Evaluation of udder firmness by palpation and a dynamometer

A. Rees, C. Fischer-Tenhagen, and W. Heuwieser
Clinic for Animal Reproduction, Faculty of Veterinary Medicine, Freie Universität Berlin, Koenigsweg 65, 14163 Berlin, Germany

ABSTRACT

Swelling of the mammary gland is an important health status sign for clinical exploration and palpation is a routine diagnostic tool for mastitis detection in dairy cows. Data on repeatability or validity of specific methods of udder palpation are rare. The overall objective was to study the validity of estimates of udder firmness generated by palpation and by using a validated dynamometer. Specifically, we set out to determine within-observer repeatability and between-observer repeatability in 2 specific experiments. Additionally, we compared a 4-point palpation scoring system with estimates obtained with a dynamometer in this study. In a pilot trial, we determined the range of udder firmness of 25 cows and developed an in vitro model for udder firmness. This model enabled training of the observers and allowed investigating a 4-point palpation scoring system. In vivo, udder firmness was determined before and after milking by palpation and by using a dynamometer. Within-observer repeatability based on estimates of udder firmness of 25 cows obtained by 3 observers on a single day by palpation was 0.968. Within-observer repeatability of estimates of udder firmness of 25 cows obtained with the dynamometer by a single observer was 0.997. The coefficient of variation of the same measures was 9.1%. To determine between-observer repeatability (palpation: 0.932; dynamometer: 0.898), udder firmness of 100 cows was measured on 4 different days by 9 observers in experiment 2. Udder firmness in dairy cows could be measured repeatably with the dynamometer and by palpation, especially when performed by a single observer. Estimates of udder firmness generated by palpation and with the dynamometer were moderately related (correlation coefficient = 0.54). Training of observers through the pilot trial or practical experience in the 4 d of the study in experiment 2 did not improve the correlation. Further research is warranted to understand how udder firmness develops in infected udders.

Key words: udder firmness, dynamometer, palpation, repeatability

INTRODUCTION

Swelling of the mammary gland is an important inflammation sign and is associated with clinical mastitis (IDF, 1999). Furthermore, udder swelling is a health status sign for clinical exploration. Clinical examination of the udder includes palpation of the udder tissue. Palpation is a routine diagnostic tool for mastitis detection in dairy cows (Hillerton, 2000; Pyörälä, 2003; IDF, 2007) and is a simple and quick method to diagnose relevant findings indicative of inflammation of the udder (i.e., nodes, heat, pain, and swelling). Palpation was used in numerous studies addressing detection (Polat et al., 2010; Petrovski et al., 2011), prevention (Runciman et al., 2010), and antibiotic treatment (Cao et al., 2007; Lago et al., 2011) to differentiate between cows with and without mastitis. Apart from visual observation of milk and implementation of a California mastitis test, palpation is an important diagnostic criterion for cow-side treatment before cytological and bacteriological laboratory results of milk samples are available (Lago et al., 2011).

Internationally recognized textbooks detail the method of palpation of the udder as an integral part of the clinical examination of an individual cow (Rosenberger et al., 1990) and the physical examination of the udder as a basic component of an udder health management program for dairy herds (Radostits et al., 2001). Two 5-point palpation scoring systems have been developed to describe the severity of symptoms (Petrovski et al., 2011) or the clinical status of the quarters (Hogan et al., 1995; Scaletti and Harmon, 2012). A 3-point scoring system for udder firmness has been assessed to examine effects of changes of milking frequency (Gleeson et al., 2007) or omission of a milking event (O’Driscoll et al., 2011) on animal welfare. To achieve more objective measures of udder firmness, some studies used different technical devices (i.e., dynamometer) for measuring the force to indent the udder tissue for studying effects related to animal welfare (Tucker et al., 2007, 2009) and dry cow management (Bertulat et al., 2013). Except for 2 studies estimating udder firmness with a technical device (Bertulat et al., 2012) or by palpation (Houe et al., 2002), data on repeatability or validity of methods to examine udder firmness have not been described. Furthermore, palpation of the udder tissue as an es-
ential part of cow-side mastitis diagnostics is not well defined in the current literature.

