J. Dairy Sci. 95:206–212
doi:10.3168/jds.2011-4603

Evaluation of vaginoscopy for the diagnosis of clinical endometritis in dairy cows

C. Leutert, X. von Krueger, J. Plöntzke, and W. Heuwieser[1]
Clinic of Animal Reproduction, Faculty of Veterinary Medicine, Freie Universität Berlin, Königsweg 65, 14163 Berlin, Germany

ABSTRACT

The objective of the study was to evaluate the visual assessment of vaginal discharge by vaginoscopy for the diagnosis of clinical endometritis (CE) in dairy cows. In an in vivo trial, inter- and intraobserver repeatability of vaginoscopic examination (VE) was determined and the effect of transrectal palpation and experience of the investigator evaluated. Holstein-Friesian cows (n = 380) were examined by vaginoscopy between 21 and 27 d in milk by 3 investigators twice. Vaginal discharge was categorized on a 4-point classification system (0 = clear mucus, 1 = mucus containing flecks of pus, 2 = discharge containing less than 50% pus, 3 = discharge containing more than 50% pus). Cows with a vaginal discharge score (VDS) of 0 were classified as healthy, whereas cows with a VDS of 1 to 3 were classified as having CE. Vaginal discharge score on a scale from 0 to 3 has moderate intra- (Cohen's kappa coefficient, κ = 0.55–0.60) and interobserver (κ = 0.44) repeatability. The prevalence of CE was comparable between the 3 investigators (first VE: 42.6, 34.8, and 38.7; second VE 46.8, 36.9, and 43.7%). Transrectal palpation (relative risk = 0.96–1.03) or experience of the investigator (relative risk = 0.9–1.1) did not affect results of VE. In an in vitro trial, sensitivity and specificity of visual assessment were determined utilizing 33 images showing yellow and pink areas in certain percentages as a reference standard. Pus was represented by yellow areas and the mucosa, including clear mucus, by pink areas. These images were visually assessed by 30 investigators via PowerPoint presentation (experiment 1) and by 23 investigators via a simulated vaginal examination (experiment 2) utilizing the same 4-point classification system. Sensitivity was 99.6 and 96.3% and specificity was 96.7 and 90.1% in experiments 1 and 2, respectively. The results provide evidence that a visual assessment conducted by vaginoscopic examination is not perfect but can be considered a reasonable measurement of vaginal discharge and is a practical tool to distinguish healthy from diseased cows.

Key words: clinical endometritis, vaginoscopy, diagnosis, test characteristic

INTRODUCTION

Clinical endometritis (**CE**) is an important disease in dairy cows and is defined as an inflammation of the endometrium occurring later than 21 DIM. Two recent studies established a scientifically sound and clinically useful case definition of CE based on factors that were prognostic for impaired reproductive performance (LeBlanc et al., 2002; Sheldon et al., 2006). Mucopurulent vaginal discharge and a cervix diameter >7.5 cm were the only clinical findings with predictive value for decreased fertility (LeBlanc et al., 2002). Although in practice transrectal palpation of the uterus was the predominant method used by veterinarians to diagnose uterine diseases, several studies have demonstrated that this method results in a large number of false-positive diagnoses (LeBlanc, 2003). Several diagnostic methods have been described to examine vaginal discharge such as the gloved hand, the Metricheck device (Metricheck, Simcro, New Zealand), and vaginoscopy (Sheldon et al., 2002; McDougall et al., 2007; Pleticha et al., 2009). A common method used to diagnose vaginal discharge is the vaginal examination with a vaginoscope (LeBlanc et al., 2002; Sheldon et al., 2006). A 4-point scoring system (0 = clear mucus, 1 = mucus containing flecks of pus, 2 = discharge containing less than 50% pus, 3 = discharge containing more than 50% pus) to classify vaginal mucus was established by Williams et al. (2005) and has been used in recent studies (Sheldon et al., 2006; Dubuc et al., 2010; Kaufmann et al., 2010).

An important obstacle in validating different diagnostic methods and describing test characteristics is the lack of a gold standard to verify inflammation of the uterus (Drillich et al., 2007). In addition, the presence of mucopurulent or worse vaginal discharge may not be reflective of endometrial inflammation (Dubuc et. al., 2010). Cytological examination is the most definitive diagnosis for CE and enables the differentiation of CE from vaginitis or cervicitis. However, this diagnostic

Received June 6, 2011.
Accepted September 16, 2011.
[1]Corresponding author: w.heuwieser@fu-berlin.de

method is time consuming and expensive for routine use (Sheldon et al., 2006). In recent studies, cytological examination has been used to calculate sensitivity and specificity of vaginoscopy, endometrial cytology, and ultrasonography for the diagnosis of CE (Drillich et al., 2007; Barlund et al., 2008; Westermann et al., 2010).

