Animal (2016), **10:2**, pp 349–356 © The Animal Consortium 2015
doi:10.1017/S1751731115001536

animal

Sequential sampling: a novel method in farm animal welfare assessment

C. A. E. Heath[1†], D. C. J. Main[1], S. Mullan[1], M. J. Haskell[2] and W. J. Browne[3,4]

[1]*School of Veterinary Sciences, University of Bristol, Langford House, Langford, Bristol BS40 5D, United Kingdom;* [2]*SRUC, West Mains Road, Edinburgh EH9 3JG, United Kingdom;* [3]*Graduate School of Education, University of Bristol, Helen Wodehouse Building, 35 Berkeley Square, Clifton, Bristol BS8 1JA, United Kingdom;* [4]*Centre for Multilevel Modelling, 2 Priory Road, Bristol BS8 1TX, United Kingdom*

(Received 3 October 2014; Accepted 6 July 2015; First published online 12 August 2015)

Lameness in dairy cows is an important welfare issue. As part of a welfare assessment, herd level lameness prevalence can be estimated from scoring a sample of animals, where higher levels of accuracy are associated with larger sample sizes. As the financial cost is related to the number of cows sampled, smaller samples are preferred. Sequential sampling schemes have been used for informing decision making in clinical trials. Sequential sampling involves taking samples in stages, where sampling can stop early depending on the estimated lameness prevalence. When welfare assessment is used for a pass/fail decision, a similar approach could be applied to reduce the overall sample size. The sampling schemes proposed here apply the principles of sequential sampling within a diagnostic testing framework. This study develops three sequential sampling schemes of increasing complexity to classify 80 fully assessed UK dairy farms, each with known lameness prevalence. Using the Welfare Quality herd-size-based sampling scheme, the first 'basic' scheme involves two sampling events. At the first sampling event half the Welfare Quality sample size is drawn, and then depending on the outcome, sampling either stops or is continued and the same number of animals is sampled again. In the second 'cautious' scheme, an adaptation is made to ensure that correctly classifying a farm as 'bad' is done with greater certainty. The third scheme is the only scheme to go beyond lameness as a binary measure and investigates the potential for increasing accuracy by incorporating the number of severely lame cows into the decision. The three schemes are evaluated with respect to accuracy and average sample size by running 100 000 simulations for each scheme, and a comparison is made with the fixed size Welfare Quality herd-size-based sampling scheme. All three schemes performed almost as well as the fixed size scheme but with much smaller average sample sizes. For the third scheme, an overall association between lameness prevalence and the proportion of lame cows that were severely lame on a farm was found. However, as this association was found to not be consistent across all farms, the sampling scheme did not prove to be as useful as expected. The preferred scheme was therefore the 'cautious' scheme for which a sampling protocol has also been developed.

Keywords: sequential sampling, dairy cows, lameness, animal welfare, prevalence

Implications

Lameness is an important welfare and production concern for dairy farms. To establish lameness prevalence on a farm, a sample of cows needs to be assessed. Larger sample sizes provide more accurate results but take longer and incur a greater cost. Sequential sampling provides a means of stopping sampling early without sacrificing accuracy, and can be used to establish whether lameness levels on a farm are acceptable. A sampling protocol is presented that can be applied as part of a targeted surveillance tool, where greater certainty is attached to declaring a farm to have unacceptable levels of lameness.

† E-mail: Cheryl.heath@bristol.ac.uk

Introduction

Animal-based measures, or welfare outcome measures are considered to provide a more direct account of welfare compared with resource-based measures (Webster *et al.*, 2004). This change in perception has influenced the way in which welfare assessments are carried out and welfare outcome measures are increasingly being included in welfare assessment protocols. For example, the holistic species specific Welfare Quality protocols consist of predominantly animal-based measures. One of the drawbacks of animal-based measures, however, is that they are time consuming to collect as they require a sample of animals to be observed individually. Lameness in dairy cows is an important concern urgently in need of attention (FAWC, 2009). It is a painful

https://doi.org/10.1017/S1751731115001536 Published online by Cambridge University Press

condition (Rushen *et al*., 2007) that has consequences in terms of production that include milk loss (Amory *et al*., 2008), fertility problems (Garbarino *et al*., 2004), culling (Booth *et al*., 2004) and costs associated with treatment (Kossaibati and Esslemont, 1997). Recent estimates of the levels of lameness suggest that UK farm prevalence levels are increasing, from an average of 21% in 1989/90 (Clarkson *et al*., 1996) to 36% in 2006/07 (Barker *et al*., 2010). There is therefore a need for an assessment, which is both accurate and feasible.

In cases where an agency would charge a fee to carry out a welfare assessment, the cost of this assessment is likely to be partly attributable to the time taken to carry out the assessment, which is a function of the number of animals that are assessed. Hence, an argument has been made for reducing sample sizes to minimise these costs (Sorensen *et al*., 2007). Sample sizes vary according to the purpose of assessment, and the level at which the estimates are intended to be used. For example, the protocol used by the Red Tractor Assured Dairy Farm Scheme (which assures over 11 000 dairy farms in the UK (CHAWG, 2013)) uses a sample size of 10 cows/farm (where the average UK farm has 125 cows (DairyCo, 2013)) and is expected to provide accurate estimates for the UK dairy herd as a whole. Similarly, sampling 50 birds/flock (where maximum flock size is 2000 birds and 16 000 birds for Soil Association and Freedom Food certified farms, respectively) provides satisfactory estimates at the farm assurance scheme level, but, this sample size is limited in terms of the accuracy of prevalence estimates it can provide at the individual farm level (Main *et al*., 2012). At the farm level, which is the focus of this study, larger sample sizes are required as the estimates produced will themselves be of interest rather than simply contribute to an aggregated estimate.