Thus, the overall objective of our study was to evaluate the validity of estimates of udder firmness determined by palpation and by using a dynamometer. Specifically, we set out to (1) determine within-observer repeatability (WOR; experiment 1) and between-observer repeatability (BOR; experiment 2) of estimates of udder firmness generated by palpation and using a dynamometer and (2) compare a 4-point palpation scoring system with measures obtained with a dynamometer.

MATERIALS AND METHODS

Animals

The study was conducted in September and October 2012 on a commercial dairy farm milking 175 dairy cows in Brandenburg, Germany. For the study, a total of 150 Holstein-Friesian and crossbreeds of Holstein-Friesian dairy cows (71 primiparous and 79 multiparous) were used. Cows were housed in a deep-bedded stall. They received a balanced TMR based on 52.7% corn silage, 24.8% grass silage, 9.3% brewers grains, 5.3% corn meal, 4.7% rapeseed, 2.2% triticale, 0.9% straw, 0.1% urea, and basic mineral mix. The TMR was delivered twice per day at 0830 and 1700 h. Cows were milked twice daily (0700 and 1500 h) in a 2 × 8 Herringbone milking parlor (System Happel GmbH, Friesenried, Germany). The rolling herd average (305 d) was 8,745 ± 2,149 kg of milk/cow per year. The actual milk yield was 28.4 ± 8.4 kg/cow per day. Cows were in different stages of lactation (147 ± 94 DIM; mean ± SD).

Measurements

Udder firmness was determined by palpation and by using a dynamometer (Penefel DFT 14; Agro Technologie, Forges-les-Eaux, France). The dynamometer was used following the standard operating procedure (SOP) described by Bertulat et al. (2012). In brief, the left hind udder quarter of cows enrolled was used to determine udder firmness. The measuring point was located in the horizontal and vertical center of the left hind quarter. This point was marked with livestock paint crayons (Raidex GmbH, Dettingen, Germany) to ensure a consistent measurement location within the udder. The cow had to stand with all 4 legs on a level surface during the whole measurement. After 5 consecutive measurements performed within 10 s, the dynamometer displayed the arithmetic mean and coefficient of variation. Values with a coefficient of variation exceeding 10% were discarded and the measurement repeated. This general procedure of measuring udder firmness with the dynamometer was identical in all experiments.

Palpation was conducted by pressing the fingertips of all fingers of 1 hand except the thumb into the udder tissue. The measuring point was identical to the marked measuring point for the dynamometer. This approach was selected to standardize the location within the udder and within the quarter to reduce bias due to inhomogeneity of the tissue.

Pilot Trial

A pilot trial was conducted to determine the range of udder firmness expectable in healthy dairy cows considering different stages of lactation before and after milking. Furthermore, we wanted to investigate if observers were able to correctly classify different firmness levels of specimens presented in vitro on a 4-point palpation scoring system with sufficient repeatability before implementing a large field study.

In total, 25 cows were fixed in the head locker before and after the evening milking and udder firmness was determined using the dynamometer on 10 consecutive days. To avoid bias due to recognizing an individual cow by the observer, cows were fixed in different positions after milking each day. All measurements were conducted within 45 min before and within 45 min after milking. Based on the range of udder firmness measured with the dynamometer, a 4-point classification system was developed by dividing the range determined by 4 and calculating the mean value of the estimates within that increment.

An in vitro trial comprising a reference standard was conducted to enable the calculation of WOR and BOR for the 4-point classification system, which was also used in the in vivo trial. As a reference standard, 4 tire tubes (30.48 cm; Schrader valve, TAQ-33 Technique + Quality; BICO Zweirad Marketing GmbH, Verl, Germany) were inflated. The firmness levels of the tires were adjusted such that measurements obtained with the dynamometer equaled the 4 means of udder firmness calculated for the 4-point classification system developed with data obtained in the pilot trial. As a result, each tire tube represented 1 distinct score of the 4-point system. To ensure repeatability of this experiment, air pressures measured with a manometer as reference standard (GDH 200-14; Greisinger Electronic GmbH, Regenstauf, Germany) were also reported (Table 1). For example, tire tube 1 was inflated until 0.766 kg was measured by the dynamometer corresponding to 2.5 kPa measured with the manometer and representing score 1.
Training of the Observers

Nine observers (3 male and 6 female) consented to participate in the study, including 3 students of veterinary medicine (fourth, fifth, and sixth year), 3 graduated veterinarians, and 3 milkers. First, the 4 inflated tire tubes were presented one by one in a random order to train palpation and the use of the dynamometer, respectively. Every observer had to palpate the tubes until 4 consecutive presented tire tubes (i.e., all 4 scores) were scored correctly. The use of the dynamometer and the guidelines of the SOP were explained and all observers had to measure firmness of all 4 tire tubes until their results were concordant with the calculated firmness.