Recent publications reported the sensitivity and specificity of vaginoscopic examination (**VE**) using the pregnancy status at 150 DIM or cytological results as the reference method (LeBlanc et al., 2002; Barlund et al., 2008). Sensitivity and specificity of VE compared with cytological findings calculated by Barlund et al. (2008) were 53.9 and 95.4%, respectively. When comparing results of the vaginal examination with pregnancy status at 150 DIM, sensitivity and specificity were 20 and 88% (LeBlanc et al., 2002) and 7.1 and 87.4% (Barlund et al., 2008), respectively. The comparison of VE findings and pregnancy status to calculate diagnostic accuracy may be regarded with some doubt due to numerous variables influencing the reproductive performance of a cow (Kasimanickam et al., 2004).

Science-based evidence, both from accepted clinical (e.g., rectal palpation) and advanced diagnostic methods (e.g., radiography, ultrasound), suggests that the investigator is a relevant source of measurement error (Kelton et al., 1991; Schneider et al., 2002; Andermann et al., 2007). Even though the vaginal examination is the most commonly used diagnostic method (LeBlanc et al., 2002), information regarding the reliability of this approach is lacking. Data describing inter- and intraobserver repeatability are not available. Therefore, the objective of this study was to evaluate the repeatability of vaginoscopy. Specifically, we set out to (1) determine repeatability (inter- and intraobserver) of scoring vaginal discharge using a vaginoscope and a 4-point classification system, (2) test the influence of transrectal manipulation of the uterus and the level of experience of the investigator on a vaginal examination, and (3) study the sensitivity and specificity of the human capability to visually assess color shades by means of 2 in vitro experiments.

MATERIALS AND METHODS

In Vivo Trial

The study began with an in vivo trial conducted on a commercial dairy farm in Brandenburg, Germany, between August 2009 and June 2010. The herd size was 750 lactating cows milked 3 times a day. Animals were housed in freestall facilities with cubicles, rubber mats, and slotted floors year round. Cows were grouped in pens holding approximately 100 cows depending on lactation and reproduction status. Calving pens were straw-bedded. Average milk yield was 10.050 kg per lactation and cow (fat 4.3%, protein 3.1%). A TMR was fed consisting of 39.7% concentrate and mineral mix, 32.9% grass silage, and 19.2% corn silage (NE_L = 1.65 Mcal/kg). Before the study, an informed consent was obtained from the owner.

Every week 10 ± 2 cows between 21 and 27 DIM were selected with a random treatment allocation plan generated with PASW (PASW statistics 18.0, SPSS Inc., Munich, Germany) and enrolled. Three investigators (**Inv**) examined the cows independently. In total, 386 cows were examined by Inv 1 and 3 and 339 cows by Inv 2. Investigator 1 (author: C. Leutert) and 2 (co-author: X. von Krueger) were the same veterinarians for the whole study period and had been trained in VE before the study. "Investigator 3" comprised 41 veterinary students who changed from week to week and had marginal experience in VE. Their experience in VE (i.e., fewer than 20 cows) was surveyed by means of a questionnaire before the herd visit.

The vulva was cleaned with a dry paper towel. The vaginoscopes (Hauptner and Herberholz, Solingen, Germany; length: 30 cm, diameter: 2.8 cm) were single packed and autoclaved. They were unwrapped, moistened with 0.9% sodium chloride solution, and inserted into the vagina up to the outer cervical os. Cervix and vagina were visually examined for presence and quality of discharge with the help of a flashlight. All 3 investigators obtained their results through the same vaginoscope using the same flashlight. For a given animal, all 3 visual assessments were conducted within 30 s or less. To ensure independent results, the investigators made their observations in the absence of the other investigators and documented their findings separately on case report forms. On average, 4 cows were examined in one batch. All cows were re-examined by means of a vaginoscope after approximately 10 min following the same examination protocol. Half of the animals were randomly selected, utilizing a random treatment allocation plan generated with PASW, and examined by transrectal palpation (**TP**; n = 191) of the uterus before the second VE, whereas the remaining cows were not examined and were used as a control group (n = 189). Investigator 1, who classified the cervical diameter, location and consistence of the uterus, and symmetry and diameter of the uterine horns exclusively performed the TP. Investigators 2 and 3 were absent during the TP conducted by Inv 1 and were unaware of the allocation. The perianal area of the cows was carefully cleaned to avoid any sign of TP conducted before the other investigators returned to perform the second VE. The second VE was conducted as described for the first VE in all cows. Vaginal discharge was categorized utilizing the 4-point classification system (0 = no or

clear mucus, 1 = mucus containing flecks, 2 = discharge containing less than 50% pus, 3 = discharge containing more than 50% pus) described by Williams et al. (2005). All cows diagnosed with a vaginal discharge score (**VDS**) 1 to 3 were treated with prostaglandin $F_{2\alpha}$ (twice within 2 wk) administered by Inv 1.

