Vurdering af dyrevelfærd i malkekvægsbesætninger på basis af egenkontrolregistreringer

Mogens A. Krogh[1] & Søren Saxmose Nielsen[2]

December 2019

[1]Institut for Husdyrvidenskab, Aarhus Universitet

[2]Institut for Veterinær- og Husdyrvidenskab, Københavns Universitet

Forord

Denne rapport er udarbejdet som afrapportering på projektet 'Vurdering af dyrevelfærd i malkekvægsbesætninger på basis af egenkontrol registreringer' igangsat af Videncenter for Dyrevelfærd. Projektgruppen har bestået af Mogens Krogh (Aarhus Universitet) og Søren Saxmose Nielsen (Københavns Universitet).

Projektet har løbet i 2019, hvor projektgruppen løbende har afholdt møder. Der er ikke blevet afholdt møder eller kommenteringer med nogle udenfor projektgruppen undervejs i forløbet. Data er stillet til rådighed i denne rapport af Arla Foods.

Intern fagfællekommentering er foretaget af Tine Rousing, Institut for Husdyrvidenskab, Aarhus Universitet.

Rapporten består af tre særskilte, faglige dele, alle skrevet på engelsk. Første del omhandler baggrund, definition og formål ved at forsøge at anvende egenkontrolregistreringer til vurdering af dyrevelfærd baseret på litteratur. Anden del omhandler overensstemmelse imellem landmandens registreringer i forhold til ekstern auditør og vurdering af, om overensstemmelse er tilstrækkeligt og til hvilket formål. Tredje del omhandler potentialet for anvendelse af disse egenkontrolregistreringer til benchmarking.

Resumé

Velfærdsvurdering på besætningsniveau ved direkte observation af dyrene (dyrebaserede indikatorer) er oftest blevet foretaget af eksterne parter, hvilket bliver betragtet som en barriere for implementering af protokoller til velfærdsvurdering. En alternativ tilgang er at anvende dyrebaserede indikatorer, der er udført og indsamlet af kvægbrugeren. Formålet med dette projekt er at vurdere kvaliteten af data fra egenkontrol med dyrevelfærd sigte på overvågning, certificering og benchmarking. Det empiriske grundlag er data fra kvalitetssikringsprogrammet Arlagaarden®Plus, der anvender dyrebaserede indikationer i egenkontrol for dyrevelfærd. Egenkontrollen foregår kvartalsvis ved kvægbrugere i Storbritannien, Holland, Luxembourg, Belgien, Tyskland, Danmark og Sverige. Deltager kvægbrugeren i egenkontrolprogrammet, gives en kompensation via mælkeprisen.

I rapporten illustrerer vi, at forskelligheder i opfattelse af dyrevelfærd udfordrer observation af dyrevelfærden, fordi observationerne skal passe til formålet med at lave velfærdsvurderingen. Når opfattelserne af dyrevelfærd varierer, så er det sandsynligt, at formålene med at vurdere dyrevelfærden også varierer. Vi understreger, at et veldefineret formål med velfærdsvurdering er nødvendigt, inden at man går i gang med egenkontrol og dataopsamling. Det specifikke formål kunne være at opnå viden om egen velfærdstatus med henblik på forbedring, certificering af produkter eller målrettet overvågning af dyrevelfærd.

Landmandens observationer i egenkontrollen er dyrebaserede indikatorer, som er væsentlige i forhold til vurdering af dyrevelfærd. Når vi sammenligner de fire anvendte dyrebaserede indikatorer med Welfare Quality® finder vi, at de delvist dækker 3 ud af 4 velfærdsprincipper med fokus på 'Good Health' med to dyrebaserede indikatorer men ingen dyrebaserede indikatorer, der beskriver 'Appropriate Behaviour'. Dog er de 4 indikatorer, der er valgt - Bevægelse, Renhed, Læsioner og Kropskonstitution –væsentlige i vurdering af dyrevelfærd, og kan anvendes til specifikke hensigter som fx at skabe opmærksomhed og sammenligning med eksterne auditører.

