perfect.
from the high-producing pen (objective 1) or using this entire pen (objective 2) was comparable to the results for all lactating cows.

When the sample size formulas were applied to either cows in the high-producing pen or to all lactating cows (objective 1), accuracy and kappa were often similar (e.g., for all values of \( d \) for severe carpal injuries with a threshold of <10%, Figure 2B; Supplemental Table S5; https://doi.org/10.3168/jds.2018-15134), but in many cases the former strategy was less accurate, particularly when using \( d = 5\% \) (e.g., moderate lameness, Figure 1A). In most cases, sensitivity was similar between the strategies within a given level of \( d \), but was lower when sampling from the high-producing pen for hock injuries (Supplemental Figure S8 and Table S9; https://doi.org/10.3168/jds.2018-15134). In contrast, specificity, PPV, and NPV were in most cases lower when sampling from the high-producing pen compared with all lactating cows, although the pattern was reversed or inconsistent for hock injuries with specificity and PPV (Supplemental Figures and Tables S8–S9 and S12–S13, https://doi.org/10.3168/jds.2018-15134), and for carpal injuries with NPV (Supplemental Figure S11 and Table S13, https://doi.org/10.3168/jds.2018-15134).

When farm classifications were compared using either the entire high-producing pen or all lactating cows (objective 2), percent agreement was high (range: 78.9 to 100%) for all measures and thresholds (Supplemental Table S17, https://doi.org/10.3168/jds.2018-15134; individual farms shown in Supplemental Figure S4, https://doi.org/10.3168/jds.2018-15134). Likewise, the level of agreement accounting for chance ranged from substantial to perfect (\( \kappa = 0.64 \) to 1.0) for all measures except moderate lameness (\( \kappa = 0.37 \), fair) and severe lameness or carpal injuries with thresholds of \( \leq 1\% \) and \( <5\% \), respectively (\( \kappa = 0 \) and 0.56, slight and moderate, respectively).

**Proportion of Total Herd.** The FARM method consistently resulted in lower accuracy (Figures 1–3; Supplemental Table S5; https://doi.org/10.3168/jds.2018-15134), kappa, sensitivity (Supplemental Table and Figures S6–S9; https://doi.org/10.3168/jds.2018-15134), and NPV (Supplemental Table and Figures S10–S13; https://doi.org/10.3168/jds.2018-15134) rela-

---

**Figure 3.** Pooled accuracy (circles) and Cohen’s kappa (triangles) for farm classifications when using thresholds of (A) \( \leq 20\% \) for moderate or (B) \( <10\% \), (C) \( <5\% \), and (D) \( \leq 1\% \) for severe injuries on the hock joint. Classifications were determined by comparing the true prevalence of moderate and severe hock injuries among all lactating cows on each farm against estimates (10,000 sample repetitions each) generated by sampling. Six sample sizes were calculated using a formula with precision (\( d \)) = 15, 10, or 5\% applied to either all lactating cows or a single pen of high-producing cows. The filled symbols indicate when the formula using \( d = 10\% \) was applied to the total herd size, with lactating cows selected in proportion to their representation in the herd. The dashed lines indicate levels of agreement based on kappa: \( \kappa \leq 0.2 \) = slight, \( \kappa > 0.2 \) and \( \leq 0.4 \) = fair, \( \kappa > 0.4 \) and \( \leq 0.6 \) = moderate, \( \kappa > 0.6 \) and \( \leq 0.8 \) = substantial, and \( \kappa > 0.8 \) = almost perfect.
tive to applying the formula with \( d = 10\% \) directly to lactating cows. The patterns were similar for specificity and PPV, with few exceptions where the pattern was reversed (specificity +0.7\% for moderate lameness, Supplemental Figure S6; specificity +18.0\% for severe carpal injuries using a threshold of ≤1\%, Supplemental Figure S7; PPV +0.9 to 1.2\% for severe carpal and hock injuries using a threshold of ≤1\%, Supplemental Figures S11 and S12; https://doi.org/10.3168/jds.2018-15134). The FARM method was also often less accurate than when the formula with \( d = 10\% \) was applied to the high-producing pen, but the pattern was more frequently reversed than when comparing to the formula applied to all lactating cows.

