Accepted Manuscript

Title: Inter-observer agreement for clinical examinations of foot lesions of sheep

Author: C.J. Phythian, P.C. Cripps, D. Grove-White, E. Michalopoulou, J.S. Duncan

PII: S1090-0233(16)30119-8
DOI: http://dx.doi.org/doi: 10.1016/j.tvjl.2016.08.005
Reference: YTVJL 4865

To appear in: *The Veterinary Journal*

Accepted date: 12-8-2016

Inter-observer agreement for clinical examinations of foot lesions of sheep

Phythian, C.J.[ac*], Cripps, P.C.[a] Grove-White, D.[a] Michalopoulou, E.[a] and Duncan, J.S.[a]

[a] *Department of Epidemiology, University of Liverpool, Institute of Global Health and Infection, and Population Health, Leahurst, Neston, CH64 7TE, United Kingdom*
[c] *NMBU Sandnes, Section for Small Ruminant Research, Department of Production Animal Clinical Science, Høyland, 4325 Sandnes, Norway*

*Corresponding author: Tel.: +47 51603510
Email address: clare.phythian@nmbu.no (C.J.Phythian)

Current address: *NMBU, Section for Small Ruminant Research, 4325 Sandnes, Norway*

Highlights

- Diagnostic evaluation of clinical examinations of ovine foot lesions.
- Inter-observer agreement and percentage disagreement were assessed.
- Some scoring disagreement occurred over the diagnosis of white line lesions.
- Trained observers can reliably score most common foot lesions of sheep.

Abstract

In sheep, the diagnosis of foot lesions is routinely based on physical examination of the hoof. Correct diagnosis is important for the effective treatment, prevention and control of both infectious and non-infectious causes of lameness. Therefore, the aim of this study was to evaluate the level of inter-observer agreement for clinical examination of ovine foot lesions. Eight observers of varying experience, training and occupation performed foot examinations on a total of 1158 sheep from 38 farms across North England and Wales. On each farm, a group of two to four observers independently examined a sample of 24 to 30 sheep to diagnose the presence or absence of specific foot lesions including white line lesions (WL), contagious ovine digital dermatitis (CODD), footrot (FR), inter-digital dermatitis (ID) and toe granuloma (TG). The inter-observer agreement of foot lesion assessments was examined using Fleiss kappa (*κ*), and Cohen's *κ* examined the paired agreement between the test standard observer (TSO) and each observer.

Scoring differences with the TSO were examined as the percentage of scoring errors and assessed for evidence of systematic scoring bias. With the exception of WL (maximum error rate 33.3%), few scoring differences with the TSO occurred (maximum error rate 3.3%). This suggests that observers can achieve good levels of reliability when diagnosing most of the commonly observed foot conditions associated with lameness in sheep.

Keywords: Foot lesions; Sheep; Clinical diagnosis; Observer agreement.

Introduction

Lameness is a significant and serious global issue for sheep because of the pain, discomfort and debilitation caused (Welsh et al., 1993; Ley et al., 1995; Fitzpatrick et al., 2006). Research has identified that, globally, footrot is the most common cause of lameness in sheep (Egerton et al., 1989; König et al., 2011). Consequently, a variety of strategies for control and elimination of footrot have been devised. These include control approaches based on the administration of systemic antibiotic treatments and culling of persistently-infected cases (Wassink et al., 2010), and elimination strategies based on prophylactic vaccination and whole-flock culling programs (Egerton et al., 2002; Egerton et al., 2004; Gurung et al., 2006) .

Whilst footrot may be a common cause of lameness (Kaler and Green, 2008a), clearly not all lameness in sheep can be attributed to the condition. Contagious ovine digital dermatitis (CODD), which results in severe lameness and loss of the hoof capsule, currently presents a serious welfare concern for sheep in the UK (Winter, 2008). To date, there is limited knowledge on the epidemiology of this disease and by comparison to footrot only a few recent trials have examined the efficacy of systemic treatments (Duncan et al., 2011, 2012). In addition, there are a number of other foot conditions, including separation and impaction of the white line of the hoof, toe granulomas, interdigital-hyperplasia, septic- and osteo-arthritis, which can also result in gait abnormalities of sheep (Winter, 2004; Winter and Arsenos, 2009; Hodgkinson, 2011). Whilst infectious foot lesions remain the most important concern for flock welfare, it has been suggested that these other hoof lesions, such as separation and impaction of the white line (also known colloquially as 'shelly hoof'), are underreported due to mis-diagnosis and confusion with footrot cases (Conington et al., 2010a). This is of great importance since the treatment and control points that are deemed to be effective for one foot condition may not be relevant or appropriate for the control of another lesion or infectious cause. The correct identification of a lesion or disease is essential not only for both animal welfare reasons but also economic considerations in order to assess both the scale and economic impact of the disease. Hence, the ability to correctly diagnose foot lesions is vital for implementing prompt and effective treatments and the long-term prevention and control of lameness in sheep flocks (Kaler and Green, 2008a; Kaler and Green 2008b).