The dilemma therefore, between the need for smaller sample sizes on the one hand, and the need for accuracy on the other, has been addressed in clinical trials through the use of sequential sampling. Smaller sample sizes reduce the financial costs and minimise the exposure of participants to treatments whose effects may still be uncertain. Unlike in a fixed size sampling scheme where a decision is made based on a single large sample, in sequential sampling, a number of smaller samples are drawn in stages, and, there is an option to stop sampling early at each stage. Based on the hypothesis testing approach, sampling error cannot be inflated through drawing multiple samples as error rates are fixed from the outset. The conditions concerning the decision made at each sampling event, whether there is a positive or a negative outcome, or whether further data need to be collected, are based on rules which change dynamically according to the data. Where regular data and safety monitoring of clinical trials is required by independent boards, interim analyses from each sampling event provide the opportunity to assess participants for adverse effects as well as monitor for efficacy or futility of treatment (e.g., Aberle *et al.* (2011) and Dember *et al.* (2008)). Any number of such interim analyses can be carried out (up to the total number of participants) at any point during the data collection. Published clinical trials have included, for example, a single interim analysis at 66.8% of the sample (Bang *et al*., 2012); two interim analyses at 25% and 50% (Ballard *et al*., 2006) and two interim analyses at 33% and 66% (Bussel *et al*., 2007).

Sample sizes for a number of sampling events used in the sequential schemes in this study will be derived from the Welfare Quality herd-size-based sampling scheme. The Welfare Quality protocol for dairy cows provides two sampling schemes for herd sizes up to 300 cows. For the more desirable option A, sample sizes range from 30 to 73 cows, and, when option A is not feasible, the smaller sampling scheme, option B, which ranges from 30 to 55 cows, is suggested (P98, Welfare Quality®, 2009). This study will be concerned only with the Welfare Quality 'A' scheme. In the protocol this sampling scheme is used to estimate prevalence levels, which are then integrated into aggregated scores. Here, the sample prevalence will be used to test for unacceptability as defined by exceeding a threshold for a given herd level prevalence. The performance of the sampling schemes will be evaluated against the standard fixed size Welfare Quality sampling scheme in terms of accuracy and time taken, where the accuracy can be understood to be the capacity of the test to detect the true state of a herd, and the time taken can be measured in average sample size over multiple simulations of the procedure.

A two-stage sequential sampling scheme involves two samples being drawn with the option of stopping after the initial sample. The decision to continue sampling after the first sample or to draw a second sample concerns rules set relative to the threshold value. These rules relate to the 'tolerance' set around the threshold, and are best illustrated by way of an example. For a farm of 120 cows, the Welfare Quality 'A' scheme suggests that 54 cows are mobility scored using a three point scale, 0 for not lame, 1 for lame and 2 for severely lame. Where in this situation the interest is only in presence or absence of lameness, both lame and severely lame cows are classified simply as being lame and no differentiation is made. The lameness scores are aggregated to form a sample prevalence, which becomes the estimate of the farm prevalence. If the threshold is set at 30%, a rule to stop early and declare a farm problem-free could be set around a limit of 25%, or at a tolerance level of 5% below the threshold (6.75 lame cows from the first sample of 27). Thus, if after scoring 27 cows, <7 are lame, then sampling could stop after scoring the 27 cows and the farm could be classified as not having a problem. If, on the other hand, >7 are lame, the decision could be taken to sample another 27 cows. Similarly, a rule could also be applied for deciding early that a farm has a problem. In this case a rule could be applied with a tolerance of 5% above the threshold, so where there is $>35\%$ lameness (or 9.45 cows from the sample of 27) a farm would be identified as a problem farm. However, if there are between seven and nine lame cows in the first sample drawn, there is less certainty and a further 27 cows would need to be scored. When a second

https://doi.org/10.1017/S1751731115001536 Published online by Cambridge University Press

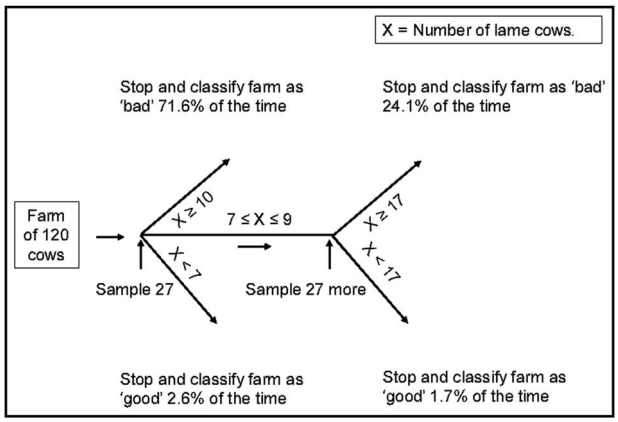


Figure 1 Illustration of a sequential sampling scheme which uses a threshold of 30% to classify a farm with 120 cows and 40% lameness.

sample is drawn the decision is then based on the original threshold of 30%. A summary of this scheme is presented in Figure 1.