For completeness, the metrics for when \( d = 5 \) and 15\% were applied to the total estimated herd size are shown in Supplemental Figures S18–S20 (https://doi.org/10.3168/jds.2018-15134), but are not discussed further.

### Objective 3: Farms with Hospital Pens

In general, using narrower precision values in the sample size formula (i.e., \( d = 5 \) vs. 10 vs. 15\%) improved all classification metrics for the 20 farms with hospital pens, similar to when considering all 38 farms. In contrast, there was no consistent effect on any of the classification metrics when cows in the hospital pen were either included or excluded from the sample, and any differences were small. For example, when the hospital pen was excluded, accuracy was slightly lower (by −0.1 to −2.9\%) for some measures (moderate lameness, Figure 5A; severe lameness using a threshold of <10\%, Figure 5B; severe carpal injuries using a threshold of <10 or 5\%, Figure 6B and 6C; moderate hock injuries, Figure 7A; severe hock injuries with a threshold of ≤1\%, Figure 7D; Supplemental Table S21; https://doi.org/10.3168/jds.2018-15134), but was slightly higher (by 0.1 to 0.6\%) for others (moderate carpal injuries, Figure 6A; severe carpal injuries using a threshold of ≤1\%, Figure 6D), and was either the same or showed no consistent pattern for the rest (severe lameness using thresholds of <5 or ≤1\%, Figure 5C and 5D; severe hock injuries using thresholds of <10 or 5\%, Figure 7B and 7C). Likewise, kappa changed by only −0.06 to +0.03 units when the hospital pen was either excluded or included, and the overall level of agreement did not change. For example, within a given level of precision, there was either substantial or almost perfect agreement for all thresholds of severe hock injuries (Figure 7B, 7C, 7D), regardless of whether cows in the hospital pen were included or excluded. Likewise, there was no consistent effect of excluding the hospital pen on sensitivity and specificity (Supplemental Table and Figures S22–S25; https://doi.org/10.3168/jds.2018-15134) or PPV and NPV (Supplemental Table and Figures S26–S29; https://doi.org/10.3168/jds.2018-15134).

For the same reasons as for the full set of 38 farms, the results for thin cows are not discussed further but are shown in Supplemental Figures S30–S32 (https://doi.org/10.3168/jds.2018-15134).

### DISCUSSION

As the number of cows sampled increased, so did the number of farms accurately classified relative to each threshold analyzed. The number of cows sampled depended on the desired precision used in the sample size calculation as well as the population of cattle to which the formula was applied (i.e., the total estimated herd size, all lactating cows, or a single high-producing pen) and from which cows were selected (i.e., all lactating cows vs. the high-producing pen). Compared with...
applying the formula to all lactating cows, applying the calculation to the total herd size, then selecting lactating cows in proportion to their representation in the herd, was consistently less accurate. However, assessing the entire high-producing pen resulted in similar farm classification compared with all lactating cows. On the farms with hospital pens, including cows from this pen type had little effect on the accuracy of farm classifications.

**Desired Precision**

Across many disciplines, the equations by Cochran (1977) are commonly used to calculate the sample size needed to estimate prevalence. When the other variables in the equations are held constant, the resulting sample size increases when the desired confidence (Z) is higher, the expected prevalence (p) approaches 50% (because the formula multiplies p by 1 – p), or the desired precision (d) is narrower. Because all 4 of the animal welfare assessments on which we based our sample size calculations used Z = 95% and p = 50%, we focused on evaluating the effects of varying the desired precision (d = 15, 10, or 5%). Using narrower precision values always results in selecting a larger number of cows to sample.