Experiment 1

The objective of experiment 1 was to determine the WOR of estimates of udder firmness determined by palpation and the dynamometer. Udder firmness of 25 cows was evaluated on a single day both by palpation with the fingertips and by using a dynamometer.

To evaluate the WOR of udder firmness scores obtained by palpation, 3 (1 student and 2 milkers) of the 9 trained observers were randomly selected. Each of the 3 observers obtained 1 measurement per cow. This procedure was repeated another 9 times within 45 min before and within 45 min after milking, resulting in overall 10 measurements per cow and 500 measurements per observer. The observers measured and recorded their results on case report forms independently of each other within 2 ± 1 min. Blank case report forms were used for each measurement run.

To evaluate the WOR of udder firmness determined with the dynamometer, 1 additional observer measured udder firmness of the same 25 cows consecutively 10 times per cow within 45 min before and within 45 min after milking by using a single dynamometer. As described above, cows were selected by chance and fixed in the head locks.

Experiment 2

To determine the BOR, udder firmness of 100 cows was determined on 4 different days by all 9 observers both by palpation and by using the dynamometer. The procedure of measuring was identical to experiment 1. Due to logistical reasons, udder firmness was recorded first either by palpation (n = 5) or by using the dynamometer (n = 4). In total, 4 different dynamometers were used.

Statistical Analysis

Data were entered into Excel spreadsheets (version 2010; Microsoft Corp., Redmond WA) and statistical analyses performed with SPSS for Windows (version 20.0; SPSS Inc., Munich, Germany) and R (version R-2.15.2.tar.gz; http://www.r-project.org/; Statistics Department of the University of Auckland, Auckland, New Zealand). A Kolmogorov-Smirnov test was performed to test whether estimates obtained with the dynamometer in experiments 1 and 2 were distributed normally. Because the estimates were not normally distributed, effect of time (before and after milking) on udder firmness was statistically determined by a Wilcoxon signed-rank test.

A Kruskal-Wallis test was performed to test the effect of study day (n = 4) and palpation score on medians of estimates of udder firmness obtained in experiment 2. Repeatability of both methods of estimating udder firmness was calculated separately for measurements taken before and after milking as well as for all data combined, regardless of time of measurement.

To determine the WOR of measurements obtained with the dynamometer by a single observer, the minimum, maximum, median, interquartile range (IQR), and coefficient of variation were calculated for the 25 measured cows individually and combined. To determine the WOR of estimates obtained by palpation and with the dynamometer, intraclass correlations (ICC; model: 2 way; type: consistency) for average (ICCAm) and single measures (ICCSm) were calculated. According to Shrout and Fleiss (1979), the ICC was calculated in a 2-way mixed model with measures of consistency due to the study design (i.e., each observer out of a fixed set of observers measured each cow with systematic variability due to observers or measures considered to be irrelevant). An ICC of 0 indicated no relationship between the measurements and their coherence could be regarded as essentially random. If the ICC was 1, all measurements showed a perfect correlation. To determine the BOR of measurements of all 9 observers obtained with the dynamometer, ICCAm and ICCSm (model: 2 way; type: consistency) were calculated. To determine the BOR of udder firmness obtained by palpation, ICCAm and ICCSm (model: 2 way; type: consistency) were calculated.

Table 1. Comparison of the 4-point palpation system with estimates of udder firmness measured with the dynamometer (n = 500) and tire pressures measured with a manometer in the pilot trial.