Where there is a known farm prevalence, it is possible to test how this scheme performs. By running 100 000 simulations of drawing random samples, the sequential scheme is able to correctly identify a problem farm 95.7% of the time. In terms of the cost associated with the number of animals scored, in the fixed size Welfare Quality sampling scheme this would have been associated with scoring 54 cows, whereas for the sequential scheme, this would be associated with scoring 34 cows on average. This is calculated from the proportion of times when sampling stops early and only 27 cows are scored plus the proportion of times a second sample is drawn and 54 cows are scored.

The example above describes a farm of 120 cows with a lameness prevalence of 40%. If a scheme is used where the total sample size for those farms that need a second sample to be drawn is equal to the Welfare Quality sample size, then, when compared to the fixed size Welfare Quality sampling scheme, (i) there are more incorrectly classified farms (4.2% as opposed to 2.8%); (ii) on average the sample size is greatly reduced (from 54 to 34); and (iii) there is the assurance that sample sizes will not be greater than Welfare Quality sample sizes on any farm.

The sequential scheme involves balancing the need for accuracy with the need for smaller sample sizes. However, if correctly classifying farms was considered to be more important, the sample size in each stage could be increased. For this example, if there were 33 cows in each half, then (i) there would be a mistaken classification 2.8% of the time (a rate equal to that of the fixed sample size scheme); (ii) the expected sample size would be 40 as opposed to 54 (still greatly reduced); however, (iii) on 21.6% of farms 66 cows would be sampled, that is, more than the Welfare Quality fixed size sampling scheme.

The aim of this study is to investigate which sequential sampling schemes are able to provide an optimum balance of accuracy and length of assessment to meet the needs of an on-farm welfare assessment.

Methods

Farm recruitment

Data were used from another much larger study, which looked at a number of different welfare outcome measures collected at different times of year, on organic and non-organic systems. Full details may be found in the study by Rutherford *et al.* (2009). A brief description of the farm visits, farm recruitment and lameness scoring is detailed below.

Farm visits

A total of 80 UK dairy farms were recruited, 40 organic and 40 matched non-organic. The farms visited here were visited once during the autumn (September to October) and once in the spring (March to June) with organic and matched non-organic farms visited within a 2-week period. Locomotion scoring was carried out using a four point scale, a simplification of one used previously (Manson and Leaver, 1988). Details of this scale and how it compares to that used by Welfare Quality is shown in Table 1. All cows in the milking herd were scored upon leaving the milking parlour at either morning or afternoon milking. Locomotion scoring was based on leg movement parameters such as stride length, tracking, adduction or abduction and weight bearing. Inter-observer reliability is reported in the study by Rutherford *et al.* (2009).

Three two-stage sequential sampling schemes of increasing complexity were investigated, and were compared with a fixed size scheme (the Welfare Quality herd-size-based sampling scheme). For the sequential schemes, the potential (maximum) number of animals sampled (over the two stages) was equal to that of the fixed size sampling scheme, and the number of animals sampled at the first stage was half that number. The sequential sampling schemes investigated were, (1) a basic two-stage scheme, (2) a scheme that placed greater emphasis on correctly identifying 'problem' farms and (3) a scheme that incorporated severe lameness as a variable in the decision criterion. Comparison of the schemes to the fixed size sampling scheme was made in terms of accuracy (misclassification rate), and mean sample size (as a proxy for time taken). Accuracy was assessed on all farms irrespective of farm status, but due to farms either being classified as 'good' or 'bad' it was only possible to calculate either sensitivity or specificity for each farm.

The analyses investigated the effects of changing the threshold and the tolerance, where the threshold was the lameness prevalence used to classify a farm in terms of acceptability, and the tolerance was the range of farm prevalence levels around the threshold used at the first sampling event, to decide whether or not a second sample would be drawn. For each of the models, random samples were drawn from each of the farms according to either the fixed size or the sequential sampling scheme, and consequently each farm was classified as to its lameness status. Both the mean sample size and the misclassification rate were calculated by averaging first across the 80 farms, then across the 100 000 simulations. All analyses were carried out using the program R (version 2.14.1).

https://doi.org/10.1017/S1751731115001536 Published online by Cambridge University Press

Table 1 *Comparison of locomotion score categories (adapted from Rutherford* et al*. (2009))*

Score	Description of score category (Rutherford *et al*., 2009)	Equivalent Welfare Quality® score
1	Cow walks freely and evenly. Cow shows even stride lengths with good tracking[a], no adduction[b] or abduction[c] and an even walking rhythm	0
2	Cow walks with small gait abnormality. Gait pattern is almost normal but cow shows slight but noticeable deviations from normal tracking or clear adduction or abduction	0
3	Cow is clearly walking abnormally, with clear deviations from normal tracking and abduction or adduction. Stride length can be reduced and stride rhythm may be irregular	1'Moderately lame'
4	Movement is clearly impaired and the cow is seen to limp or hobble as she moves. The cow shows uneven stride length and timing and appears reluctant to bear weight on one or more limbs	2 'Severely lame'

[a]Tracking: proximity of the hind foot to the print of the fore foot, normal walking would entail having the hind foot placed in the print of the fore foot.
[b]Adduction: the limb is turned outwards, away from the body, while the hock is turned inwards.
[c]Abduction: the limb is turned inwards towards the body.