Consistent with our predictions, when d decreased from 15 to 10 to 5%, classification metrics improved, whether the calculation was applied to all lactating cows (as with proAction, Dairy Farmers of Canada, 2017; Validus, Urbandale, IA, personal communication) or to just those in the high-producing pen (as with Dairy Well, Walker et al., 2017). In the context of conducting prevalence research, Naing et al. (2006) recommended using d = 5% when p is expected to fall between 10 and 90%. However, depending on the herd size, using narrower precision in the formula may require a very large sample size. In our study, this effect was particularly noticeable when using d = 5 instead of 10%. Reducing d from 15 to 10% increased the sample by only 5 to 46 cows, depending on the farm and whether the formula was applied to the high-producing pen or all lactating cows, but reducing d from 10 to 5% required sampling up to 183 additional cows. In practice, the choice of

![Figure 5](image-url)

*Figure 5.* Pooled accuracy (circles) and Cohen’s kappa (triangles) for farm classifications when using thresholds of (A) ≤15% for moderately or (B) <10%, (C) <5%, and (D) ≤1% for severely lame cows. Classifications were determined by comparing the true prevalence of moderate and severe lameness among all lactating cows on each farm against estimates (10,000 sample repetitions each) generated by sampling. Six sample sizes were calculated using a formula with precision (d) = 15, 10, or 5% applied to lactating cows either including or excluding (filled and open symbols, respectively) those in hospital pens. The dashed lines indicate levels of agreement based on kappa: κ ≤ 0.2 = slight, κ > 0.2 and ≤0.4 = fair, κ > 0.4 and ≤0.6 = moderate, κ > 0.6 and ≤0.8 = substantial, and κ > 0.8 = almost perfect. Missing values for kappa occurred when both the true and estimated prevalence on all farms were above the threshold and thus all agreements were by chance (denominator = 0).
When assessment programs set low (i.e., <5%) thresholds of acceptability for animal-based measures, this suggests they expect to find low prevalence. In turn, this implies that narrower precision values should be used in the sampling formula and that more cows should be sampled to accurately estimate prevalence. However, in practice, the widespread use of some assessments in the industry is relatively new, and the thresholds currently serve more as goals (e.g., National Milk Producers Federation, 2016; Dairy Farmers of Canada, 2017). If these programs succeed in driving continuous improvement in the industry, we expect prevalence to decrease toward these thresholds. When this happens, assessment programs should reevaluate sampling methods to account for lower expected prevalence.

**Sampling from Only the High-Producing Pen (e.g., Dairy Well)**

Previous research (Endres et al., 2014) concluded that selecting from only certain subpopulations of lactating cows (e.g., a high-producing pen) affects prevalence es-
estimates for lameness and hock injuries because $R^2$ was $< 0.9$ ($R^2 = 0.52$ to 0.65 and 0.67 to 0.71, respectively). However, other studies have used less strict criteria to define the strength of associations (e.g., strong $= R^2 > 0.76$ in Madruga et al., 2017; strong and fair $= R^2 > 0.80$ and 0.70, respectively, in Miller-Cushon et al., 2011). In our study, the associations between true prevalence among high-producing versus all lactating cows were fairly strong for hock injuries ($R^2 = 0.69$ to 0.86) but were weaker for lameness ($R^2 = 0.43$ to 0.47).

Until the current study, no previous research evaluated the effects of sampling from the high-producing pen on farm classifications relative to thresholds of acceptability. Dairy Well (Walker et al., 2017) recommends assessing all cows from a single high-producing pen as the “ideal” strategy or, if the former method is infeasible, sampling cows from this pen using $d = 5\%$. Applying the sample size formula to the high-producing pen (objective 1) resulted in smaller or equal sample sizes compared with when applying the calculation to all lactating cows, and the former method had lower classification metrics in many cases. Nonetheless, including all cows from the high-producing pen (objective 2) resulted in farm classifications similar to when using all lactating cows.

The current study suggests that differences in prevalence between high-producing and all lactating cows do not affect farm classifications relative to thresholds for animal-based measures. This finding supports the strategy in some assessment programs (Dairy Well, Walker et al., 2017) of evaluating all high-producing animals as a proxy for the wider population of lactating cows. Farms use a variety of grouping strategies, such as separating pens of cows by parity, stage of lactation, pregnancy status, or other characteristics. Future research could further explore whether groups based on other strategies are representative of the lactating herd.