<table>
<thead>
<tr>
<th>Score</th>
<th>Dynamometer (kg)</th>
<th>Manometer (kPa)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Minimum</td>
<td>Maximum</td>
</tr>
<tr>
<td>1</td>
<td>0.306</td>
<td>1.225</td>
</tr>
<tr>
<td>2</td>
<td>1.226</td>
<td>2.144</td>
</tr>
<tr>
<td>3</td>
<td>2.145</td>
<td>3.062</td>
</tr>
<tr>
<td>4</td>
<td>3.063</td>
<td>3.981</td>
</tr>
</tbody>
</table>

Table 1. Comparison of the 4-point palpation system with estimates of udder firmness measured with the dynamometer (n = 500) and tire pressures measured with a manometer in the pilot trial.
To evaluate the relationship between the 4-point palpation scoring system and measures obtained with the dynamometer, the Spearman rank correlation coefficient was calculated. Overall, 1,800 paired observations (i.e., 1,800 measurements obtained by palpation and 1,800 measurements obtained with the dynamometer) were used (9 observers, 100 cows, and 2 replicates).

We calculated ICC to analyze the BOR and WOR of measurements conducted by multiple observers and both continuous and ordinal data for the following reasons. Intraclass correlation values are equivalent to the weighted kappa (Fleiss and Cohen, 1973) often used to assess agreement of measurements for categorical data (Landis and Koch, 1977). Additionally, ICC is applicable to continuous data (Shrout and Fleiss, 1979). Therefore, comparability of both data sets was ensured. Estimates of udder firmness (n = 1,800) obtained by the dynamometer in experiment 2 were reclassified into a 4-point scoring system to generate ordinal data and to allow comparison of correlated ICC.

**RESULTS**

**Measurements**

In total, 6,100 udder firmness measurements performed by palpation and by using the dynamometer were documented in the pilot trial (n = 500), experiment 1 (n = 2,000), and experiment 2 (n = 3,600). Using the dynamometer by applying the SOP took considerably more time than determining udder firmness by palpation. Specifically, the mean time needed to measure udder firmness of a single cow by palpation and with the dynamometer was 12 ± 1.2 s and 48 ± 18 s (P = 0.04), respectively.

**Pilot Trial**

Minimum and maximum udder firmness of the 25 cows enrolled in the pilot trial measured with the dynamometer are summarized in Table 1. Median (and IQR) of udder firmness was 1.296 kg (0.947 to 2.355 kg) and 0.572 kg (0.465 to 0.782 kg) before milking and after milking, respectively. Median milking-induced decrease [i.e., difference (and IQR)] in udder firmness was 0.763 kg (0.426 to 1.570 kg; P = 0.00). Graduated veterinarians and milking personnel needed an average of 3 replicates, whereas the students needed an average of 2 replicates to correctly diagnose the firmness of all 4 tire tubes.

**Experiments 1 and 2**

Twenty-five and 100 cows were enrolled to determine WOR and BOR of measuring methods in experiments 1 and 2, respectively. Estimates of udder firmness measured with the dynamometer by all observers in experiments 1 and 2 ranged from 0.302 to 5.309 kg. The median (and IQR) udder firmness was 1.373 kg (0.996 to 1.845 kg) and 0.659 kg (0.507 to 0.905 kg) before and after milking, respectively. The median milking-induced decrease [i.e., difference (and IQR)] in udder firmness was 0.605 kg (0.350 to 1.017 kg; P = 0.00).

To enable comparison of udder firmness values measured in this and a previous study, median udder firmness of cows >304 DIM (n = 5) was calculated. The median (and IQR) udder firmness of this subset of cows was 0.840 kg (0.604 to 1.135 kg) and 0.443 kg (0.363 to 0.506 kg) before and after milking, respectively. Medians of measurements obtained with the dynamometer in experiment 2 differed (P = 0.00) between the 4 d of the study (Figure 1) as well as between different palpation scores (Figure 2).