In Vitro Trial

An in vitro trial using a reference standard was conducted to enable the calculation of sensitivity and specificity for the 4-point classification system used in the in vivo trial. To create reference standards representing different percentages of pus within vaginal mucus according to the classification system, 33 images were designed with a computer program (Microsoft Paint 5.1, Microsoft Deutschland GmbH, Munich, Germany). The images showed yellow and pink areas in certain percentages. Pus was represented as yellow areas (red = 255, green = 255, blue = 128; color = 40; saturation = 240; intensity = 180) and the mucosa including clear mucus as pink areas (red = 255, green = 128, blue = 128; color = 0; saturation = 240; intensity = 180). A VDS of 0 was represented by images displaying exclusively a pink area, and VDS 1, 2, and 3 were represented by 4 to 10%, 11 to 50%, and >50% of yellow areas (i.e., pus) on the images, respectively. Of the 33 images, 2, 10, 10, 11 represented VDS 0, 1, 2, and 3, respectively.

The images were applied in 2 experiments. First, a PowerPoint presentation (Microsoft Office 2003, Microsoft Deutschland GmbH) was presented to one investigator at a time. In total, 30 investigators (14 final-year veterinary students and 16 licensed veterinarians) were enrolled. They classified the images using the 4-point classification system. The 33 images were shown in a random order for 3 s each. In the second experiment, the same images were printed in high-quality color and presented in a wooden case (30 cm × 23 cm × 18 cm). On the front side, a 2-mm slot allowed the images to be changed by sliding them in or out. On the opposite side of the wooden case, a vaginoscope was inserted through a hole (diameter = 3.5 cm). Each investigator had to assess 28 images (i.e., 7 for each score) using a vaginoscope and a flashlight. Both, vaginoscope and flashlight were identical with the instruments used in the in vivo trial. Again, images were presented in a random order for 5 s each. Twenty-three investigators were enrolled.

Ten investigators were enrolled in both experiments twice, with a time interval of approximately 4 wk. The same images were presented to the investigators twice in a different randomized order. Before the experiments, an informed consent was obtained from each investigator.

Statistical Methods

The analyses were performed with PASW Statistics for Windows (PASW statistics 18.0, SPSS Inc.) and with Excel (Microsoft Office 2003, Microsoft Deutschland GmbH).

The randomization for the selection of study animals was conducted with the random sample function of PASW. The overall prevalence of CE was calculated from the VE results of investigator 1 at first examination, regarding cows with score 0 as healthy and with score 1, 2, and 3 as affected with CE. Relative risks (**RR**) of CE were calculated including time of examination (0 = first examination, 1 = second examination), transrectal palpation (0 = no transrectal palpation, 1 = transrectal palpation), and experience of investigator, with investigator 3 as inexperienced (0 = experienced, 1 = inexperienced) as cofactors. Adjusted RR, confidence intervals, and *P*-values are reported. The CI were set at 95% and the level of significance was set at $\alpha = 0.05$.

The interobserver repeatability between the 3 investigators was calculated using the Fleiss κ test. Data were analyzed with PASW using an additional syntax obtained from http://www.spsstools.net/Syntax/Matrix/CohensKappa.txt. The intraobserver repeatability was calculated with Cohen's kappa test using "crosstabs, statistics" in PASW. The results of the kappa test were interpreted according to the classification κ <0.00 = poor, 0.00–0.20 = slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, and 0.81–1.00 = almost perfect agreement, created by Landis and Koch (1977).

The order of the images presented in the in vitro trial was randomized by using the random number function in Excel (Microsoft Deutschland GmbH). In the in vitro trial, sensitivity and specificity of the 4-point classification system were calculated using the percentage distributions of pus and mucus of the images as a reference standard. Sensitivity was calculated as the proportion of images correctly diagnosed as positive for CE by the visual assessment. Specificity was calculated as the proportion of images correctly diagnosed as negative by the visual assessment. The intra- and interobserver repeatability were calculated using Cohen's kappa test.