**Applying the Formula to the Total Herd (e.g., FARM Program)**

Although some programs evaluate all measures on the same cows, other programs score different animals for some measures. The FARM program (National Milk Producers Federation, 2016) evaluates lactating and dry cows for lameness and leg injuries, but this program

![Figure 7](image-url)
scores body condition for all age groups of cattle. The current FARM sampling method applies the formula once to the total herd size, with lactating and dry cows then selected in proportion to their representation in the herd (version 3.0; National Milk Producers Federation, 2016, and the National Dairy FARM Program, Arlington, VA, personal communication). In the current study, applying the formula (using \(d = 10\%\)) to the estimated total herd size resulted in selecting 15 to 51 fewer cows compared with applying the formula directly to the number of lactating cows on each farm. As a consequence, the FARM method resulted in poorer classification metrics for all thresholds. This finding indicates that applying the sample size calculation directly to the population of interest (i.e., lactating cows) results in more accurate farm classifications. When the population of interest is the entire herd (i.e., the FARM program standards for body condition, National Milk Producers Federation, 2016), applying the sampling formula to the whole herd may generate accurate classifications. Therefore, assessment programs may need to use different sample sizes for different measures, depending on the populations of interest.

**Hospital Pens**

We predicted that excluding cows in hospital pens would change farm classifications relative to sampling from all lactating cows. However, excluding the hospital pens had little or no effect on classifications for the subset of 20 farms with this pen type. To encourage farmers to uphold their commitment to provide proper care for their animals, some assessment programs exclude the hospital pen from the sample for animal-based measures (i.e., proAction, Dairy Farmers of Canada, 2017; Validus, Urbandale, IA, personal communication). In theory, excluding the hospital pen from an assessment could incentivize farms to strategically move problem animals to that group in anticipation of an evaluation. Another approach is to assess the hospital pen but exclude it from the thresholds to avoid giving farmers the impression of being penalized for identifying and treating those animals (i.e., Dairy Well, Walker et al., 2017). Our results suggest that excluding cows in the hospital pens has little effect on farm classification.

**Tradeoffs Among Metrics**

Across all thresholds for lameness and leg injuries, all sampling strategies resulted in relatively high accuracy (at least 4 in 5 assessments would result in an accurate classification), sensitivity, and positive predictive value, whereas specificity and NPV, while often high, were more variable. There is an inherent tradeoff between sensitivity versus specificity and PPV versus NPV, and assessment programs may prioritize these metrics differently than dairy producers. Assessment programs seeking to provide assurance to consumers may value sampling strategies that generate high sensitivity and NPV, whereas dairy producers may prefer high specificity and PPV. In many assessment programs, farms failing to meet a threshold of acceptability would be expected to take corrective action and show improvement at the next evaluation. From the perspective of assessment programs, high sensitivity indicates that most farms which would fail to meet a threshold of acceptability based on true prevalence would be captured in the sample, meaning most farms needing improvement would be identified. Likewise, high NPV indicates most farms meeting a threshold in the sample would indeed pass based on true prevalence and would not warrant further attention. From the perspective of dairy producers, high PPV and specificity indicate that being asked to take corrective action would likely be warranted: most farms failing to meet thresholds of acceptability in the sample would indeed fail based on true prevalence, and most farms that would meet thresholds based on true prevalence are indeed classified as passing (i.e., a low number of FP), respectively.