Classification of farms in terms of acceptability was based on a threshold set at the mean lameness prevalence (18%). This was an arbitrary value with the potential to bias the results to higher rates of misclassification. Normally the welfare relevance of a threshold would be incorporated through expert opinion. Misclassification rates and mean sample sizes were used as metrics to compare sequential schemes with a fixed size sampling scheme (the herd-size-based Welfare Quality sampling scheme). For the sequential sampling scheme at the first sampling event, half the total sample of the fixed size scheme was drawn. If at that stage, the sample prevalence fell outside a given distance (tolerance) from either side of the threshold, sampling stopped and a farm was classified. Otherwise, if it fell between those limits a second sample was drawn, so that the total number of animals sampled was equal to that of the fixed size scheme. For example, for a farm of 120 cows, a sample of 54 cows was drawn. If the tolerance was set at 10% either side of the threshold, then when using the threshold of 18%, sampling stopped when there were three or fewer lame cows (0.08×27) and the farm was declared 'good', or if there were more than eight lame cows (0.28×27) and the farm was classified as a problem farm. If the number was between those limits then a second sample was drawn and a decision was made based on the prevalence from the 54 cows. At that stage, a total of 10 or fewer lame cows in the sample of 54 classified a farm as 'good' (0.18×54), whereas >10 lame cows, a problem farm.

Results

The data used in this study consisted of the lameness scoring of 28 559 cows from 80 farms. Lameness prevalence as a proportion, varied from 0.04 to 0.42, with a mean value of 0.18 and the third quartile at 0.22. Severe lameness prevalence (score 4 s) varied from 0 to 0.15 with a mean value of 0.04 and the third quartile at 0.05.

Figures 2a and b show misclassification rates and (relative) mean sample sizes for a basic two-stage sampling scheme compared with the fixed size sampling scheme. By increasing the tolerance around the threshold a second sample is drawn more often. This increases average sample sizes, and hence reduces the misclassification rate. Figure 2a shows that the misclassification rate of the simple and the sequential scheme are visually indistinguishable from each other beyond a tolerance of 0.05, and even at that level, a difference can only be perceived for farms where there is a lameness prevalence of approximately between 0.2 and 0.3. More generally, misclassification peaks where the true farm prevalence is close to the threshold, as these farms are the most difficult to classify. Figures 2a and b show that a tolerance of 0.05 or 0.075 would be the best choice for the basic two-stage sequential scheme depending on the relative importance of mean sample size and misclassification rate. Based on the small additional misclassification rate associated with when the tolerance is at 0.05, all further analyses will assume a tolerance of 0.075.

A second (cautious) sequential scheme was investigated that placed greater emphasis on correctly classifying 'bad' farms. This was achieved by making an adjustment to the basic sequential model, so that sampling only stopped early when the initial sample drawn indicated that it was a 'good' farm. Figure 3a shows lower misclassification rates for the 'cautious' scheme, and for a tolerance of 0.075 below the threshold, Figure 3b shows mean sample sizes are greater for farms with higher prevalence compared with the basic sequential scheme (though still much smaller than the fixed size sampling scheme).

A cautious scheme thus provides a way of assuring greater confidence when declaring a farm to be 'bad' while still benefitting from smaller average sample sizes on good farms. However, accuracy may be further improved by incorporating a second predictive variable, prevalence of severely lame cows (score 4 s), for a cautious scheme, which is also 'weighted'. The motivation here is that as there is a correlation of 0.58 between the proportion of cows that are lame on a farm, and, the proportion of the lame cows that are severely lame, then detecting more severely lame cows in a sample should be associated with a 'bad' farm. Severely lame cows are given a weight (w) and the sample size is adjusted to accommodate this by the following calculation:

$$\text{Severity adjusted prevalence} = \frac{((w \times \text{number of severely lame cows}) + \text{number of lame cows})}{\left(\frac{w+1}{2}\right) \times \text{sample size}}$$

https://doi.org/10.1017/S1751731115001536 Published online by Cambridge University Press