Et centralt spørgsmål i vurdering af egenkontroldata er, hvor nøjagtigt kvægbrugeren er i stand til at opnå det samme resultat som andre observatører (overensstemmelse). Grænserne for overensstemmelse i egenkontrol bør fastsættes af dem, der skal anvende observationerne til et specifikt formål. I denne rapport foreslår vi anvendelse af Coverage Probability, som er er et transparent udtryk for overensstemmelse, der let kan forstås. Beregning af Coverage Probability med et accept-niveau på +/- 5% point ved en sand prævalens på 10% i besætningen giver resultater imellem 6% og 28% for de 4 dyrebaserede indikatorer. Dette tolkes som: Hvis 10% af køerne har afvigende Kropskonstitution, så vil 28% af kvægbrugerne være i stand til at observere en prævalens imellem 5% og 15%. De Coverage Probabilities, vi finder, er så lave, at vi har svært ved at identificere et formål, hvor egenkontrolobservationerne er brugbare på besætningsniveau. Vi vurderer, at det er urealistisk at lave certificering, målrettet audit af dyrevelfærd eller benchmarking på besætningsniveau uden en betydelig forbedring af overensstemmelse. Det er muligt, at intensiv træning, kommunikation og auditering kan forbedre kvaliteten af egenkontroldata udført af kvægbrugeren. For fremtidig anvendelse af egenkontroldata på besætningsniveau forudsætter, at bedre overensstemmelse kan dokumenteres.

Analyserne viser, at benchmarking på gruppe- eller regionsniveau for tre af de dyrebaserede indikatorer - Bevægelse, Læsioner og Kropskonstitution - er mulig inden for rimelig præcision. Renhed udviser derimod stor systematisk bias (massiv underrapportering af kvægbrugerne), som gør denne indikator uanvendelig.

Der er regionale forskelle imellem regioner på besætningsgennemsnit af prævalenserne af Bevægelse, Læsioner og Kropskonstitution. Disse forskelle kan anvendes til at følge regionale udviklingstrends, men yderligere studier af validitet og/eller eksterne auditeringer er nødvendige for at øge nytteværdien.

Summary

Welfare assessment on herd level using animal-based measures is most commonly done by external personnel, which is considered a limit to implementation of welfare assessment schemes. An alternative approach is to use animal-based measures that are observed and recorded by the farmer. The purpose of this project is based on data recorded by a large number of dairy farmers to assess the reliability of data from self-assessment used for animal welfare assessment for the purpose of surveillance, certification and benchmarking. The empiric basis for this project is data from the quality assurance scheme Arlagaarden®Plus on the animal-based measures done quarterly by the farmers in United Kingdom, Netherlands, Luxembourg, Belgium, Germany, Denmark and Sweden. The incentive for collecting the Arlagaarden®Plus self-assessments was a compensation in the milk price.

In the report, we set forth that variation in animal welfare perceptions challenges the monitoring of dairy cow welfare, which should be fit-for-purpose. When perceptions vary, the purpose(s) may vary as well. We emphasize that well-defined specific purpose(s) are needed before embarking on self-assessment and data collection. The specific purposes could include local appraisal of animal welfare with the intent for farmers to improve, certify or target surveillance of animal welfare.

The observations recorded by the farmers in self-assessment are animal-based measures, which are important in animal welfare assessment. Using Welfare Quality® as a standard, the four animal-based measures partly cover three out of four welfare principles, with focus on the principle 'good health' by two animal-based measures and no animal-based measures that describes the welfare principle 'Appropriate behavior'. However, the four animal-based measures chosen - mobility, cleanliness, lesions and body condition – are important in evaluation of animal welfare and can be used for specific purposes e.g. raising awareness and comparison to external auditors.

A central question in evaluation of self-assessment data is how precisely farmers are able to achieve the same results as e.g. an auditor is (agreement). The variance allowed in self-assessment should be set by the one(s) to use the results given a specific purpose. In this report we suggest using the Coverage Probability as a very transparent measure of agreement, which can easily be understood. Estimating the Coverage Probabilities with a level of acceptability of +/- 5%-point at a true prevalence of 10% in the herd were between 6% and 28% for the four animal-based measures. It is interpreted as: Given a true prevalence in the herd of 10% of cows with low body condition, only 28% of the farmers are able to observe a prevalence between 5% and 15%. The observed Coverage Probabilities of all four animal-based measures are so low, that it is questionable for which purpose self-assessment observations can be useful on herd level. We find it unrealistic to do certification, targeted audits or herd-level benchmarking (comparing herds) without a considerable increase in agreement. It is likely that intensive training, communication and external audits can improve the quality of the self-assessment done by the farmers. Further studies on agreement/variance in the future are needed to demonstrate better agreement before the self-assessment data can be used on herd level.