If continuous improvement in the industry results in farms that currently have high prevalence moving closer to the threshold, the challenge of classifying farms correctly may increase. When true prevalence is close to the threshold, the task of determining appropriate sample size for classification becomes similar to evaluating how closely sample estimates reflect true prevalence (e.g., Endres et al., 2014; Van Os et al., 2018). To classify farms near the threshold with greater certainty, further study is warranted on multi-step sampling using a decision tree to balance feasibility with accuracy. First, a tolerance level above the threshold is established, and an initial sample is taken. If the initial prevalence estimate falls above the tolerance level (e.g., an estimate of >21% prevalence for a <10% threshold of acceptability in severe carpal injuries in our data set), the farm can be confidently classified as failing to meet the standard; however, if the farm is closer to the threshold, an additional sample of cows is assessed. Heath et al. (2016) investigated multi-step sampling approaches for lameness, but their gold standard for classifying farms used prevalence estimates based on the Welfare Quality (2009) sampling calculation, which has limited accuracy (see Van Os et al., 2018). Further studies using the true prevalence to classify farms are needed to evaluate multi-step sampling strategies for various animal-based measures.
CONCLUSIONS

All sampling strategies classified farms with ≥79% accuracy for all thresholds of acceptability for lameness and leg injuries, and sensitivity and PPV were also relatively high. When more cows were sampled, classification performance improved. Because using narrower precision values can dramatically increase sample size on larger herds, assessment programs need to consider feasibility constraints in addition to their tolerance for misclassification. Relative to selecting from all lactating cows, excluding cows in the hospital pen had little effect on farm classification. However, assessing the entire high-producing pen or all lactating cows resulted in similar classifications. Therefore, assessing the high-producing pen could serve as a reasonable proxy for the larger population of lactating cows for the purposes of classifying farms relative to thresholds of acceptability.

ACKNOWLEDGMENTS

We are grateful to Cassandra Tucker (University of California, Davis) for helpful discussions on the topic of this study. Many thanks are extended to Tracy Burnett (Vancouver, BC, Canada), Paulo Henrique Manske Doering, Vanessa Groenwold (Chácará Douwe, Castro, PR, Brazil), Luís Andréz Gonzalez, Gabriela Marquete, and Angélica Roslindo (Florianópolis, SC, Brazil) for their help with data collection. We thank the Castrolanda Cooperativa Agroindustrial (Castro, PR, Brazil), Capal Cooperativa Agroindustrial (Arapoti, PR, Brazil), Cooperativa Agropecuária Witmarsum (Witmarsum, PR, Brazil), and the farmers, local extension agents, and veterinarians that participated in the study, especially Hugo Richard Dick and Gunther Schartner (Cooperativa Agropecuária Witmarsum, Witmarsum, PR, Brazil), Rodrigo Navarro (Capal Cooperativa Agroindustrial, Arapoti, PR, Brazil), and Junio Fabiano dos Santos (Castrolanda Cooperativa Agroindustrial, Castro, PR, Brazil), for their help with identification of farms and for sharing their knowledge about the milk production systems in the region.

The data collection portion of this study was funded by the Science Without Borders Program (CNPq–National Council for Scientific and Technological Development, Brazil, grant no. 400850/2013-3), which also provided funds to M. A. G. von Keyserlingk for visits to Brazil that enabled this outstanding collaboration. Maria J. Hötzel also received support from CNPq (grant no. 311509/2015-0), and J. H. C. Costa was supported by a postdoctoral fellowship from CAPES (Ministry of Education, Brazil) and the University of British Columbia Animal Welfare Program throughout the development of the project. J. M. C. Van Os was supported through an Natural Sciences and Engineering Research Council (NSERC) Industrial Research Chair awarded to M. A. G. von Keyserlingk and D. M. Weary with industry contributions from the Dairy Farmers of Canada (Ottawa, ON, Canada), British Columbia Dairy Association (Burnaby, BC, Canada), Westgen Endowment Fund (Milner, BC, Canada), Interact Canada Corporation (Kirkland, QC, Canada), Novus International Inc. (Oakville, ON, Canada), Zoetis (Kirkland, QC, Canada), BC Cattle Industry Development Fund (Kamloops, BC, Canada), Alberta Milk (Edmonton, AB, Canada), Valacta (St. Anne-de-Bellevue, QC, Canada), and CanWest DHI (Guelph, ON, Canada).

REFERENCES