RESULTS

In total, 386 cows were examined during the in vivo trial. Six cows were excluded from analysis due to diseases affecting examination outcomes (4 with vaginal injuries and 2 with pyometra). Because of illness, 41 cows were not examined by investigator 2. In total, 339 paired examination outcomes were available for

analysis for all 3 investigators, and 380 paired observations were available for analysis for investigator 1 and 3, respectively.

Overall prevalence of CE was 42.6%. The prevalence of CE was similar at the first and second VE (42.6 vs. 46.8%). Considering the 3 investigators separately, the prevalence of CE was 42.6, 34.8, and 38.7% at the first VE and 46.8, 36.9, and 43.7% at the second VE for investigators 1, 2, and 3, respectively. Frequency distribution of VDS was similar at first and second VE and for the 3 investigators (Table 1). When VE outcomes were summarized to distinguish between diseased (scores 1, 2, and 3) and healthy (score 0) cows, agreement between first and second VE was 84.7, 84.4, and 83.9% for investigators 1, 2, and 3, respectively. The 3 investigators agreed in the diagnosis of diseased and healthy cows in 67.9% (first VE) and 65.5% (second VE), respectively. Defining score 0 and 1 as healthy, as suggested by LeBlanc et al. (2002), the VE outcomes by Inv 1, 2, and 3 agreed between the first and second VE in 92.4, 92.9, and 90.3%, respectively, and the 3 investigators agreed in 81.2% (first VE) and 78.6% (second VE), respectively.

The RR for the diagnosis of CE was similar at first and second VE (Inv 1: RR = 1.1, 95% CI: 0.94–1.29, $P > 0.05$; Inv 2: RR = 1.06, 95% CI: 0.87–1.3, $P > 0.05$; Inv 3: RR = 1.13, 95% CI: 0.95–1.34, $P > 0.05$)

Frequency distribution of VDS was similar for cows examined with and without TP considering the VDS of the first examination (Table 2). The RR for a cow to be diagnosed with CE by vaginoscopic examination was not affected by transrectal palpation (Inv 1: RR = 0.96, 95% CI: 0.77–1.19, $P > 0.05$; Inv 2: RR = 1.03, 95% CI: 0.78–1.36, $P > 0.05$; Inv 3: RR = 1.0, 95% CI: 0.8–1.26, $P > 0.05$).

Results of the experienced (Inv 1 and 2) and inexperienced investigators (Inv 3) were similar (Tables 1 and 2), and the investigator's experience did not influence the likelihood of CE (Inv 1 vs. Inv 3: RR = 1.1, 95% CI: 0.93–1.31, $P > 0.05$; Inv 2 vs. Inv 3: RR = 0.9, 95% CI: 0.74–1.09, $P > 0.05$).

The kappa test revealed an agreement of $\kappa = 0.55$ to 0.60 between the first and second VE for all 3 investigators (Inv 1: n = 380, $\kappa = 0.60$, $P < 0.001$; Inv 2: n = 339, $\kappa = 0.56$, $P < 0.001$; Inv 3: n = 380; $\kappa = 0.55$; $P < 0.001$). The κ coefficient for the interobserver repeatability for all 3 investigators was 0.44 (n = 339, $P < 0.001$).

In the first experiment of the in vitro trial, overall sensitivity and specificity (30 investigators) were 99.6 and 96.7% (n = 990), respectively, when VDS 0 was considered as healthy and VDS 1 to 3 as CE. Defining score 0 and 1 as healthy and score 2 and 3 as CE, the sensitivity and specificity decreased to 95.2 and 85.5% (n = 990), respectively.

Kappa statistics could not be calculated for 3 investigators for interobserver repeatability, as they did not use all scores (0 to 3) for classification of the images. Thus, interobserver repeatability was analyzed for 27 investigators. Overall, median interobserver repeatability was $\kappa = 0.55$ (n = 351). Classifying interobserver repeatability, 12, 124, 165, and 50 investigator comparisons were in the κ ranges of 0.81 to 0.90, 0.61 to 0.80, 0.41 to 0.60, and 0.21 to 0.40, respectively. Overall, median intraobserver repeatability considering 10 investigators was $\kappa = 0.82$. Vaginal discharge scores 0 to 3 were diagnosed correctly in 96.7, 83.0, 63.7, and 72.0% of the cases, respectively.