We find that group level or regional benchmarking is possible for the three animal-based measures of Mobility, Lesions and Body Condition. Cleanliness demonstrated large systematic bias (consistent underreporting by the farmers) that invalidate the use of that animal-based measure. Regional differences in average herd level prevalences were found in the three animal-based measures of mobility, lesion and

body condition. These can be used to direct focus at specific regions and find risk factors that can be associated with the differences. We think that these regional estimates can be used to monitor regional development, but still more validation studies and/or external audits would be highly beneficial.

General introduction

Assessment of animal welfare on herd level is often based on animal-based measures done by external personnel. An alternative approach is to use animal-based measures observed and recorded by the farmer. The purpose of this project is based on data recorded by a large number of dairy farmers to assess the reliability in data from self-assessment used for animal welfare for the purpose of surveillance, certification and benchmarking. The empiric basis for this project is data from the quality assurance scheme Arlagaarden®Plus on the animal-based measures done quarterly by the farmers in Unted Kingdom, Netherlands, Luxembourg, Belgium, Germany, Denmark and Sweden. Implementation of welfare assessment based on animal-based measures have so far only been widely applied in commercial herds in fur animals (WelFur: https://www.sustainablefur.com/animal-welfare/) and based on external inspections. The incentive for collecting self-assessments in Arlagaarden®Plus is is a compensation of the milk price.

The project consists of three different components, each with its' own part (Part I to Part III) and finally an overall conclusion on the project across the three parts.

Part I: Self-assessment of animal welfare in dairy herds. In Part I, we characterize some conceptual relations between requirements for reliability and the intended purpose of the welfare assessment. Initially, the overall frame of welfare perception(s), possible stakeholders and the 'fit-for-purpose' concept is illustrated. Within the quality assurance system Arlagaarden®Plus, four animal-based measures are chosen and these are analyzed in terms of utility.

Part II: Assessment of the quality of farmers' observations of animal-based measures. Part II concerns analytical methods to assess the reliability of the self-assessment data provided by the farmers. The data used for this part were based on 83 herds within Arla Foods with both farmer self-assessment and an external auditor. Based on the results, the quality of the farmer self-assessment were evaluated for each of the four animal-based measures depending on potential purposes of doing the self-assessment.

Part III: Benchmarking farmers' cow assessment data. The objective of Part III is using the farmers' observations to monitor developments over time within region and to compare regions.

Part I: Self-assessment of animal welfare in dairy herds

Introduction

The objective of Part I was to describe the utility of measuring indicators of animal welfare for different purposes (monitoring, benchmarking and certification) with special emphasis on farmers' self-assessment. The paper is structured with a definition of different perceptions of animal welfare in order to set an overall frame relating to animal welfare. The next section deals with why and who should/could be interested in spending resources and efforts to improve animal welfare. The next part introduces fitness-for-purpose with a comparative aspect from diagnostic tests to address the importance of strict relations between what is observed and what is intended to be observed. We then introduce, what is currently measured in self-assessment of animal welfare within the quality control system of Arla Foods and the justification for doing these measures. Finally, we analyze the utility of the different measures for various purposes.

Perceptions of animal welfare

Animal welfare derives from animal ethics which have developed over centuries and focuses on the responsibility of farmers when they use animals for production. Multiple textbooks are written about animal ethics, but one key point is that the discussion often becomes philosophical. At the same time, it can also be challenging to operationalize the thoughts and definitions from animal ethics. From the 1960's to the 1980's there was an increasing focus on moving from not only animal protection, but also to focus on the animals' ability to express normal behavior (FAWC, 1979).

Based on thoughts and research in 1980's and 1990's, animal welfare evolved into a multidimensional entity, thereby emphasizing that animal welfare as a whole cannot be addressed by just one specific view on animals. Different people may have very different perceptions of animal welfare and it is important to –at least – realize this (Vanhonacker et al., 2008; Sandøe et al., 2011). Currently, three different perceptions on animal welfare exist that are used to conceptually define how different recordings/measurements or observations are relevant from the different views. The three perceptions are (in relation to dairy cows):

Naturalness: There is a focus on the cow having the possibility to act as they were living in the wild or fulfill the animal's natural habits. What exactly is meant by 'natural' is debatable but Yeates (2018) suggests looking at species-specific behavior in the closest populations of wild animals. The possibility to be on pasture, not being dehorned, living in naturally sized groups, interact with offspring and perform natural mating are events that are considered important characteristics of high welfare. Rollin (1993) is an exponent of this view of animal welfare. Often the welfare according to this view is measured by resource-based measures such as access to nature, feed, possibility to interact with offspring or to form naturally sized groups.