The coefficient of variation of repeated measures obtained with the dynamometer by a single observer (experiment 1) was 9.1% (mean ± SD = 1.176 kg ± 0.107 kg; n = 500). Within-observer repeatability of the same measures of udder firmness was 0.997 (ICCam; 95% CI = 0.996–0.998) and 0.971 (ICCCsm; 95% CI = 0.959–0.981) for all measures (n = 500). Within-observer repeatability of these measures was 0.996 (ICCam; 95% CI = 0.993–0.998) and 0.960 (ICCCsm; 95% CI = 0.935–0.978) before and 0.996 (ICCam; 95% CI = 0.993–0.998) and 0.957 (ICCCsm; 95% CI = 0.930–0.977) after milking. Within-observer repeatability for measures of udder firmness obtained by palpation via 3 observers was 0.968 (ICCam; 95% CI = 0.960–0.975) and 0.753 (ICCCsm; 95% CI = 0.706–0.798) for all measures (n = 500). Within-observer repeatability of the same measures was 0.931 (ICCam; 95% CI = 0.905–0.952) and 0.575 (ICCCsm; 95% CI = 0.489–0.666) before and 0.942 (ICCam; 95% CI = 0.920–0.960) and 0.618 (ICCCsm; 95% CI = 0.535–0.704) after milking. Between-observer repeatability of measures obtained with the dynamometer by all 9 observers of experiment 2 was 0.898 (ICCam; 95% CI = 0.867–0.925) and 0.329 (ICCCsm; 95% CI = 0.266–0.407) for all measures (n = 1,800). Between-observer repeatability of the same measures was 0.894 (ICCam; 95% CI = 0.860–0.922) and 0.483 (ICCCsm; 95% CI = 0.405–0.569) before and 0.884 (ICCam; 95% CI = 0.846–0.915) and 0.458 (ICCCsm; 95% CI = 0.379–0.545) after milking. Between-observer repeatability of measures obtained by palpation via all 9 observers was 0.932 (ICCam; 95% CI = 0.917–0.945) and 0.603 (ICCCsm; 95% CI = 0.550–0.657) for all measures (n = 1,800). The calculated BOR of the same measures was 0.905 (ICCam; 95% CI = 0.874–0.930) and 0.513 (ICCCsm; 95% CI = 0.435–0.597) before and 0.835 (IC-
Cam; 95% CI = 0.781–0.879) and 0.359 (ICCsm; 95% CI = 0.284–0.447) after milking.

After reclassification of estimates of udder firmness (n = 1,800) obtained by the dynamometer in experiment 2, the recalculated BOR was 0.391 (ICCsm; 95% CI = 0.336–0.452) and 0.852 (ICCam; 95% CI = 0.820–0.881).

**Relationship Between 2 Methods of Measuring Udder Firmness**

Spearman rank correlation coefficients quantifying the relationship between 900 udder firmness estimates per study day (n = 4) obtained by palpation and measured with the dynamometer were 0.49, 0.58, 0.56, and 0.55 on d 1, 2, 3, and 4, respectively. On all days, a difference (P = 0.00) between the 2 measurement methods existed.

Overall, the Spearman rank correlation coefficient was 0.54 (P = 0.00; n = 3,600; Figure 2). The coefficient of correlation between the 2 methods to determine udder firmness was lower before (r = 0.24; P = 0.00; n = 1,800) than after milking (r = 0.32; P = 0.00; n = 1,800).

**DISCUSSION**

**Measurements**

We assumed that a validated technical device (i.e., dynamometer) would produce less subjective values than palpation. Therefore we used the dynamometer as a reference device to examine if an objective and repeatable estimate of udder firmness by palpation could be established.

Whereas in a previous study 9.3% of the measurements with the dynamometer exceeded the cut point of 10% coefficient of variation and were regarded as invalid (Bertulat et al., 2012), in the current study, 24% of the measurements had to be repeated. We speculated that the high number of measurements conducted on a given day (n = 50 to 62) and the considerable time commitment (2.5 h per day of study) by each of the observers might have decreased SOP compliance and caused a higher repetition rate.
Previous studies implementing palpation of udder tissue to detect signs of inflammation did not provide any details of the method of palpation or the location within the udder (Polat et al., 2010; Petrovski et al., 2011). Other studies conducting udder firmness measurements and applying a firmness scoring system (Gleeson et al., 2007; O’Driscoll et al., 2011) did only vaguely describe the location of palpation (i.e., between the hind legs). Houe et al. (2002) compared clinical evaluations of udder health characteristics among observers using a defined scoring system and the palpation method described by Rosenberger et al. (1990). Those authors quantified the agreement of evaluation of udder hardness via a 5-point score (i.e., categorical data by calculating above-mentioned $\kappa$ and weighted $\kappa$ values). They compared recordings of observers 2 by 2 and found poor agreement between clinical parameters not directly related to pathological conditions, such as udder hardness ($\kappa = 0.31$), but good agreement for pathological changes such as nodes ($\kappa = 0.7$). In their study, palpation encompassed a superficial and deep palpation of every quarter after milking with the palm of the hand, whereas in our study observers determined udder firmness just with their fingertips at the same point at which the dynamometer was applied. Previously, it has been demonstrated that udder firmness changes considerably between front and hind quarters as well as within the location of a given quarter (Bertulat et al., 2012). To ensure objective and repeatable measurements with the dynamometer, Bertulat et al. (2012) outlined in their SOP that the cow must stand still with all 4 legs on a level ground during measurement. We observed that a measurement with a coefficient of variation less than 10% could be ensured when following this procedure. The measuring point was located in the horizontal and vertical center of the left hind quarter. To address the objective of our study comparing palpation with a technical device, it was mandatory to conduct the measurements at the same quarter and at the same level to exclude bias. Therefore, our method of palpation was chosen to enable a more standardized comparison of measurements between observers and between the 2 methods. Other variables could have biased the results and decreased repeatability and correlation between the 2 methods. Before milking, the weight of the milk ex-