The ease and accuracy of using diagnoses based on the clinical appearance of lesions needs to be further considered given that there is considerable variation in the visual appearance of ovine foot lesions (Kaler and Green, 2008a). Furthermore, there are recognised differences in the interpretation and assessment of different foot lesions amongst differing assessors, such as veterinary surgeons, farmers and researchers (Kaler and Green, 2008b). Microbiological culture (Pitman et al., 1994) and PCR-based techniques (Moore et al., 2005; Frosth et al., 2012) can be employed to complement clinical examination in the diagnosis of some hoof pathologies. However, the time and financial cost of such methods preclude their routine use. Thus, clinical examination by the producer or a veterinary surgeon remains the mainstay for diagnosis of foot conditions in sheep. Consequently, the practical experience and training of farm professionals and veterinarians in the recognition and correct diagnosis of common foot lesions of sheep is an area that warrants further attention.

The diagnostic abilities of different observers can be examined in terms of the level of inter-observer agreement or reliability. The reliability of both binary and categorical scoring measures can be evaluated using agreement analysis methods such as the kappa coefficient (κ) (Kaler et al., 2009). The agreement analysis presents the degree of observed agreement compared to the agreement expected by chance (Feinstein and Cicchetti, 1990) and has been widely used in veterinary research applications, for example to assess the observer reliability for equine health and welfare indicator assessments (Burn et al., 2009) or lameness scores of sheep (Kaler et al., 2009). The type of κ selected depends on the number of observers involved. Fleiss's κ determines the reliability of multiple observers ($n > 2$) (Fleiss, 1981), whereas Cohen's κ (Cohen, 1960) examines the reliability of paired assessments ($n = 2$) such as the level of agreement between a study observer and a reference observer, such as the trainer (Burn et al., 2009). κ can also be used to assess the level of agreement between each study observer and a reference observer, such as the trainer (Burn et al., 2009). Several categorical systems for scoring ovine hoof health conditions, and specifically footrot, have been developed and tested (Egerton and Roberts, 1971; Raadsma et al., 1994; Conington et al., 2008; Foddai et al., 2012). However, for routine on-farm assessments as conducted by producers and veterinarians it may not be necessary to use such detailed scoring systems since a binary scale (presence or absence) could provide sufficient information.

The objective of this study was to examine the level of inter-observer agreement for specific ovine foot lesion conditions, using κ agreement analysis statistics and percentage error rate results.

Materials and methods

Study population

The investigation was a cross-sectional study in which 38 farms, located in a 120 mile radius of the University of Liverpool, School of Veterinary Science, Leahurst were recruited through contact with their local veterinary practice. Once the informed consent of farmers was obtained, each farm was requested to gather a sample of approximately 100 sheep for assessment during July to November 2008. On the day of assessment, each sheep was then assigned a numeric identifier in the order they entered the assessment pen and on each farm 30 sheep were selected for examination using a pre-determined random number system.

Observer population

A pool of eight observers was recruited from the University of Liverpool, School of Veterinary Science comprising undergraduate veterinary and animal science students (n = 3) and veterinary surgeons (n = 5). Observers were classified as 'experienced' if they had undertaken clinical examinations and foot lesion diagnosis of sheep in the previous year (Table 1), those that did not meet these criteria were classified as inexperienced. On the basis of their experience and role in the design and conduct of the study, observer 1 was designated the 'test standard observer' (TSO) and used as the reference test for comparison. All observers were provided with a scoring definition for each lesion, which they were requested to familiarise themselves with together with example images of the specific lesions. In addition, observers classed as 'trained' (n = 5) attended a one-day on-farm training session at the University of Liverpool sheep farm in the diagnosis of foot lesions in sheep. The TSO performed assessments on all study farms and was accompanied at each assessment visit by one to two observers who performed independent clinical examinations of the same sheep on the same day.