Functionality: Here, focus is on functionality. If the cow performs well, then she is considered to have good welfare. Broom (1996) defines this as the animal's ability to cope with the environment. It is a kind of performance measure of the cow, and people would look at milk production and reproductive rates as measures to access this perception of welfare. This perception of animal welfare is more prominent among people working directly with the cows (farmers/herdsmen). Measurements of functionality could be described as physiological measures and often include analyses of hormones like cortisol etc.

Emotional/affective state: In relation to this conception, focus is on what the animal feels or the experience of the animal. A central reference is Duncan (1993). This perception is the one used in most European

legislation concerning animal rights and welfare. It is important that an animal does not suffer from disease, but it should also not experience fear and anxiety. Positive emotions, that could be observed as play-behavior) are important as well. This is often measured by animal-based measures, which are observations on the individual animal for signs of disease, suffering or positive behaviour.

Fraser et al. (1997) combined the different perceptions of animal welfare, indicating that all three different mind sets of animal welfare somehow need to be addressed to give a comprehensive overall assessment of animal welfare. This is in line with the '5 freedoms' that also include all three perceptions of animal welfare (FAWC, 1979). A central point is that none of the perceptions of animal welfare can be considered superior to the others. It should be noted, that any one person hardly ever fit into one of these views of animal welfare and they are also overlapping. Lameness is an often used example, where lame cows will not be able to express their *natural* behavior like walking on pasture, they will not *function* very well because of reduced feed intake and subsequent reduced production, and lameness is the symptom of a condition that is painful, and the cow will then be judged to have poor welfare regardless of the perception of animal welfare. However, there will be areas where the concepts of good welfare will conflict, or where a measure is judged as positive by one view of welfare and negative by another view of animal welfare. One example is natural mating of cows on pasture. This results in pregnant cattle, which have to be slaughtered as part of production (the purpose of beef production is beef), and some may consider it unethical to slaughter the pregnant animal. Some measures are considered entirely irrelevant if one have a specific understanding of animal welfare, i.e. housing and other resource-based measures are normally considered irrelevant if you have an affective state point of view on animal welfare.

Interest in animal welfare

As mentioned above, different people can have different perceptions or views of animal welfare, just like society, governmental institutions, organizations and companies can have different reasons for being interested in animal welfare.

A basic requirement is food security, and the agricultural production in the European Union has developed to secure inexpensive and safe food since the 2nd World War. That said, the general society's interest in animal welfare will, as such, reflect the underlying distribution of the different perceptions of animal welfare. Most people in Western Europe have very limited contact with production animals and animal production systems. When faced with intensive production systems, they often will oppose to the confinement of the animals and will be directed to the naturalness perception of animal welfare.

Governmental institutions have the obligation of implementing the current legislation on animal welfare, and will subsequently assess if the welfare is as required. Legislation is the outcome of longer processes considering all aspects related to the effects of the legislation, balancing e.g. animal welfare, producer concerns, food security and national economy. The interest for the governmental institutions is to have clear and operational welfare measures that can be used for assess welfare with focus on what is acceptable and what is not.

Companies such as dairies and abattoirs also have an interest in animal welfare, because animal welfare could be an important attribute that could provide an advantage in the market, or retailers have specific demands, e.g. McDonald's in Germany requires that beef is not derived from pregnant cattle. Their interests are both external communication to the consumers but also assuring the quality of the produced product

from a welfare perspective. External communication can be thought of as providing solid information about management routines that are considered positive for animal welfare - like grazing - to more direct certification of animal welfare and monitoring development in key variables within animal welfare. The quality of the products concerning animal welfare could be to assure that no products were delivered from production sites with poor animal welfare.