In the second experiment of the in vitro trial, sensitivity and specificity of 23 investigators were 96.3% and 90.1% (n = 644), respectively, when VDS 0 was considered as healthy and VDS 1 to 3 as CE. Defining score 0 and 1 as healthy and score 2 and 3 as CE, both sensitivity and specificity were 92.9% (n = 644).

Kappa statistics could not be calculated for 1 investigator for inter- and intraobserver repeatability, as he did not use all scores (0–3) for the classification of images. Thus, interobserver repeatability was analyzed for

Table 1. Frequency distribution of vaginal discharge scores in cows examined at 21 to 27 DIM by 3 independent investigators twice within 10 min

Investigator	Vaginal examination[1]	Number of cows	Vaginal discharge score,[2] % (n)			
			Score 0	Score 1	Score 2	Score 3
1	First	380	57.4 (218)	17.9 (68)	14.5 (55)	10.3 (39)
	Second	380	53.2 (202)	20.3 (77)	15.8 (60)	10.8 (41)
2	First	339	65.2 (221)	13.3 (45)	11.2 (38)	10.3 (35)
	Second	339	63.1 (214)	15.9 (54)	9.7 (33)	11.2 (38)
3	First	380	61.3 (233)	16.6 (63)	11.8 (45)	10.3 (39)
	Second	380	56.3 (214)	21.8 (83)	10.3 (39)	11.6 (44)

[1]Second vaginal examination was conducted within 10 min after the first examination.

[2]Vaginal discharge score: 0 = clear mucus, 1 = mucus containing flecks of pus, 2 = discharge containing less than 50% pus, 3 = discharge containing more than 50% pus.

Table 2. Frequency distribution of vaginal discharge scores (VDS) of the second vaginal examination considering the VDS of the first examination by 3 independent investigators (Inv) and the effect of transrectal palpation of the uterus

VDS at second exam[1]	With transrectal palpation, % (no.)			Without transrectal palpation, % (no.)		
	Inv 1	Inv 2	Inv 3	Inv 1	Inv 2	Inv 3
VDS 0 at first exam						
0	89.7 (96)	88.2 (97)	81.8 (99)	76.6 (85)	84.7 (94)	83.9 (94)
1	8.4 (9)	9.1 (10)	16.5 (20)	19.8 (22)	13.5 (15)	12.5 (14)
2	1.9 (2)	1.8 (2)	0.8 (1)	3.6 (4)	1.8 (2)	1.8 (2)
3	0	0.9 (1)	0.8 (1)	0	0	1.8 (2)
VDS 1 at first exam						
0	21.1 (8)	33.3 (7)	19.2 (5)	33.3 (10)	41.7 (10)	21.6 (8)
1	52.6 (20)	47.6 (10)	69.2 (18)	60.0 (18)	50 (12)	54.1 (20)
2	26.3 (10)	14.3 (3)	11.5 (3)	6.7 (2)	8.3 (2)	18.9 (7)
3	0	4.8 (1)	0	0	0	5.4 (2)
VDS 2 at first exam						
0	0	8.7 (2)	9.1 (2)	6.1 (2)	20 (3)	13.0 (3)
1	13.6 (3)	17.4 (4)	22.7 (5)	15.2 (5)	6.7 (1)	26.1 (6)
2	72.7 (16)	47.8 (11)	40.9 (9)	60.6 (20)	53.3 (8)	43.5 (10)
3	13.6 (3)	26.1 (6)	27.3 (6)	15.2 (6)	20.0 (3)	17.4 (4)
VDS 3 at first exam						
0	0	0	4.5 (1)	0	0	11.8 (2)
1	0	5.9 (1)	0	6.7 (1)	11.1 (2)	0
2	29.2 (7)	23.5 (4)	22.7 (5)	6.7 (1)	16.7 (3)	11.8 (2)
3	70.8 (17)	70.6 (12)	72.7 (16)	86.7 (13)	72.2 (13)	76.5 (13)

[1]Vaginal discharge score at second vaginal examination: 0 = clear mucus, 1 = mucus containing flecks of pus, 2 = discharge containing less than 50% pus, 3 = discharge containing more than 50% pus.

22 investigators and intraobserver repeatability for 9 investigators. As in the first experiment, variation was observed between the investigators (n = 22). Overall median interobserver repeatability was $\kappa = 0.47$ (n = 231). Classifying interobserver repeatability, 10, 58, 88, 68, and 7 investigators were in the kappa ranges of $\kappa = 0.81$ to 0.90, $\kappa = 0.61$ to 0.80, $\kappa = 0.41$ to 0.60, $\kappa = 0.21$ to 0.40, and $\kappa = 0.01$ to 0.20, respectively. Overall, median intraobserver repeatability considering 9 investigators was $\kappa = 0.61$. Vaginal discharge scores 0 to 3 were diagnosed correctly in 90, 75.2, 72.7, and 57.8% of the cases, respectively.