Foot examination

Each observer independently performed a clinical examination of each foot of all sample animals as described by Hodgkinson (2010). The absence or presence of any foot lesion in each sheep was recorded. The following specific diagnoses were made based on the descriptions of Winter (2004): white line lesion (WL) - separation and detachment of the white line ('shelly hoof') with impaction or infection present: inter-digital dermatitis (ID) - a raw to white, moist hairless area, progressing to inflammation, infection and necrosis of the inter-digital skin: footrot (FR) - separation of the horn of the hoof, beginning at the junction of the skin and horn near the heel, through to invasion of the sole with separation of insensitive and sensitive laminae: contagious ovine digital dermatitis (CODD) - ulceration around the coronary band, with or without loosening of the claw through to the total loss of the hoof capsule and presence of a raw stump of sensitive laminae: toe granuloma (TG) - strawberry-like growth of proud flesh, which may be covered with loose horn: inter-digital hyperplasia (IH) - folds or protrusions of the skin of variable size located within the inter-digital cleft, and pedal joint sepsis (PJS) - presence of heat, swelling and hair loss above the coronary band, with or without discharging tracts of pus above the coronary band or interdigital cleft. No diagnosis was recorded if it was not possible to make a specific diagnosis based solely on the visual appearance of the foot. Each observer manually recorded their findings on pre-tested recording charts. Observers were not provided with any clinical or production information before each visit. During the visit, each study observer performed an independent clinical examination and observers did not disclose or discuss their foot scores at any stage. The study was approved by the University of Liverpool Ethics Committee (RETH000287).

Data analysis

Data was analysed using Minitab version 16 and Stata version 13 (StataCorp LP). The prevalence (percentage) and 95 percent confidence interval (95% CI) of each foot condition was determined as the total number of sheep observed by the TSO with each foot condition divided by the total number of sheep assessed.

The overall level of inter-observer reliability of multiple observer assessments ($n \geq 2$) was determined by Fleiss's κ (Fleiss et al., 2003). As Fleiss's κ analysis provides an overall agreement value and does not take account of observer characteristics i.e. 'experienced' versus 'inexperienced' assessors, the paired agreement between the TSO and each observer was estimated using Cohen's κ statistic (Cohen, 1960). All κ results were interpreted according to Fleiss et al., (2003), whereby values ≥ 0.75 suggested 'excellent', κ 0.40 - 0.75 indicated 'fair–good', and $\kappa \leq 0.40$ suggested 'poor' levels of agreement.

As the κ analytical approach cannot identify whether systematic scoring differences occur between pairs or groups of multiple observers, additional approaches were employed to assess the level of observer disagreement in terms of scoring divergence from the TSO. Firstly, scoring differences between the TSO and each observer (TSO score minus observer score) were graphically represented and visually examined for evidence of systematic scoring bias i.e. if an observer consistently scored one unit higher or lower than the TSO. For each observer, the total number of lesions diagnosed by the TSO during paired assessments was calculated and the number of paired scoring differences with the TSO divided by the total number of sheep examined was expressed as a percentage (percentage error rate). Secondly, the proportion of scoring differences with the TSO on each farm visit was plotted to assess if there was any effect of increasing experience of foot assessments on the amount of scoring disagreements. Observers were not provided with any clinical or production information before each visit. During the visit, each study observer performed an independent clinical examination and observers did not disclose or discuss scores at any stage.

Results

A total of 4632 feet from 1158 sheep were examined for the presence of specific foot lesions. From the pool of eight observers, a varying group of two to three observers, including the TSO, independently examined the feet of 24 to 30 animals on each farm. Data recorded by the TSO indicated that over half of the population (*n* = 610, 52.6%) was observed to have at least one recorded condition in one or more feet. The most frequently observed lesion was WL (49.1%) and few cases of FR, ID, TG and CODD were recorded (Table 2). No cases of pedal joint sepsis cases were identified and there were insufficient observations of interdigital hyperplasia recorded by observers 2 – 8 to permit evaluation of the reliability, error rates for these foot conditions.

Overall level of inter-observer reliability was interpreted as 'excellent' for assessments of CODD (κ 0.72, 95% CI 0.71 - 0.77) being 'fair-good' for WL (κ 0.47, 95% CI 0.46 - 0.47) and TG (κ 0.65 95% CI 0.46 - 0.85). Fleiss κ values for FR (κ 0.49, 0.35 - 0.63) and ID (κ 0.49, 95% CI 0.37 - 0.65) diagnoses were 'fair to good' but the lower 95% confidence intervals for these lesions indicated some 'poor' levels of κ agreement occurred (Table 3).

With the exception of WL assessments, the majority of paired assessments with the TSO showed a low level (≤ 3.3% error rate) of scoring disagreement (Table 4). Graphical representation of the frequency of scoring differences for the diagnosis of white line lesions suggested there were some systematic scoring differences in the diagnosis of WL by several observers (Fig. 1). The evaluation of the effect of time on agreement with the TSO was limited to three observers (observers 3, 4, 7) who each performed ten or more study visits. Other study observers performed insufficient visits to facilitate this evaluation. Graphical representation of the proportion of scoring differences between the TSO and observers 3, 4 and 7 suggested there was no effect of increasing number of farm visits on the level of scoring disagreements (Fig. 2).