Farmers' interest in animal welfare and observing animal welfare can be separated into use and non-use values (McInerney and Defra, 2004), where use-values are products and services from the animal production and non-use values are ethical values on treating animals. Focusing on the use-values, McInerney (2004) describes this in a conceptual framework, where he states that basic economic principles optimizing human benefit will drive the animal benefit (perceived animal welfare) towards what is, at minimum, acceptable by society (minimal welfare). This is illustrated in Figure 1, where increasing livestock production (moving left on the x-axis) will drive the animal welfare (y-axis) towards point D. This is an economy driven model that emphasizes that society have to set some limits on the production systems to safeguard animal welfare, like animal welfare acts. Legislation on animal welfare can be seen as moving the blue line, thereby increasing/decreasing what is socially acceptable. An example could be the milk-fed veal calves production systems, where calves are fed entirely on milk replacers and deprived of most bedding material. This is an illegal production system in Denmark, but a production elsewhere produces a product, for which there is an international demand.

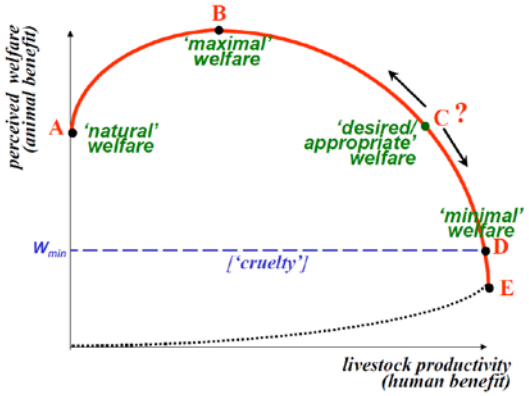

Figure 1: Conceptual relation between perceived welfare and livestock productivity (McInerney, 2004)

Farmers are not only interested in non-use values as demonstrated in Swedish dairy production by Hansson and Lagerkvist (2015). They identified 'avoidance of suffering' and 'doing the right thing' as important non-use values for the dairy farmers. This implies that farmers have dual interests in animal welfare and that farmers in general cannot just be seen as opponents to the rest of society concerning animal welfare.

Fit for purpose

Assessment of animal welfare relies on at least two aspects; how animal welfare is measured (an indicator) and how these measurements or observations are interpreted. Sandøe and Simonsen (1992) have discussed this in some detail with focus on the interpretation of animal welfare. Instead, we focus on the first part, trying to assess if an indicator can be considered generally valid. For this, we consider welfare indicators as a diagnostic test that is intended to provide information about something else that cannot be directly observed. Fit-for-purpose is a concept that has been adopted by the OIE in 2003 concerning evaluations of diagnostic tests (OIE, 2012 p 36). The idea is that the validity of a diagnostic test can only be assessed if the intended purpose of the indicator is known and precisely defined. For diagnostic tests, there are multiple

different purposes for infectious diseases like confirmation of diagnosis, eradication, certification etc. Often the same test can be used for different purposes, but its validity can be very different depending on the purpose. In Figure 2, the Justification – purpose – context – target-condition –rationale – utility complex is shown as a diagram.

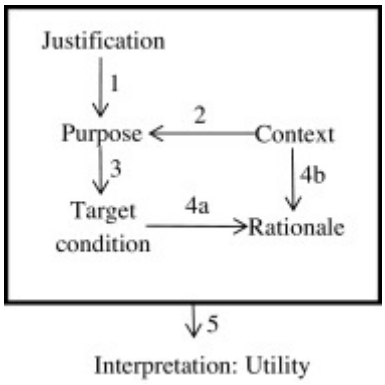

Figure 2: The justification – purpose – context – target – condition-rationale - utility complex (Kostoulas et al., 2017)

Figure 2 illustrates that the utility of systematically observing an entity is dependent on the initial justification, like concern for animal welfare. This leads to a purpose that could be to monitor or improve animal welfare and a target condition that could be a specific area where problems with animal welfare are more likely to occur. The context, like a specific production system and specific welfare problems associated with them, will also influence not only the purpose but also the subsequent target condition. The rationale for observing animal welfare thus needs to be evaluated within this context. Finally, the utility of observing animal welfare relies on the decisions that can be drawn from the result of this complex. This stresses the importance of a clearly defined purpose for doing observations that is closely related to the specific context in order to achieve the benefits of observation.