DISCUSSION

The objective of the present study was to evaluate the validity of vaginoscopic examination for the diagnosis of clinical endometritis in dairy cattle. To date, science-based information is not available on the repeatability of this diagnostic method. Based on the current literature (Kelton et al., 1991; McDougall et al., 2007), we speculated that findings generated by vaginoscopic examination are influenced by inter- and intraobserver variability, the investigator's experience, and preceding transrectal manipulation of the uterus.

The overall prevalence of CE in this study (42.6%) was similar to that in a previous trial conducted in 2 herds (39.7%; 42.2%) under comparable conditions (Westermann et al., 2010). Even though it is controversial whether cows with mild purulent uterine exudate in the vagina (i.e., VDS 1) require treatment or not (LeBlanc et al., 2002; Williams et al., 2005; Dubuc et al., 2010), animals with a VDS 1, 2, and 3 were classified as affected with CE and treated with $PGF_{2\alpha}$ twice within 2 wk. To reduce bias and to ensure virtually identical conditions for each investigator and for both VE, we minimized the time lag between investigators to less than 1 min and between the 2 VE to less than 10 min. On the other hand, our experimental design with a short time interval between first and second examination did not allow us to completely exclude memorization of VE results for a given cow.

Previous reports found that a stimulation of uterine contractions caused by previous vaginal diagnostic tests improved detection of vaginal discharge (McDougall et al., 2007; Pleticha et al., 2009). Thus, it may be assumed that the first VE stimulated uterine contractions and led to the slight increase in prevalence of CE at the second VE, regardless of the investigator. Furthermore, these reports lead to the hypothesis that uterine stimulation can also be caused by previous transrectal palpation. As current reports argue that it is important to perform a transrectal examination at the same time as a vaginoscopy to increase the sensitivity of the diagnostic process (LeBlanc et al., 2002; Runciman et al., 2008), the effect of previous transrectal palpation was analyzed. The relative risk analysis did not reveal a significant effect of a transrectal palpation of the uterus on the prevalence of CE diagnosed by VE ($P > 0.05$). These data clearly demonstrated that the detection of

vaginal discharge by vaginoscopy cannot be enhanced by a preceding palpation of the uterus before VE.

Interestingly, frequency distribution of vaginal findings of the inexperienced investigators was similar to that of the experienced investigators. Relative risk analysis did not reveal an effect of the investigator's experience on the prevalence of CE identified by VE ($P > 0.05$). A new investigator was assigned weekly to ensure that experience was limited to fewer than 20 animals for each of the investigators designated as Inv 3.

Intra- and interobserver repeatability in vivo may have suffered from limitations due to the nature of a field trial conducted on a commercial dairy farm. Intraobserver repeatability might have been influenced by memorization of cows and the corresponding vaginal findings because cows were not randomly regrouped between the 2 VE to avoid additional stress and movement of the cow, which might itself have influenced vaginal findings. Because 4 cows, on average, were examined and findings were documented in one batch by each investigator spending only a short time per cow (10 ± 5 s), one can speculate that the likelihood of memorization was limited. Communication between investigators was eliminated as much as possible (e.g., by documentation of findings on separated data capture forms and treatment by one person in the absence of the others).

Based on the limitations of the in vivo trial, inter- and intraobserver repeatabilities were calculated in an additional in vitro experiment. Thus, factors were minimized that could bias the results of the visual assessment. In addition, a greater number of investigators could be enrolled for the in vitro trial compared with a field study. Interobserver agreement was as high as in the vivo experiment, as demonstrated by the median κ coefficients (experiment 1: $\kappa = 0.55$, experiment 2: $\kappa = 0.47$). Intraobserver repeatability was higher in the vitro experiment compared with the in vivo trial, as demonstrated by the median κ coefficients. Intraobserver repeatability showed substantial to almost perfect agreement in the first in vitro experiment and showed a substantial agreement for the majority of the investigators in the second experiment (experiment 1: $\kappa = 0.82$, experiment 2: $\kappa = 0.61$). These data confirm the assumption that diagnostic outcomes by examinations in the field can be influenced by various factors (Vyskocil et al., 2008), such as time constraints or movement of the animal. The inter- and intraobserver repeatabilities of the second in vitro experiment were lower compared with those of the first experiment. This can be explained by the reduced lighting provided by flashlight and limited field of vision through the vaginoscope in the second in vitro trial. The repeated examination in the in vitro trial was conducted after 4 wk to exclude potential memorization of findings as discussed for the field study. For most investigators, the intraobserver repeatability calculated in both in vitro trials was higher compared with that in the in vivo trial. Therefore, we assume that the short time interval between first and second VE in the vivo trial had only a limited effect on the relatively high repeatability ($\kappa = 0.55$–60).