Discussion

The objective of this study was to evaluate the performance of eight assessors of varying experience, training and occupation on their ability to agree on the diagnosis of a range of specific foot conditions of sheep based solely on clinical examination and by using a binary scoring system. For the purposes of this study, farmers provided a group of sheep for assessment from which 30 animals were randomly selected for foot lesion examination. Previous work by Foddai et al. (2012) found high levels of inter-observer reliability when three observers used a combination of video, photographic images and post-mortem feet specimens to score lesions and foot shape using an ordinal scoring system. Assessments of lesion images produce higher levels of observer agreement compared to assessments of cadaver foot specimens (Foddai et al., 2012) and may reflect the more controlled observational conditions provided by image-based studies or the selection of lesion images that illustrate clear, characteristic signs of specific disease. When comparing the reliability and diagnostic test results of different studies, as well as considering the type of scoring system used it is also important to evaluate the conditions for assessment. It is possible that the levels of observer agreement that are attainable during on-farm studies may be with vastly different from those of image-based studies given the less controlled observational conditions and the need for handling of animals, which may introduce observational errors (Foddai et al., 2012). Therefore, a key aspect of this study was to test diagnostic abilities under conditions that producers and veterinary surgeons, who routinely conduct ovine foot examinations, are exposed to. Whilst the reliability of footrot scoring systems has been previously examined (Conington et al., 2008; Foddai et al., 2012), to the best of our knowledge, this is the first study to examine the inter-observer agreement of eight observers who examined and scored a large number of feet (*n* = 4632) from sheep managed under differing farm production systems for eight specific lesions.

Overall, Fleiss kappa results indicate that acceptable levels of reliability were achieved for the combined FR and ID scores, CODD and TG. Limitations in the availability of methods of agreement analysis and issues with the use of κ for the evaluation of observer reliability are well-recognised (Feinstein and Cicchetti, 1990; Burn et al., 2009; Foddai., 2012). The strong influence of lesion prevalence on κ estimates can be a particular issue for reliability studies conducted under field conditions (Burn et al., 2009). Therefore, these results should be viewed in light of the low prevalence of certain foot conditions, since this can reduce the level of κ and subsequent interpretation of the degree of inter-observer agreement achieved (Feinstein and Cicchetti, 1990; Burn et al., 2009). Another limitation with κ is the inability to quantify the level of scoring disagreement. In addition, whilst Fleiss's κ gives a useful indication of agreement between multiple observers the method does not take account of any biases due to observer characteristics, for example, 'experienced' vs. 'trained' vs. inexperienced' that may arise in multiple observer combinations. Observer 1 was selected as a 'pseudo-gold standard' reference standard using the approach of Burn et al., (2009) in order to compare paired inter-observer agreement and scoring divergence with the trainer and to assess for evidence of systematic scoring bias (Bland and Altman, 1986). Since clinical examinations are subjective, in some cases, it is possible that some observer scoring divergence from their trainer could represent a closer approximation to the true (latent) foot condition. For example, the paired κ agreement with the TSO ranged from poor to good for assessment of ID and FR but few scoring errors (maximum error rate 3.3%) were found for both of these lesions. Here, the κ results for ID and FR are considered to reflect the low number of animals observed in the study and this likely affected the cross-tabulation of results, required for agreement analysis. Kappa values are generally provided on a scale of 0 to 1 but negative values do arise and indicate poorer agreement than that expected by chance alone (Cohen, 1960). A negative κ value arose in the 95% CI for the paired assessments of TSO and observer 7, which reflected the very low number of animals that were observed with footrot during the paired scoring sessions.

Other studies have examined the diagnostic abilities of other assessors including farmers and veterinarians. Direct comparison of reliability studies can be complicated by differences in the scoring systems used, selection of material used for assessment; ranging from photographic images, video clips, post-mortem specimens (Foddai et al., 2012) or live sheep (Conington et al., 2010b), and the context or conditions for assessment. Earlier research into the diagnostic abilities of sheep veterinarians and producers found that ≥ 94% of veterinarians correctly diagnosed ID, FR, CODD and TG (Kaler and Green, 2008a). By contrast, only 26% of sheep farmers could correctly diagnose the same lesions (Kaler and Green, 2008a). Reliability results from the present study appear to concur with results of Kaler and Green (2008a), suggesting that when present these infectious foot lesions can be readily identified by experienced and/or trained assessors.