Self-assessment and its purposes

Self-assessment is the process of looking at one self – or judging our own achievements or progress within a specific field. However, the term 'self-assessment' is somewhat widely used and could be said to involve two different processes. The first process of self-monitoring by observing and recording a specific part of the production process, and a second process where these observations are evaluated or judged and possible actions are taken based on these. Often the two processes are not clearly defined and especially the last part can appear very subtle. Also, self-assessments can be performed for two different purposes. Internal self-assessment where the purpose is to learn about your own achievements (the animal welfare of the herd) and thereby setting a base-line where future progress can be measured against. This is in contrast to external self-assessment, where the results from the self-assessment are used by external parties for e.g. declaration of products, monitoring of the performance on higher levels. Strictly speaking, this is more of self-monitoring on herd level and self-assessment on higher levels, because that is where the assessment should take place. The self-monitoring often includes some kind of external audit of the self-assessment. Doing self-assessment and subsequent benchmarking is possible on both internally (between herds) and externally (between groups of herds).

In 2010, self-assessment of animal welfare was made mandatory in herds with pigs and cattle in Denmark. The focus within this self-assessment was compliance to the current legislation on animal protection and subsequently incorporated into production codes by the industry. The role of the farmer was to evaluate if

they complied with the legislation, but not to quantify aspects of animal welfare. The legislation was repealed in 2015. Lassen et al. (2012) did a quantitative study on the self-assessment of animal welfare, which revealed large differences between farmers in their attitude towards self-assessment.

In Germany, the Animal Welfare Act from 2014 requires that livestock farmers do on-farm self-assessments and guidelines for potential indicators, and protocols are available (Zapf et al., 2017). The stated purpose is mostly to raise self-awareness of the livestock keepers on animal welfare. Similarly, Arla Foods has initiated a voluntary self-assessment protocol, where the farmers are compensated for providing the results from self-assessment following a specific protocol. The purpose is stated to be the ability to actively work on improving dairy cow welfare. Both self-assessment schemes are based on predefined direct observations of individual or groups of cows, which is also known as animal-based measures.

Animal-based measures

Animal-based measures are central to measure the affective state. The key point is that observations are specifically done for welfare assessment (primary information source) and done directly on the individual animal. Secondary information sources or information gathered for other purposes have been shown to be relatively poor in predicting animal welfare (Otten et al., 2016). Some animal-based measures such as changes to the hair coat and skin are a result of the interactions between the animal and the local environment. The animal-based measures do however also have some limitations. Animal-based measures are observations directly on the animal on one particular day or point in time (cross-sectional in nature). This provides prevalence measures that can be followed over time and it may work very well when observing conditions that develop over a longer timer period and/or remain for some time (like months). However, it is quite poor in relation to the detection of severe conditions that only occurs with a very low prevalence and only for a very short time period. A number of different research projects have been conducted in Europe in the last decade to provide sound scientific basis for welfare assessments using animal-based measures in dairy cattle. The comprehensive EU funded Welfare Quality® project (2006-2010) included 13 European countries (Welfare Quality®, 2009). In principle, the Welfare Quality program for quantification of cattle welfare consists of approximately 30 measures that are aggregated into 12 criteria, further into 4 principles and finally to an overall assessment. The Welfare Quality® program was intended to be based entirely on animal-based measures, but did not entirely succeed, i.e. grazing and access to water and other resource-based measures were included in the assessment. The Welfare Quality® assessment protocol is also very time-consuming as demonstrated by (Knierim and Winckler, 2009) and hence not feasible for more practical purposes like self-assessments, but supply a detailed description of animal-based measures that could potentially be included in self-assessment.

Animal-based Measures in Arlagaarden®Plus

Arlagaarden®Plus is an extension of the current quality assurance scheme that also includes assessment of animal welfare. Today, animal welfare within Arlagaarden®Plus is observed using a combination of register data (like mortality) and animal-based measures as described below. The animal-based measures to be included were taken from the gross list of animal-based measures in Welfare Quality®. Welfare Quality® is based on external audit, so not all the animal-based measures were expected to be possible for farmers to do with limited training. Other aspects used for choosing the animal-based measures were acceptability of the measure by the farmer and expectations about prevalence and reliability. From a more practical point of view on implementation in >7.000 herds, it seems useful to start with a few animal-based measures that the farmers could easily relate to. If the initial results were promising then the list of animal-based measures could be expanded. The initial 4 animal-based measures, Body Condition, Mobility, Abrasions/Lesion and

Cleanliness were chosen to be included in Arlagaarden®Plus. These animal-based measures are all found to be valid, reliable and feasible for the purpose of describing (some of) the affective state of cattle (Winckler et al., 2003), even though "feasible" indicates feasible from a research perspective.