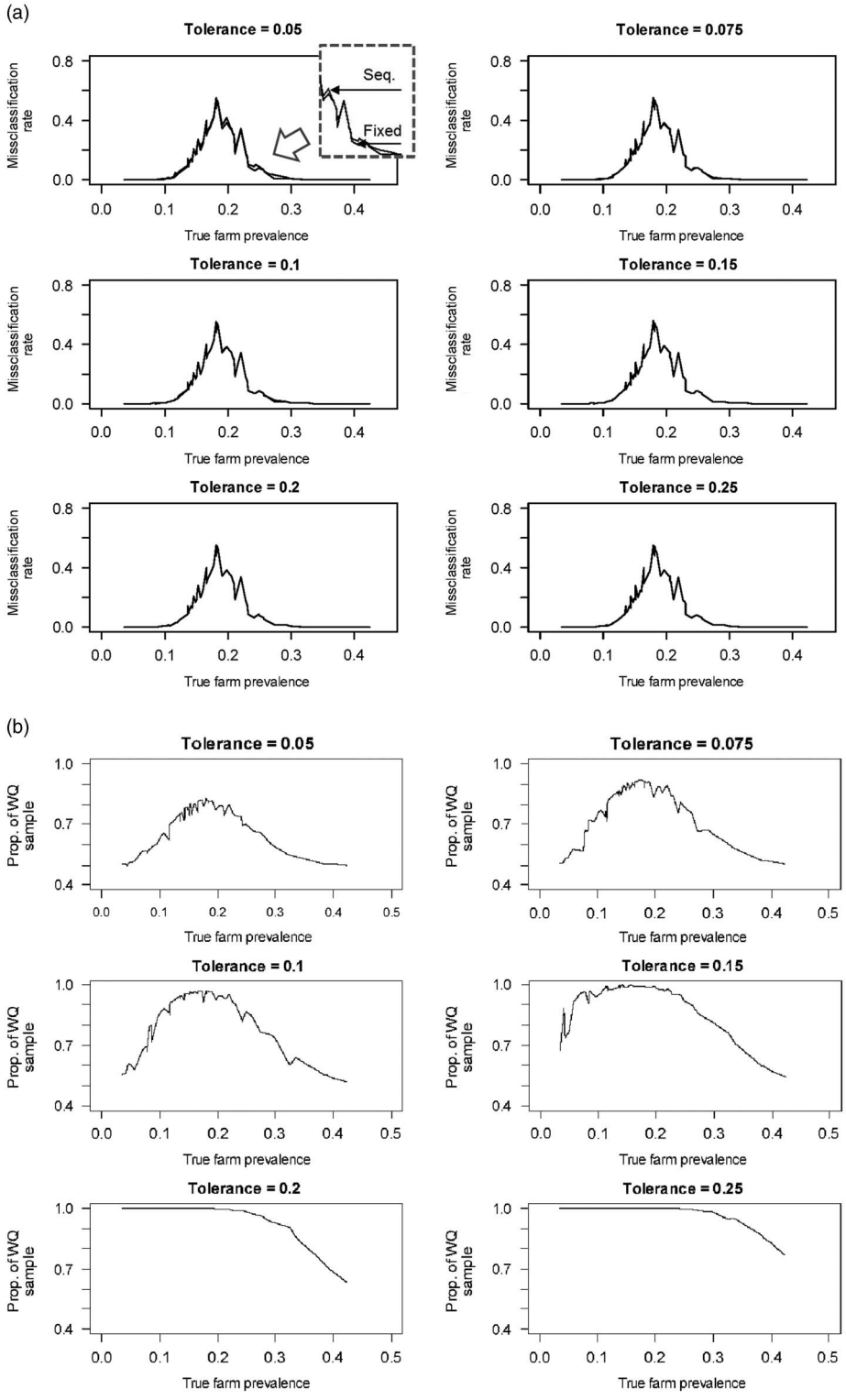


Figure 2 A comparison of fixed and sequential sampling schemes, in terms of (a) misclassification rate, and (b) relative mean sample size. Sampling schemes used to classify 80 dairy farms, based on 100 000 randomly drawn samples, where a threshold 18% lameness has been applied.

Despite the association, Figure 4 shows that when the threshold is set at the mean (18%) and the tolerance is at 0.075, the misclassification rate shows only a negligible reduction at a weighting of 1.2 before it increases. Upon closer inspection of the data, when the threshold is set at the mean, it can be seen that 24 farms are located within 3% of the threshold. These farms are the most likely to be misclassified and the aforementioned association with

https://doi.org/10.1017/S1751731115001536 Published online by Cambridge University Press

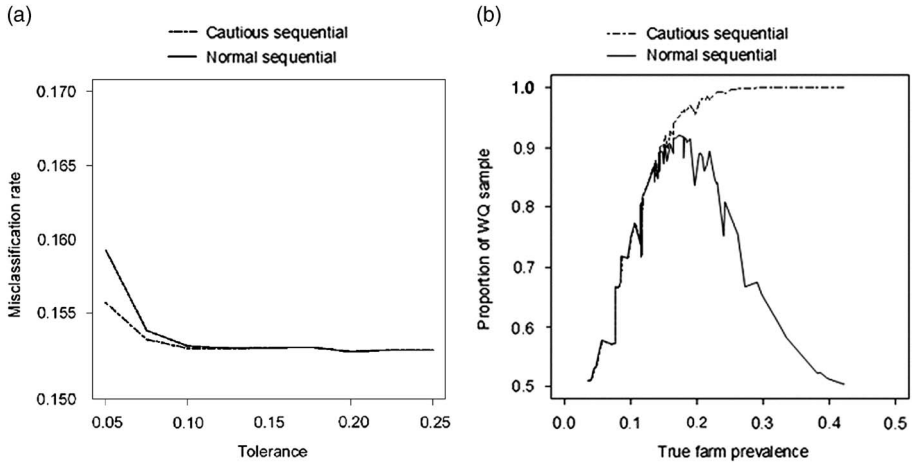


Figure 3 A comparison of (a) misclassification rates (with a variable tolerance) and (b) mean sample sizes (with a fixed tolerance of 0.075) for two sequential sampling schemes used to identify 'good' and 'bad' farms, based on a lameness prevalence threshold of 18%. Sampling schemes used to classify 80 dairy farms, using 100 000 randomly drawn samples, where a threshold 18% lameness has been applied.