---

**Figure 2.** Box plot of relationships between estimates of udder firmness obtained by palpation (palpation score 1 to 4) and with the dynamometer (kg). The heavy black line inside each box marks the median; whiskers end at the smallest and largest statistical values that are not outliers or extreme values. $\bigcirc =$ outlier; $^* =$ extreme value.
panded the system of udder suspension, which most likely affected udder firmness. Discomfort and pain has been described for the cow lying with a filled udder due to external pressure on the udder (Österman and Redbo, 2001). We speculate that the position of hind legs relative to the hind quarter could also affect udder firmness due to increasing external pressure on the udder. Because we did not lift the udder quarter while estimating udder firmness, our palpation method probably was more susceptible to changes in udder firmness due to higher tension of the udder suspension system before milking or the position of the hind legs and ankle joints exerting external pressure onto the udder, as described previously (Österman and Redbo, 2001).

**Pilot Trial**

A wide range of udder firmness was determined by the 500 conducted measurements. Minimum values of udder firmness obtained with the dynamometer were almost identical in the pilot trial and the experiments. This may be due to the lower threshold of 0.3 kg of the dynamometer; but only 3.4% of all values were between 0.3 and 0.4 kg. The maximum values, however, were different between the pilot study and the experiments. This is probably due to the larger sample size of experiment 2 compared with the pilot trial and the larger number of observers. It is unfortunate that data to compare our findings with are not yet available.

As no gold standard exists for the diagnosis of udder firmness, the objective of the in vitro trial was to generate a reference standard having properties (size and firmness) similar to an udder quarter. Except for isolated perfused bovine udders used in pharmacodynamic studies (Ehinger et al., 2006; Kietzmann et al., 2008, 2010), in vitro approaches establishing defined reference standards to investigate characteristics of diagnostic methods have not been described. Other studies, however, mentioned the lack of a gold standard for udder health (Houe et al., 2002) and for udder examination and found evidence that the sensitivity and specificity of diagnosing clinical mastitis differ among observers (McDougall, 1999). By evaluating different materials such as foams, rubbers, and inflatable objects, we found that the pressure of an inflated tire tube could be adapted to represent a certain firmness and measured accurately, thus creating an artificial reference for udder firmness. This allowed training of multiple observers with specimens of identical characteristics and to study whether the described method of palpation is adequate to categorize firmness on a 4-point classification system. We used an even-numbered score to avoid biased results due to a middle option of an odd-numbered score (e.g., point 3 of a 5-point score; Clark and Watson, 1995). Additionally, the range of udder firmness was suitable for a 4-point score. An increase in score points (e.g., 6-point palpation scoring system) could have led to reduced validity (Clark and Watson, 1995). It is noteworthy that after only a few replicates, each observer was able to correctly classify the 4 firmness levels. It is obvious that the complex anatomy of a mammary gland could not be perfectly simulated, but an in vitro model with a reference standard that could be described specifically (i.e., pressure applied in kilopascals) was necessary to evaluate and standardize palpation.