Body Condition: Body condition in Arlagaarden®Plus is recorded as normal, thin and very thin, whereas the condition score in the Welfare Quality® protocol is recorded as thin (0), normal (1) and fat (2). Low body condition scores indicate a prolonged situation, where the cow is in energy deficiency. They can be caused by disease, or be a result of a cow being provided with insufficient and/or low quality feed. Regardless of the cause (disease or hunger) it is problematic from a welfare perspective. The robustness of the body condition scoring was evaluated within Welfare Quality® (Leach et al., 2009b) and found to be useful for welfare assessment. Body condition scoring is commonly used in herd management, with a far more detailed scale, where robustness can be an issue (Kristensen et al., 2006). It is well known that body condition changes through lactation (e.g. Ferguson et al. (1994)), which could invalidate the use of body condition as a welfare measure. In this project, it is only the low (and very low) body condition that are recorded, e.g. conditions that exceed normal fluctuations in body condition. The associations between these low body conditions and milk production, reproduction, diseases are all negative (Roche et al., 2009).

Cleanliness: Measurements of cleanliness of the cows are included in the Welfare Quality® protocol. Cleanliness, soiled hair and skin may induce itching, reduce thermoregulatory capacity, be associated with superficial inflammation and reduce the microbiological defense (Winckler et al., 2003). Cleanliness reflects the cows' interaction with the local environment (and management), so cleanliness of the cows can be considered a superior measure to hygiene observations of the local environment. Leach et al. (2009a) reviews the associations between cleanliness and other infectious diseases and find positive association to mastitis. In a study of 201 dairy herds, Barkema et al. (1998) found that herds with more focus on hygiene generally had a lower bulk milk somatic cell count. In a subsequent study they concluded that herds with lower bulk milk somatic cell count were associated with a management practice that can be characterized as accurate, precise and clean (Barkema et al., 1999). Schreiner and Ruegg (2003) found that hygiene scores of the udder was associated with higher somatic cell scores and cows with higher udder hygiene score had more infection with major pathogens and hind leg hygiene was associated with udder hygiene. Similar results have been found under French conditions, where more than 1,000 cows were hygiene scored and an association between increasing composite hind-leg /udder hygiene scores and individual somatic cell counts (Reneau et al., 2005). In general, the studies documented associations between cow hygiene and somatic cell count, and the closer to the teats the dirt is, the more likely it is that the cow has as elevated somatic cell count.
Associations between cow cleanliness and claw lesions are also described. In a Swedish study (Hultgren and Bergsten, 2001) an experiment was conducted where cows were randomly allocated to different beddings (drained/not drained) in tie stalls. The results showed that cows on drained surfaces were both cleaner and had a lower prevalence of heel horn erosions and digital dermatitis. Similarly, a French trial found improved hind leg cleanliness to be a very beneficial factor for preventing digital dermatitis on herd level (Relun et al., 2013). Winckler et al. (2007) demonstrated that the prevalence of hind leg cleanliness was consistent over a 1-year period in dairy herds. This is in contrast with udder cleanliness that seem to fluctuate more on herd level.

Mobility*:* Mobility is assessed as deviations from the normal mobility pattern or simply as lameness. Lameness is an abnormal behavioral pattern that indicate a painful condition within the claws. Lameness is widely accepted as a major cow welfare issue (Von Keyserlingk et al., 2009). Multiple locomotive scoring systems exist but generally a fairly good correlation is found between the locomotive score and pathological

lesions within the claws at trimming (Winckler and Willen, 2001; Thomsen et al., 2012). In addition, multiple studies have been done to access the inter- and intra-observer agreement i.e. Thomsen et al. (2008), which find the observations robust. Despite these results, research also suggests that there could be an issue related to the sensitivity of the lameness scoring (Otten et al., 2013), which would indicate that the prevalence of lame cows are systematically underestimated. Numerous scientific articles demonstrate negative effects of lameness on milk production, fertility and culling are substantial. Huxley (2013) reviewed the economic impact of lameness and concluded that the economic loss associated with lameness is very substantial. The range of estimated economic losses are however also large, most likely due to differences in case definitions of lameness between studies. Since there is a positive association between milk production and the risk of lameness, estimates of production loss will most likely be underestimated.
Assessment of mobility in tie-stalls are however a challenge.