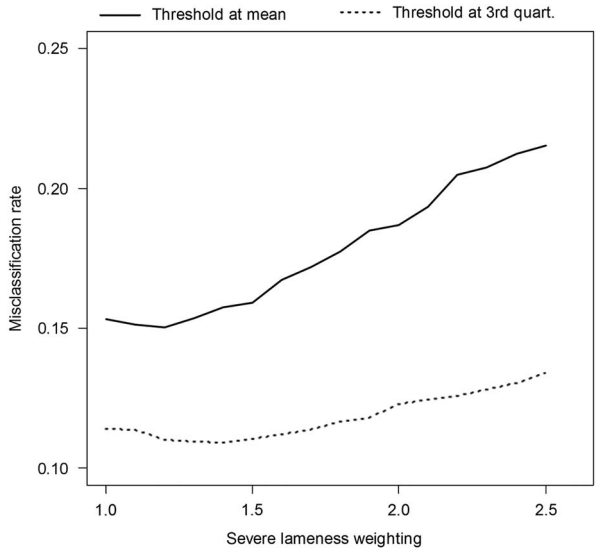


Figure 4 Misclassification rates for a 'weighted' sequential scheme, based on 100 000 randomly drawn samples from 80 dairy farms, at two different thresholds.

severe lameness has been reduced to a correlation of 0.09 for those farms. When the threshold is moved to the third quartile, the number of farms likely to be misclassified is reduced to 18 and the correlation of 0.14 is somewhat stronger, leading to a small reduction in misclassification rate at a weighting of 1.4.

Table 2 presents a summary of the schemes and shows that the sequential schemes are able to achieve comparable levels of accuracy to the fixed size scheme, but with smaller sample sizes. The effect of the threshold can also be seen. The sensitivity of the test, the proportion of farms that have lameness prevalence levels above the threshold which also test as 'bad', decreases for the cautious scheme. This is because by no longer stopping early for bad farms, the risk of misclassifying 'good' farms as 'bad' based on a smaller sample is eliminated. Likewise, the specificity, the proportion of 'good' farms which are also classified as 'good' increases. This is because there are fewer 'good' farms that have been misclassified as 'bad'. Sensitivity is further increased and specificity is further decreased when severe lameness is incorporated as an additional variable.

This study has focused on achieving comparable levels of accuracy for sampling schemes that give smaller average sample sizes when compared with a fixed size scheme. An alternative approach might have been to instead focus on increased accuracy. For the 'sequential scheme plus', the sample sizes from the basic sequential scheme have been inflated. Here, they have been multiplied by a factor of 1.2 as this provides a reasonable compromise in terms of average sample size. Thus, in comparison with the fixed size sampling scheme, the sequential scheme plus has greater accuracy, and smaller average sample sizes, but, for some farms closer to the threshold, if a second sample is drawn, the sample size for these farms will then be greater.

Discussion

The aim of this study was to develop a feasible and accurate sequential sampling scheme, which could be used to classify dairy farms according to lameness prevalence. While the emphasis could have been on either accuracy or sample size, this study focused on schemes that were as good as a fixed size sampling scheme but benefited from smaller average sample sizes. By using a diagnostic approach, a threshold was used to classify farms in terms of acceptability. Farms with a lameness prevalence close to this threshold were the most difficult to classify, and had the greatest rates of misclassification. In this study, the mean value from the dataset of 80 farms was chosen for the threshold for illustration purposes. As there were generally more farms found around the mean than, for example, the third quartile and it was farms near the threshold that were often misclassified, this resulted in higher rates of misclassification for all methods.

https://doi.org/10.1017/S1751731115001536 Published online by Cambridge University Press

Table 2 *Average sample size, accuracy, sensitivity and specificity for different sampling schemes using thresholds set at the mean (18%) and third quartile (22%) for classifying problem farms on 80 UK farms*

	Simple scheme[a]		Sequential scheme[b]		Sequential scheme plus[c]		'Cautious' sequential scheme[d]		'Cautious weighted' sequential scheme[e]	
Threshold	18%	22%	18%	22%	18%	22%	18%	22%	18%	22%
Mean sample size	69.8	69.8	54.8	49.7	63.9	60.0	61.6	54.4	60.3	53.1
Accuracy	84.8	88.6	84.6	88.4	86.2	89.6	84.7	88.6	85.0	89.1
Sensitivity	80.6	79.5	80.5	79.1	83.7	81.4	80.2	78.5	76.7	70.1
Specificity	88.0	91.7	87.9	91.4	88.1	92.3	88.2	92.0	91.3	95.4

[a]Sample sizes based on Welfare Quality.
[b]Sample sizes based on Welfare Quality with interim sampling at 50%.
[c]1.2 × (sample sizes based on Welfare Quality with interim sampling at 50%).
[d]Sample sizes based on Welfare Quality with interim sampling at 50%, stopping early only for farms with prevalences below the threshold.
[e]Sample sizes based on Welfare Quality with interim sampling at 50%, stopping early only for farms with prevalences below the threshold. Severe lameness prevalence weighted as an additional variable. Weighting 1.2 for analysis at the mean and 1.4 for analysis at the third quartile.