**Observers**

The professional background (students, veterinarians, and milkers) and an assumed different experience did not affect the ability to classify the 4 different firmness levels by palpation. In vitro, only 2 to 3 attempts were needed to categorize 4 firmness levels correctly in the 4-point classification system. We speculate that a training effect resulted from the possibility of comparing all different firmness levels in a short time and directly one after another. In vivo, the Spearman rank correlation coefficient did not change over the days of the trial. Therefore, we did not observe an improvement of the correlation between estimates of udder firmness obtained by palpation and measurements obtained with the dynamometer over time. This observation differed from the results of an earlier study (Houe et al., 2002) that found increased κ correlation values for examinations conducted on different days. Those authors assumed this to be an effect of practicing clinical examination. In their study, however, only sense-based udder examinations (i.e., inspection and palpation) were conducted, whereas in our study, a technical device (i.e., dynamometer) was compared with palpation. We assume that improvement of correlation over time can be realized when sense-based and, therefore, trainable methods are used. The observers in their study were given a vivid description of the scoring system, but training was not conducted before the initiation of the study. We speculate that in our study training could have contributed to a more consistent diagnostic performance, reducing variation that might have affected correlation. Interestingly, as a result of their findings Houe et al. (2002) stated that a need seems to exist for an increase in training and calibration of score values.

**Repeatability of Estimates of Udder Firmness in Experiments 1 and 2**

We are aware of potential criticism regarding the use of the coefficient of variation and ICC in the context...
of non-normally distributed data. However, we chose ICC instead of weighted $\kappa$ to facilitate comparison of results. Additionally, we suggest a mixed model to be an appropriate statistical method for future studies due to the possibility of estimating the contributions to variability from cows, for example.

**Dynamometer**

The measurements were conducted at the same time of the day and within 45 min before and after milking in each replicate. We assume that this consistency in time contributed to repeatability of estimates of udder firmness. This finding substantiates the recommendation of a previous study conducted by our group (Bertulat et al., 2012). Significantly different medians of measurements (Figure 1) were obtained with the dynamometer in experiment 2 simply because different animals were used on the 4 d of the study.

Our results indicate that measurements obtained with the dynamometer by a single observer were highly repeatable. Whereas averaged measurements of different observers also had high repeatability, a single measurement of one observer was poorly related to the measurements of other observers. These data are contrary to a previous study (Bertulat et al., 2012) in which a high BOR with an average inter-observer variation for all measurements of 11.3% ($r = 0.94$) was described. The different statistical calculations (i.e., comparing 3 pairs of observers separately vs. ICC), a smaller number of observers (2 vs. 9), or biasing conditions (movement of cows between measurements and variation in ankle joint position) can help explain this discrepancy. As in our study, Tucker et al. (2007, 2009) described a standardized measuring point and defined the penetration depth for the device used. Repeatability, however, was not calculated and udder firmness values were obtained with a different measuring device, resulting in values with a different unit (i.e., gram of force). Therefore, our estimates of udder firmness are not directly comparable with their study. Bertulat et al. (2012) validated the same device as used in our study and applied an SOP, but did not mention absolute udder firmness values. In a previous trial, we enrolled only cows within the last week before dry-off ($343 \pm 39$ DIM) and until 9 d after dry-off to study udder firmness after dry-off (Bertulat et al., 2013). Due to the different DIM and higher daily milk yield in the current study, data of the 2 studies are also not directly comparable. Cows >304 DIM ($n = 5$) in the current study had very similar average udder firmness before and after milking to the baseline values of cows before dry-off in a previous study (Bertulat et al., 2013).

**Palpation**

Categorizing udder firmness by palpation on a 4-point scale was highly repeatable for a given observer both for a single or average measurement. Averaged measurements of different observers showed high repeatability and single measurements between observers were moderately correlated. These results indicate that a single classification of udder firmness on a 4-point scale obtained by palpation, carried out once and conducted by different observers, is more comparable than measurements obtained by dynamometer by different observers. Between-observer repeatability of average measurements of both measuring methods was similar.

Even though ICC was calculated for both variables, it must be noted that estimates of firmness obtained by palpation were on an ordinal scale (1 to 4) and values obtained by the dynamometer were continuous variables. A direct comparison of ICC using different data scales, however, is critical. Therefore, we reclassified estimates of udder firmness ($n = 1,800$) obtained by the dynamometer in experiment 2 into a 4-point scoring system to generate ordinal data and to allow a comparison of correlated ICC. The recalculated interclass correlation remained similar to the first calculated correlation, still indicating a better correlation for a single measurement obtained by palpation and different observers.