The high prevalence of WL identified in this study is in contrast to other studies from Australia (Egerton et al., 1989), England (Grogono-Thomas and Johnston, 1997) and Sweden (König et al., 2011), which suggest that footrot is the most commonly diagnosed foot condition of lambs and sheep. Indeed, according to farmer reports, footrot is the most prevalence foot lesion identified in English sheep flocks with a reported within-flock prevalence of 10% (Kaler and Green, 2008a). However, the WL prevalence findings of the present on-farm study (49%) do concur with those of Conington et al. (2010b), who assessed foot health scores from 27 flocks across the UK. With the exception of the Texel breed, white line separation was the lesion found at highest (40%) prevalence (Conington et al., 2010b). In the present study, participating farms were a convenience sample selected according to farm type and consent thus the presented results cannot be interpreted as prevalence estimates. However, these farms were considered to be representative of commercial sheep farming systems in England and Wales and these results may highlight some interesting regional trends in sheep managed in these flocks. To the author's knowledge, these farms had not been involved in previous research or training on sheep lameness. Although, it is possible that the low level of FR and ID identified here may suggest that farms with good ovine footrot control programs were recruited. These findings may also reflect the management, environmental and climatic conditions at the time of assessment that resulted in few sheep being diagnosed with these infectious lesions. The vast majority of WL lesions observed in the current study were restricted to separation of the hoof without impaction and infection of the white line (Winter and Arsenos, 2009). These observations are in agreement with a single-flock trial, which identified a high prevalence of WL of relatively minor degree of separation and an absence of other foot lesions (Wheeler et al., 2013). It is possible that many sheep have a mild degree of white line separation, which may be considered clinically insignificant, or missed during routine foot inspections. Co-existing minor WL lesions might also not be recorded during inspections focused on the detection of other ovine foot lesions, which might explain the prevalence findings reported here. In spite of a high proportion of white line lesions, there does not appear to be a strong association with a high level of flock lameness (Phythian et al., 2013) as often occurs with footrot (Kaler et al., 2011). The significance of seemingly minor WL separation on foot health and sheep welfare is not fully understood. Whilst there is some genetic heritability to ovine white line degeneration (Conington et al., 2010a), currently the prevention and control of this condition, predisposition to other hoof diseases and subsequent flock lameness prevalence is unknown.

Interestingly, the data revealed systematic scoring differences consistently arose over WL assessments. The seemingly minor degree of separation of the white line and potential poor differentiation of mild WL lesions might explain the level of scoring disagreement found here. Although a large number of feet were examined during the course of the study, the varying number of farm visits conducted by all observers limited the ability of the study to fully assess the effect of training and experience on diagnostic performance and no conclusions can be reached in this respect. There are some trends in the data to suggest that following on-farm training, inexperienced and trained observers (observers 3, 4 and 7) did not become more reliable in WL diagnoses over the course of examining more than 1000 feet. However, the results are limited to observers 3, 4 and 7 since they were the only observers that undertook ten or more farm visits, which facilitated the evaluation of reliability over time. Further evaluation of the effect of experience gained over a longer period of time and assessing whether a re-calibration session is beneficial for inexperienced observers would be valuable here.

In addition to observer experience and training, scoring errors can also arise due to misclassification that may be associated with the type of scoring system used. Misclassification of disease errors may have arisen here over the diagnosis of inter-digital dermatitis and footrot due to splitting of the scoring system into two distinct categories. With footrot lesions the different clinical outcomes that can arise are due in part to the strain of *Dichelobacter nodosus* involved (Moore et al., 2005), host susceptibility and genetic resistance (Emery et al., 1984). No infectious disease model of inter-digital dermatitis has yet been demonstrated, although some consider ID to be a continuum of clinical signs of a single disease (virulent and benign footrot) (Egerton and Roberts, 1971). In the present study, simple binary scoring scales were used to score benign and virulent footrot separately. However, a simple presence and absence binary scoring system may clearly not accurately describe the continuum of disease signs observed in footrot cases. With further training, assessors could be trained to grade the severity of these footrot lesions using more detailed and categorical footrot scoring systems, such as those of Egerton and Roberts (1971), Raadsma et al. (1994), and Nieuwhof et al. (2008). This may be desirable for examining the effectiveness of different treatments or disease elimination program (Egerton et al., 2004).

Very few cases (n = 3) were recorded by the TSO with no diagnosis. These cases were considered to represent developing and early lesions that could not be defined as a specific condition based solely on the visual appearance of the foot. Therefore, in some instances, microbiological and molecular biological testing (Moore et al., 2005; Frosth et al., 2012) may be required to support clinical examinations. The high levels of inter-observer kappa agreement achieved for CODD may be attributed to the clear scoring definition and training provided, or the ease of recognising this foot condition in sheep. Further training of observers in a recently developed categorical CODD scoring system (Angell et al., 2015) could facilitate clinical trials and further research into this condition, which is recognised to be of increasing importance in UK flocks and presents serious concerns for sheep welfare (Duncan et al., 2011, 2012). Despite this, on-farm experiences suggest that outbreaks of CODD are frequently mis-diagnosed as footrot by producers who are unaware of this condition and/or the different physical features of the disease (personal observation, CJ Phythian). However, in this study observers who were previously unaware and unfamiliar with this foot lesion became competent at diagnosing CODD. Such findings could inform disease awareness campaigns and highlights the value and role of sheep veterinarians in the prompt diagnosis, treatment and control of flock lameness.