In this study, three two-stage sequential sampling schemes were investigated, all of which had comparable levels of accuracy to the fixed size Welfare Quality sampling scheme, while using smaller average sample sizes. The total potential sample sizes used in the sequential sampling schemes were the Welfare Quality herd-size-based sample sizes, under the assumption that these are in fact acceptable. The first scheme examined here, illustrated the role of the tolerance on both mean sample sizes and accuracy, where it was possible to stop sampling early for both 'good' and 'bad' farms. Declaring a farm to have unacceptable levels of welfare may potentially have negative financial consequences for farmers. For this reason, there is a greater need for accuracy in making this decision than in declaring a farm to be 'good'. By only stopping early for 'good' farms in the second 'cautious' scheme, the decision to classify farms as 'bad' was always made based on the larger sample size. This led to a small increase in average sample size, but sample sizes were still greatly reduced compared with the fixed size sampling scheme. The third scheme extended the cautious scheme, and sought to take advantage of the presence of severely lame cows as a risk factor for farms with high levels of overall lameness (Main *et al*., 2010), and the association found in this study between the proportion of cows which were lame and the proportion of lame cows that are severely lame. However, when the number of severe lame cows was weighted, there were only minimal gains in terms of accuracy and reduction in average sample size. Thus, although the decision to include severe lameness as a factor in the decision criterion was based on a moderately strong overall correlation (0.53), this was not always found to be the case for the group of farms around the threshold. Consequently, for many farms little additional predictive value was added, and higher weightings increased misclassification, especially for farms with lameness prevalence close to the threshold. Therefore, the scheme identified as the most suitable for use as part of a welfare assessment tool is the cautious scheme for which a protocol is included as Supplementary Table S1.

A number of practical considerations should be made if this protocol were to be used. First, as it has been shown that there is no strong association between different welfare indicators (Heath *et al*., 2014) it is likely that it could only be used for a single welfare indicator and may be best suited as part of a targeted lameness surveillance scheme. Second, a challenge posed by welfare outcome-based assessment in general, is the assumption of random sampling. The most convenient way to sample dairy cows is to mobility score the cows as they are leaving the milking parlour, as this causes minimal disruption to the farm, and an association between milking order and lameness has been identified (Main *et al*., 2010). Based on this association, an alternative to random sampling has been proposed by Main *et al*. (2010) who suggest locomotion scoring up to 100 cows from the middle of the milking order. This scheme has since been found to be able to provide accurate herd level prevalence when also applied to large herds with multiple groups (Hoffman *et al*., 2013). In general, though, the question of how a random sample may be achieved on farm can to some extent be considered to be a logistical one, and is common to most sampling schemes. Finally, as assessing the whole herd is rarely feasible, by only scoring a sample of animals, sampling error will inevitably be introduced, and this will be in addition to measurement error associated with inter-assessor reliability (D'Eath, 2012). Automatic lameness detection systems have the potential to offer a solution on both these accounts and remain an interesting area for future development, for example, the GAITWISE system proposed in the study by Maertens *et al*. (2011).

The schemes investigated here have used the Welfare Quality sampling scheme and have considered only two stopping points. As the same Welfare Quality sampling scheme is used for a number of different welfare indicators, as it stands, this sequential model could easily be applied to other measures with similar distributions such as body condition score, with careful selection of the threshold. However, the potential for assessing these jointly in a more complex protocol, such as the Welfare Quality protocol, is limited due to the lack of association found between different indicators. Future work could involve including further stopping points, and at the extreme, this could be as many stopping points as

https://doi.org/10.1017/S1751731115001536 Published online by Cambridge University Press

there are cows on a farm. By combining sample size inflation from the sequential sampling plus scheme, with additional stopping points, more advanced schemes that optimised both accuracy and sample size could be developed to provide the best of both worlds.

Acknowledgements

Funding for this project was provided by the Agriculture and Horticulture Development Board DairyCo division and AssureWel. The AssureWel project is a collaboration between the RSPCA, the Soil Association and the University of Bristol and is funded by the Tubney Charitable Trust. The authors would also like to thank the following for the collection of the data, along with the farmers that took part, Kenneth Rutherford, Fritha Langford, Mhairi Jack and Lorna Sherwood.

Supplementary material

For supplementary material/s referred to in this article, please visit http://dx.doi.org/10.1017/S1751731115001536

References

Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD and Team NLSTR 2011. Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine 365, 395–409.

Amory JR, Barker ZE, Wright JL, Mason SA, Blowey RW and Green LE 2008. Associations between sole ulcer, white line disease and digital dermatitis and the milk yield of 1824 dairy cows on 30 dairy cow farms in England and Wales from February 2003-November 2004. Preventive Veterinary Medicine 83, 381–391.

Ballard RA, Truog WE, Cnaan A, Martin RJ, Ballard PL, Merrill JD, Walsh MC, Durand DJ, Mayock DE, Eichenwald EC, Null DR, Hudak ML, Puri AR, Golombek SG, Courtney SE, Stewart DL, Welty SE, Phibbs RH, Hibbs AM, Luan XQ, Wadlinger SR, Asselin JM and Coburn CE 2006. Inhaled nitric oxide in preterm infants undergoing mechanical ventilation. New England Journal of Medicine 355, 343–353.