Vaginal discharge scores 2 and 3 were clearly defined by Williams et al. (2005) as having a given percentage of pus, whereas a specific percentage was not specified for VDS 1. In our images representing VDS 1, a yellow (pus) area of 4 to 10% was used. In the first and second in vitro experiments, 83.0 and 75.2%, respectively, of the images were diagnosed correctly. Therefore, we conclude that visual assessment can distinguish small increments of differently colored areas and that the thresholds used were adequate.

The accuracy of a diagnostic test can be defined by comparing the outcome of the test with an established standard diagnosis, the gold standard (Knottnerus et al., 2002). As no gold standard exists for the diagnosis of CE (Sheldon et al., 2006), the objective of our in vitro trial was to generate a reference standard to evaluate the accuracy of a semiquantitative visual assessment of different color shades, which is the basic principle of vaginoscopy using VDS. This method was chosen to have precise information about the number of true positive and true negative cases. The in vitro approach enabled us to calculate exactly the sensitivity and specificity of the visual assessment and to quantify limitations (i.e., false-positive and false-negative cases) of a diagnostic method based on visual assessment. The importance of sensitivity and specificity of VE has been emphasized recently (Westermann et al., 2010).

Sensitivity and specificity were high in both in vitro experiments. In the second experiment, the visual assessment of the reference standard images was conducted through a vaginoscope and illuminated with a flashlight to adapt the in vitro approach as closely as possible to the vaginoscopic examination in the cow. Sensitivity and specificity in the second experiment were slightly lower compared with the first experiment (99.6 and 96.7% vs. 96.3 and 90.1%). This observation shows that the use of a vaginoscope and a flashlight complicates the visual assessment and increases error rate. We suspect that this was caused by a reduced field of vision provided by the vaginoscope and the limited brightness of the flashlight.

Furthermore, the sensitivity and specificity of visual assessment in our in vitro experiments were considerably higher compared with results generated in vivo when pregnancy status at 150 DIM or cytological results was used as reference method (LeBlanc et al.,

2002; Barlund et al., 2008), except for our specificity in the experiment 1 (87.4 vs. 85.5%). In these studies, the sensitivity and specificity were calculated defining animals with mild purulent uterine discharge (VDS 1) as not being associated with reduced pregnancy rate and thus as healthy (LeBlanc et al., 2002; Barlund et al., 2008). Thus, for this comparison, our calculation of sensitivity and specificity was adjusted, defining VDS 0 and 1 as healthy. These studies and our data demonstrated that the sensitivity and specificity of a vaginoscopic examination of vaginal discharge is high. The low sensitivity based on the pregnancy status at 150 DIM as reference can be explained by the considerable time lag between diagnosis and pregnancy confirmation and the multitude of factors that may influence pregnancy status (Kasimanickam et al., 2004; Barlund et al., 2008).

CONCLUSIONS

Our data provide evidence that a visual assessment conducted by vaginoscopic examination, although imperfect, does provide a reasonable measurement of vaginal discharge. Vaginoscopy can be seen as a practical tool to distinguish healthy from CE diseased cows. The diagnostic results are not influenced by the experience of the investigator or a preceding transrectal palpation. Sensitivity and specificity of visual assessment determined in vitro is high. Inter- and intraobserver repeatability utilizing the 4-point scoring system was 0.55 to 0.60 and 0.44 (moderate), respectively. The diagnosis of healthy and sick cows showed high agreement between and within different investigators and increased when VDS 1 was considered as healthy. Thus, repeatability of diagnoses of different investigators or at different times was acceptable.

ACKNOWLEDGMENTS

We thank the manager and staff members of the farm for their support. Furthermore, we thank Fabian Lotz (Department of Biometry and Statistics, Faculty of Veterinary Medicine, Berlin, Germany) for invaluable advice on the statistical analyses and the participants of the in vitro experiments for their support.

REFERENCES

Andermann, P., S. Schlogl, U. Mader, M. Luster, M. Lassmann, and C. Reiners. 2007. Intra- and interobserver variability of thyroid volume measurements in healthy adults by 2D versus 3D ultrasound. Nuklearmedizin 46:1–7.