Conclusion

FR, ID, CODD and TG were consistently diagnosed by observers (maximum error rate 3.3%) while WL, the lesion most commonly recorded in this study, was missed or misdiagnosed by some observers (maximum error rate 33.3%). The consequences for researchers and veterinary practitioners may be that in spite of training and experience a degree of measurement error and scoring disagreement can occur when using clinical examinations to diagnose common foot lesions in sheep. This may result in under- or over-reporting of prevalence estimates of some foot lesions during field studies, which needs to be considered when assessing the treatment, control and prevention of lameness in sheep to ensure that the optimal plans and advice are targeted at the correct lesion(s). In addition to further training of assessors, in some cases, diagnoses based on visual inspections of ovine feet may need to be supplemented by other tools such as molecular diagnostic testing.

Conflict of interest statement

None of the authors has any financial or personal relationships that could inappropriately influence or bias the content of the paper.

Acknowledgements

The authors gratefully acknowledge the funding provided by Defra as part of the AW1025 project (Development of indicators for the on-farm assessment of sheep welfare), the support of the expert panel, study farms and on-farm assistance of Phillip Jones, Daniel Hughes, Stephen Beer and Rachel Wyllie. We also thank the two anonymous reviewers for their constructive comments on the earlier versions of this manuscript. Preliminary results were presented in part as an Abstract at the First European College of Small Ruminant Health Management Conference, Athens, Greece, 29-30 October 2011.

References

Angell, J.W., Bludell, R., Grove-White, D. W., Duncan, J.S., 2015. Clinical and radiographic features of contagious ovine digital dermatitis and a novel lesion grading system. Veterinary Record doi:10.1136/vr.102978.

Bland, J.M., Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1, 307-310.

Burn, C.C., Pritchard, J.C., Whay, H.R., 2009. Observer reliability for working equine welfare assessment: problems with high prevalences of certain results. Animal Welfare 18, 177-187.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological, Measurement 20, 37-46.

Conington, J., Hosie, B., Nieuwhof, G., Bishop, S., Bünger, L., 2008. Breeding for resistance to footrot – the use of hoof lesion scoring to quantify footrot in sheep. Veterinary Research Communications 32, 583-589.

Conington J., Nicoll, L., Mitchell, S., Bünger, L., 2010a. Characterisation of white line degeneration in sheep and evidence for genetic influences on its occurrence. Veterinary Research Communications 34, 481– 489.

Conington, J., Speijers, M.H.M., Carson, A., Johnston, S., Hanrahan, S., 2010b. Foot health in sheep – prevalence of hoof lesions in UK and Irish sheep. In: Proceedings of the British Society of Animal Science Annual Conference, Belfast, Northern Ireland.

Duncan, J.S., Grove-White, D., Phythian, C.J., Oultram, J., Van Dijk, J., Carter, S.D., Cripps, P.J., Williams, H., 2011., Effects of parenteral amoxicillin on recovery rates and new infection rates for contagious ovine digital dermatitis in sheep. Veterinary Record 169, 606.

Duncan, J.S., Grove-White, D., Moks, E., Oultram, J.W., Phythian, C.J., Williams, H.W., 2012. Impact of footrot vaccination and antibiotic therapy on footrot and contagious ovine digital dermatitis: a split flock randomised treatment trial. Veterinary Record 70, 462.

Egerton, J.R., Roberts, D.S., 1971. Vaccination against ovine foot-rot. Journal of Comparative Pathology 8, 179-185.

Egerton, J.R., Yong, W.K., Riffkin, G.G., 1989. Footrot and foot abscess in ruminants. CRC Press Inc, Florida, USA.

Egerton, J.R., Ghimire, S.C., Dhungyel, O.P., Shrestha, H.K., Joshi, H.D., Joshi, B.R., Abbott, K.A., Kristo, C., 2002. Eradication of virulent footrot from sheep and goats in an endemic area of Nepal and an evaluation of specific vaccination. Veterinary Record 151, 290-295.

Egerton, J.R., Seaman, J.T., Walker, R.I., 2004. Eradication of virulent footrot from New South Wales. In: Proceedings of the 13th Intentional Symposium and 5th Conference on Lameness in Ruminants, 11-15th February 2004, Maribor, Slovenija, 215-221.

Emery, D.L, Stewart, D.J., Clark BL., 1984. The comparative susceptibility of five breeds of sheep to foot-rot. Australian Veterinary Journal 61, 85-88.

Feinstein, A.R., Cicchetti, D.V., 1990. High agreement but low kappa .1. The problems of 2 paradoxes. Journal of Clinical Epidemiology 43, 543-549.

Fitzpatrick, J., Scott, M., Nolan, A., 2006. Assessment of pain and welfare in sheep. Small Ruminant Research 62, 55-61.