Bang YJ, Kim YW, Yang HK, Chung HC, Park YK, Lee KH, Lee KW, Kim YH, Noh SI, Cho JY, Mok YJ, Kim YH, Ji JF, Yeh TS, Button P, Sirzen F, Noh SH and CLASSIC trial Investigators 2012. Adjuvant capecitabine and oxaliplatin for gastric cancer after D2 gastrectomy (CLASSIC): a phase 3 open-label, randomised controlled trial. Lancet 379, 315–321.

Barker ZE, Leach KA, Whay HR, Bell NJ and Main DCJ 2010. Assessment of lameness prevalence and associated risk factors in dairy herds in England and Wales. Journal of Dairy Science 93, 932–941.

Booth CJ, Warnick LD, Grohn YT, Maizon DO, Guard CL and Janssen D 2004. Effect of lameness on culling in dairy cows. Journal of Dairy Science 87, 4115–4122.

Bussel JB, Cheng G, Saleh MN, Psaila B, Kovaleva L, Meddeb B, Kloczko J, Hassani H, Mayer B, Stone NL, Arning M, Provan D and Jenkins JM 2007. Eltrombopag for the treatment of chronic idiopathic thrombocytopenic purpura. New England Journal of Medicine 357, 2237–2247.

Chawg 2013. Dairy cow welfare strategy, year 1 progress against targets. Cattle Health and Welfare Group, Malmesbury, UK.

Clarkson MJ, Downham DY, Faull WB, Hughes JW, Manson FJ, Merritt JB, Murray RD, Russell WB, Sutherst JE and Ward WR 1996. Incidence and prevalence of lameness in dairy cattle. Veterinary Record 138, 563–567.

DairyCo 2013. Dairy statistics an insiders guide 2013. The Agriculture and Horticulture Development Board, Kenilworth, UK.

D'Eath RB 2012. Repeated locomotion scoring of a sow herd to measure lameness: consistency over time, the effect of sow characteristics and inter-observer reliability. Animal Welfare 21, 219–231.

Dember LM, Beck GJ, Allon M, Delmez JA, Dixon BS, Greenberg A, Himmelfarb J, Vazquez MA, Gassman JJ, Greene T, Radeva MK, Braden GL, Ikizler TA, Rocco MV, Davidson IJ, Kaufman JS, Meyers CM, Kusek JW, Feldman HI and Dialysis Access Consortium Study Group 2008. Effect of clopidogrel on early failure of arteriovenous fistulas for hemodialysis: A randomized controlled trial. Journal of the American Medical Association 299, 2164–2171.

FAWC 2009. Farm animal welfare in Great Britain: past, present and future. Farm Animal Welfare Council, London, UK.

Garbarino EJ, Hernandez JA, Shearer JK, Risco CA and Thatcher WW 2004. Effect of lameness on ovarian activity in postpartum Holstein cows. Journal of Dairy Science 87, 4123–4131.

Heath CAE, Lin Y, Mullan S, Browne WJ and Main DCJ 2014. Implementing Welfare Quality® in UK assurance schemes: evaluating the challenges. Animal Welfare 23, 95–107.

Hoffman AC, Moore DA, Wenz JR and Vanegas J 2013. Comparison of modeled sampling strategies for estimation of dairy herd lameness prevalence and cow-level variables associated with lameness. Journal of Dairy Science 96, 5746–5755.

Kossaibati MA and Esslemont RJ 1997. The costs of production diseases in dairy herds in England. Veterinary Journal 154, 41–51.

Maertens W, Vangeyte J, Baert J, Jantuan A, Mertens KC, De Campeneere S, Pluk A, Opsomer G, Van Weyenberg S and Van Nuffel A 2011. Development of a real time cow gait tracking and analysing tool to assess lameness using a pressure sensitive walkway: the GAITWISE system. Biosystems Engineering 110, 29–39.

Main DCJ, Barker ZE, Leach KA, Bell NJ, Whay HR and Browne WJ 2010. Sampling strategies for monitoring lameness in dairy cattle. Journal of Dairy Science 93, 1970–1978.

Main DCJ, Mullan S, Atkinson C, Bond A, Cooper M, Fraser A and Browne WJ 2012. Welfare outcomes assessment in laying hen farm assurance schemes. Animal Welfare 21, 389–396.

Manson FJ and Leaver JD 1988. The influence of dietary protein intake and of hoof trimming on lameness in dairy cattle. Animal production 47, 191–199.

Rushen J, Pombourcq E and de Passille AM 2007. Validation of two measures of lameness in dairy cows. Applied Animal Behaviour Science 106, 173–177.

Rutherford KMD, Langford FM, Jack MC, Sherwood L, Lawrence AB and Haskell MJ 2009. Lameness prevalence and risk factors in organic and non-organic dairy herds in the United Kingdom. Veterinary Journal 180, 95–105.

Sorensen JT, Rousing T, Moller SH, Bonde M and Hegelund L 2007. On-farm welfare assessment systems: what are the recording costs? Animal Welfare 16, 237–239.

Webster AJF, Main DCJ and Whay HR 2004. Welfare assessment: indices from clinical observation. Animal Welfare 13, s93–s98.

Welfare Quality® 2009. Welfare Quality® assessment protocol for cattle. Welfare Quality® Consortium, Lelystad, The Netherlands.

https://doi.org/10.1017/S1751731115001536 Published online by Cambridge University Press