In contrast to our results, a previous study (Houe et al., 2002) demonstrated poor agreement of udder examination when classifying nonpathological conditions, such as hardness of the udder parenchyma. This observation was based on 178 paired observations of clinical evaluations performed by 5 clinicians in 4 herds within 90 min after milking. A 5-point scale was used to classify hardness of udder tissue ($1 = $soft, $3 = $firm, and $5 = $hard). The calculated $\kappa$ value for the parameter udder hardness was 0.31, indicating low repeatability. This discrepancy can be explained by the different methods used to determine udder firmness or hardness. Whereas observers in our study used their fingertips and indented the udder tissue at identical locations, a palpation of the whole quarter and udder was conducted in the study of Houe et al. (2002). We speculate that quarters with inhomogeneous tissue composition will cause different findings. Our results demonstrate that studies working with more than 1 observer to perform udder examination could be biased and of low informative value. Additionally, poor observer agreement could lead to false conclusions regarding dairy herd management and therapy of udder diseases due to a lack of a standardized decision basis.
**Relationship Between 2 Methods of Measuring Udder Firmness**

It must be mentioned that comparison of 2 measurement methods via calculation of a correlation coefficient could be critical for several reasons. It has been described that sensitivity of correlation methods to sample heterogeneity could result in wrong conclusions regarding the agreement of measurements (Bland and Altman, 1995; Atkinson and Nevill, 1997). Additionally, correlation coefficients were described as a measure of association, not a measure of agreement (Altman and Bland, 1983). In our study, an ordinarily scaled 4-point palpation scoring system was compared with continuous-scale dynamometer values. We used the Spearman rank correlation coefficient to assess the relationship between both methods of measuring udder firmness. We are aware of the limitations of the correlation coefficient as a measure of agreement (Altman and Bland, 1983), but the scale of our data limited use of more adequate methods. A comparison of results with other studies was not possible simply due to a lack of publications investigating a similar approach.

Association of estimates of udder firmness on the 4-point scoring system obtained by palpation and with the dynamometer was moderate. Although medians of estimates of udder firmness differed significantly between the 4 classes of the palpation scoring system, the boxes and whiskers of the box plots overlap (Figure 2). Therefore, it is not possible to classify an udder into 1 specific score by measuring firmness in kilograms and vice versa. Besides the subjective nature of palpation, one explanation for the poor relationship could be movements of the cow between measurements leading to different postures of the hind legs relative to the udder and influencing udder firmness. Furthermore, when comparing 2 methods, it is important to recognize that neither approach may be ideal. Nevertheless, data of our and a previous study (Houe et al., 2002) provide evidence that palpation of the udder is a method with limited repeatability when multiple observers are involved. It is noteworthy that studies comparing multiple diagnostic methods to examine udder health are scarce.

Correlation between palpation and dynamometer before milking was lower than the correlation after milking. We assume that some of observers were reluctant to press the measuring tip of the dynamometer or their fingertips with adequate pressure into the tissue of hard udders found before milking, thus confounding the measurements before milking. Additionally, the volume of the milk or the weight effect on the udder suspension system could have affected measurements. Our data support the recommendation of an udder examination after milking and provide science-based information for a description in a textbook (Rosenberger et al., 1990) in which a palpation after milking was advised.

In our study, all measurements were conducted using healthy udders to exclude additional bias through inhomogeneous swelling. Considering pathological conditions, further research is warranted to understand how udder firmness develops in infected udders.

**CONCLUSIONS**

Our results demonstrate that estimates of udder firmness generated by palpation and with the dynamometer were lowly related. Although observers were trained in palpation and scoring udder firmness on a 4-point scale and in the use of the dynamometer before determination of udder firmness in vivo, correlation was limited. Udder firmness in dairy cows can be measured repeatedly by palpation and with the dynamometer, especially when performed by a single observer. Although imperfect, a 4-point palpation scoring system provides a feasible and easy-to-use classification system to estimate udder firmness.

**ACKNOWLEDGMENTS**

A. Rees was partly funded by Tiergyn Berlin e.V. (Berlin, Germany). The authors thank the participating farm, the observers, and Marcus Groß (Department of Statistics, Freie Universität Berlin, Berlin, Germany). Special thanks go to Annika Mahrt and Onno Burfeind (both from Clinic for Animal Reproduction, Faculty of Veterinary Medicine, Freie Universität Berlin) for their support.

**REFERENCES**