Fleiss, J. L., Levin, B. and Paik, M. C. 2003. The Measurement of Interrater Agreement, In: Statistical Methods for Rates and Proportions, Third Edition, John Wiley & Sons, Inc., Hoboken, New Jersey, USA.

Foddai, A., Green, L.E., Mason, S.A., Kaler, J., 2012. Evaluating observer agreement of scoring systems for foot integrity and footrot lesions in sheep. BMC Veterinary Research 8, 65.

Frosth, S., Slettemeås, J.S., Jørgensen, H.J., Angen, O., Aspán A., 2012. Development and comparison of a real-time PCR assay for detection of *Dichelobacter nodosus* with culturing and conventional PCR: harmonisation between three laboratories. Acta Veterinaria Scandinavica 54, 6.

Grogono-Thomas, R. & Johnston, A. M. 1997. A study of ovine lameness. MAFF Final Report. MAFF open contract OC59 45K.

Gurung, R.B., Dhungyel, O.P., Tshering, P., Egerton, J.R., 2006.The use of an autogenous *Dichelobacter nodosus* vaccine to eliminate clinical signs of virulent footrot in a sheep flock in Bhutan Veterinary Journal 172, 356-363.

Hodgkinson, O., 2010. The importance of feet examination in sheep health management. Small Ruminant Research 92, 67-71.

Kaler, J., Green, L.E., 2008a. Naming and recognition of six foot lesions of sheep using written and pictorial information: a study of 809 English sheep farmers. Preventive Veterinary Medicine 83, 52-64.

Kaler, J., Green, L.E., 2008b. Recognition of lameness and decisions to catch for inspection among sheep farmers and specialists in GB. BMC Veterinary Research 4, 41.

Kaler, J., Wassink, G.J., Green, L.E., 2009. The inter- and intra-observer reliability of a locomotion scoring scale for sheep. The Veterinary Journal 180, 189-194.

Kaler, J., George, T.R.N., Green, L.E. 2011. Why are sheep lame? Temporal associations between severity of foot lesions and severity of lameness in 60 sheep. Animal Welfare 20, 433-438.

König, U., Nyman, A.K., de Verdier K., 2011. Prevalence of footrot in Swedish slaughter lambs. Acta Veterinaria Scandinavica. 53, 27.

Ley, S.J., Waterman, A.E., Livingston, A., 1995. A field study of the effect of lameness on mechanical nociceptive thresholds in sheep. Veterinary Record 137, 85-87.

Moore, L.J., Wassink, G.J., Green, L.E., Grogono-Thomas, R., 2005. The detection and characterisation of Dichelobacter nodosus from cases of ovine footrot in England and Wales Veterinary Microbiology 108, 57-67.

Nieuwhof, G.J., Conington, J., Bünger, L., Haresign, W., Bishop, S.C., 2008. Genetic and phenotypic aspects of foot lesion scores in sheep of different breeds and ages. Animal 2, 1289-1296.

Pitman, D.R., Palmer, M.A., Depiazzi, LJ. 1994. The laboratory culture of *Dichelobacter nodosus* in a footrot eradication program. Australian Veterinary Journal 71, 109-112.

Phythian, C.J., Cripps, P.C., Grove-White, D., Jones, P.H., Michalopoulou, E., Duncan, J.S. 2013. Observing lame sheep: evaluating test agreement between group-level and individual animal methods of assessment. Animal welfare 22, 417-422.

Raadsma, H.W., Egerton, J.R., Wood, D., Kristo, C., Nicholas, F.W., 1994. Disease resistance in Merino sheep. III. Genetic variation in resistance to footrot following challenge and subsequent vaccination with a homologous rDNA pilus vaccine under both induced and natural conditions. Journal of Animal Breeding and Genetics 111, 367–390.

Wassink, G.J., King, E.M., Grogono-Thomas, R., Brown, J.C., Moore, L.J., Green L.E., 2010. A within farm clinical trial to compare two treatments (parenteral antibacterials and hoof trimming) for sheep lame with footrot. Preventive Veterinary Medicine 96. 93-103.

Welsh, E.M., Gettinby, G., Nolan, A.M., 1993. Comparison of a visual analogue scale and a numerical rating scale for assessment of lameness, using sheep as a model. American Journal of Veterinary Research 54, 976-983.

Winter, A., 2004. Lameness in sheep: 1. Diagnosis. In Practice 26, 58-63.

Winter, A.C., 2008. Lameness in sheep. Small Ruminant Research 76, 149-153.

Winter, A.C., Arsenos, G., 2009. Diagnosis of white line lesions in sheep. In Practice 31, 17–21.

Wheeler, K., Phythian, C.J., Phillips, K., 2013. Effect of oral biotin supplementation on the observed prevalence and severity of white line lesions in a lowland flock. In: Proceedings of the 13th International Symposium and 9th Conference on Lameness in Ruminants, 11-14th August 2013, Bristol, UK.