Barlund, C. S., T. D. Carruthers, C. L. Waldner, and C. W. Palmer. 2008. A comparison of diagnostic techniques for postpartum endometritis in dairy cattle. Theriogenology 69:714–723.

Drillich, M., N. Klever, and W. Heuwieser. 2007. Comparison of two management strategies for retained fetal membranes on small dairy farms in Germany. J. Dairy Sci. 90:4275–4281.

Dubuc, J., T. F. Duffield, K. E. Leslie, J. S. Walton, and S. J. LeBlanc. 2010. Risk factors for postpartum uterine diseases in dairy cows. J. Dairy Sci. 93:5764–5771.

Kasimanickam, R., T. F. Duffield, R. A. Foster, C. J. Gartley, K. E. Leslie, J. S. Walton, and W. H. Johnson. 2004. Endometrial cytology and ultrasonography for the detection of subclinical endometritis in postpartum dairy cows. Theriogenology 62:9–23.

Kaufmann, T. B., S. Westermann, M. Drillich, J. Plontzke, and W. Heuwieser. 2010. Systemic antibiotic treatment of clinical endometritis in dairy cows with ceftiofur or two doses of cloprostenol in a 14-d interval. Anim. Reprod. Sci. 121:55–62.

Kelton, D. F., K. E. Leslie, W. G. Etherington, B. N. Bonnett, and J. S. Walton. 1991. Accuracy of rectal palpation and of a rapid milk progesterone enzyme-immunoassay for determining the presence of a functional corpus luteum in subestrous dairy cows. Can. Vet. J. 32:286–291.

Knottnerus, J. A., C. van Weel, and J. W. Muris. 2002. Evaluation of diagnostic procedures. BMJ 324:477–480.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33:159–174.

LeBlanc, S. 2003. Field study of the diagnosis and treatment of clinical endometritis in dairy cattle. Cattle Pract. 11:255–261.

LeBlanc, S. J., T. F. Duffield, K. E. Leslie, K. G. Bateman, G. P. Keefe, J. S. Walton, and W. H. Johnson. 2002. Defining and diagnosing postpartum clinical endometritis and its impact on reproductive performance in dairy cows. J. Dairy Sci. 85:2223–2236.

McDougall, S., R. Macaulay, and C. Compton. 2007. Association between endometritis diagnosis using a novel intravaginal device and reproductive performance in dairy cattle. Anim. Reprod. Sci. 99:9–23.

Pleticha, S., M. Drillich, and W. Heuwieser. 2009. Evaluation of the Metricheck device and the gloved hand for the diagnosis of clinical endometritis in dairy cows. J. Dairy Sci. 92:5429–5435.

Runciman, D. J., G. A. Anderson, J. Malmo, and G. M. Davis. 2008. Use of postpartum vaginoscopic (visual vaginal) examination of dairy cows for the diagnosis of endometritis and the association of endometritis with reduced reproductive performance. Aust. Vet. J. 86:205–213.

Schneider, W., R. Csepan, M. Kasparek, O. Pinggera, and K. Knahr. 2002. Intra- and interobserver repeatability of radiographic measurements in hallux surgery: Improvement and validation of a method. Acta Orthop. Scand. 73:670–673.

Sheldon, I. M., G. S. Lewis, S. LeBlanc, and R. O. Gilbert. 2006. Defining postpartum uterine disease in cattle. Theriogenology 65:1516–1530.

Sheldon, I. M., D. E. Noakes, A. N. Rycroft, and H. Dobson. 2002. Effect of postpartum manual examination of the vagina on uterine bacterial contamination in cows. Vet. Rec. 151:531–534.

Vyskocil, M., T. Palenik, R. Dolezel, S. Cech, and M. Vecera. 2008. Systematic clinical examination of early postpartum cows and treatment of puerperal metritis did not have any beneficial effect on subsequent reproductive performance. Vet. Med. (Praha) 53:59–69.

Westermann, S., M. Drillich, T. B. Kaufmann, L. V. Madoz, and W. Heuwieser. 2010. A clinical approach to determine false positive findings of clinical endometritis by vaginoscopy by the use of uterine bacteriology and cytology in dairy cows. Theriogenology 74:1248–1255.

Williams, E. J., D. P. Fischer, D. U. Pfeiffer, G. C. England, D. E. Noakes, H. Dobson, and I. M. Sheldon. 2005. Clinical evaluation of postpartum vaginal mucus reflects uterine bacterial infection and the immune response in cattle. Theriogenology 63:102–117.