Figure legends

Fig. 1. Frequency of scoring differences between the paired foot examinations of the test standard observer (TSO) and each study observer (2 - 8) for diagnosis of white line (WL).

Fig. 2. Frequency of scoring differences between paired examinations of trained and inexperienced observers 3, 4 and 7 with the test standard observer (TSO) for white line lesion (WL) diagnosis.
Data is presented only for observers who conducted foot examinations on ten or more farms to permit evaluation of the effect of time on observer reliability.

Table 1. Description of the observer population

Observer	Training	Experience	Occupation
1	Trainer	Experienced	Veterinary surgeon
2	Trained	Inexperienced	Veterinary surgeon
3	Trained	Inexperienced	Animal science student
4	Trained	Inexperienced	Veterinary science student
5	Untrained	Inexperienced	Veterinary surgeon
6	Untrained	Experienced	Veterinary surgeon
7	Trained	Inexperienced	Animal science student
8	Trained	Experienced	Veterinary surgeon

Table 2. Total number and percentage of sheep (n = 1158) observed with each foot lesion by the Test Standard Observer (TSO).

Diagnosis	Total *n* observed	Percentage (%) observed (95% CI)
White line lesion	569	49.1 (46.3 – 52.0)
Inter-digital dermatitis (ID)	11	0.9 (0.4 – 1.5)
Footrot	14	1.0 (0.6 – 1.8)
Contagious ovine digital dermatitis	16	1.4 (0.7 – 2.1)
Toe granuloma	16	1.4 (0.7 – 2.1)
Interdigital hyperplasia	5	0.4 (0.1 – 0.8)
No diagnosis	3	0.3 (0 – 0.6)

Table 3. Inter-observer agreement (Fleiss's kappa, 95% confidence interval), and paired agreement between the test standard observer and observers 2 – 8 (Cohen's kappa, 95% confidence interval) for diagnoses of specific ovine foot lesions

Diagnosis	Fleiss κ (95% CI)	Cohen's κ (95 % CI) by observer						
		2	3	4	5	6	7	8
WL	0.47 (0.46 - 0.47)	0.63 (0.35 - 0.90)	0.42 (0.36 - 0.48)	0.46 (0.40 - 0.52)	0.28 (0.04 - 0.52)	0.34 (0.02 - 0.67)	0.53 (0.43 - 0.64)	0.70 (0.52 - 0.88)
ID	0.49 (0.35 - 0.63)	[a]	0.25 (0.02 - 0.49)	0.73 (0.51 - 0.96)	0.78 (0.49 - 1)	[a]	0.67 (0.05 - 1)	[a]
FR	0.49 (0.37 - 0.65)	[a]	0.58 (0.34 - 0.82)	0.55 (0.27 - 0.83)	[a]	[a]	0.40 (-0.15 - 0.94)	[a]
CODD	0.72 (0.71 - 0.77)	1 (0.99 - 1)	0.55 (0.27 - 0.83)	0.75 (0.51 - 0.99)	1 (0.99 - 1)	[a]	0.68 (0.47 - 0.89)	1 (0.99 - 1)
TG	0.65 (0.46 – 0.86)	[a]	0.71 (0.44 - 0.98)	0.57 (0.26 - 0.88)	0.65 (0.20 - 1)	[a]	0.56 (0.25 - 0.88)	1 (0.99 - 1)

[a] Insufficient observations of foot condition for kappa analysis
WL, white line lesion; ID, inter-digital dermatitis; FR, footrot; CODD, contagious ovine digital dermatitis; TG, toe granuloma

Table 4. Observer error rate expressed as the percentage (%) of scoring differences between the paired examinations with the test standard observer (TSO) and the total number (*n*) of lesions diagnosed by the TSO for each of the observer paired examinations

		Observer						
		2	3	4	5	6	7	8
Total *n* sheep examined		86	907	771	60	30	270	60
WL	percentage error	6.9	28.8	27.4	31.7	33.3	21.9	15.0
	n by TSO	9	391	372	37	12	175	9
ID	percentage error	1.2	1.9	0.7	3.3	0	0.4	0
	n by TSO	1	7	9	4	0	1	0
FR	percentage error	1.2	1.1	1.0	1.7	3.3	1.1	0
	n by TSO	1	11	10	0	0	3	0
CODD	percentage error	0	0.9	0.5	0	0	2.9	0
	n by TSO	0	7	6	1	0	9	3
TG	percentage error	1.1	0.3	0.4	0	0	0	0
	n by TSO	2	3	5	1	0	2	1

WL, white line lesion; ID, inter-digital dermatitis; FR, footrot; CODD, contagious ovine digital dermatitis; TG, toe granuloma.