Lesions/Abrasions: Abrasions and lesions to the skin are results of the cows' interaction with the local environment. Abrasion where only the hair is missing is also included in the Welfare Quality® protocol which at least indicate that experts in the field agree on these as a welfare issue. Similarly, the majority of farmers presented with pictures on hock lesions agree that this would have some impact on production and welfare of the cow (Potterton et al., 2011). It has not been possible to find solid evidence that clearly states that hairless patches at the tarsus impair welfare significantly, which is also discussed by Knierim and Winckler (2009). Abrasions and lesions where the skin is excoriated or with other signs of inflammation like swelling will be associated with discomfort for the cows. A recent study with welfare benchmarking also included hock lesions (Trillo et al., 2017). A significant negative effect on milk production was found in cows with severe (medically treated) cases of hock lesions (Bareille et al., 2003).

Utility of the 4 animal-based measures

The utility of the four animal-based measures (mobility, cleanliness, lesions and body condition) should be assessed by their ability to describe cow animal welfare. Farm animal welfare can be defined in terms of four principles taken from Welfare Quality®: Good Feeding, Good Housing, Good Health and Appropriate behavior. These principles are then subdivided into 12 welfare criteria. The concept is visualized in Figure 3.

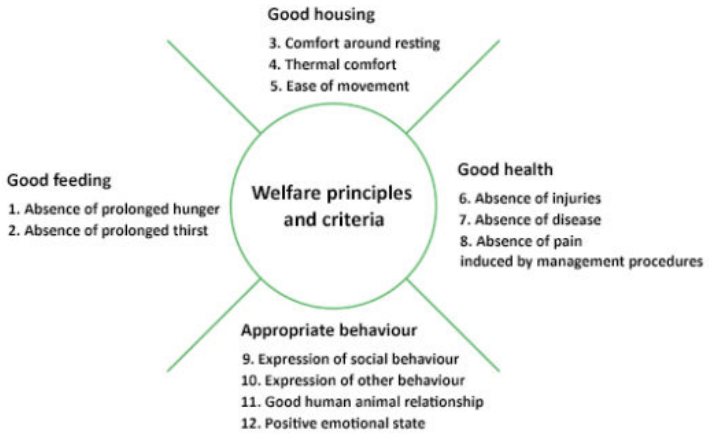

Figure 3: The four principles and 12 criteria that are the basis for the Welfare Quality® farm animal welfare assessment protocols (from https://www.luke.fi/ruokafakta/en/meat-and-fish/animal-welfare-and-health/)

The animal-based measures that are chosen for self-assessment only cover a minor part of the 31 measures that are used in the Welfare Quality® protocol. It is also obvious that the measures cannot cover all welfare aspects. However, the observation of body condition is an indicator of prolonged hunger and hence the principle of good feeding (Criterion 1). The criteria included in good housing are included by 1 animal-based measure, where cleanliness is an indicator of comfort around resting (Criterion 3). The measures of mobility and abrasions/lesions are both included in good health (Criterion 6). The criteria that are related to behavior like expression of social behavior, human-cows interaction and the positive emotional state of the cows are not included. Given that the purpose of doing self-assessment of animal welfare is external certification and declaration of animal welfare on milk, then the four animal-based measures will have some limitations, because they do not reflect the behavioral part of animal welfare. It could also be stated that focusing on Body Condition, Mobility and Lesions is actually a functionality perception of animal welfare even though the measures had an offset in Welfare Quality®, which should be based within affective state perception. There are a number of reasons for choosing these four animal-based measures, where expected prevalence of the measure is important, but also the acceptability of the measure by the farmer. Expanding the number of animal-based measures should focus on measures that can describe the social interactions between herd mates and people to include the behavioral part of animal welfare to provide a better coverage of animal welfare. But there are challenges in measures describing animal behavior; i.e. stereotypies and other abnormal behaviors low prevalent and only occurs for a short period of time.

This limitation of the chosen animal-based measures to fully describe dairy cow welfare is ealso reflected in monitoring both within and across herds as well as benchmarking; it is not actually monitoring welfare. It is self-assessment of 4 key indicators of dairy cow welfare. However, accepting that these animal-based measures do not cover dairy cow welfare, literature clearly acknowledges that these measures are important and fulfill the needed requirements for good indicators that could be used for monitoring and benchmarking at least within research.

Conclusions

- Variation in perceptions of animal welfare challenges the monitoring dairy cow welfare, which should be fit-for-purpose. When the perceptions vary, the purpose(s) may vary as well.
- Collection of data for use in self-assessment should address specific purposes. These could include local appraisal of animal welfare with the farmer's intent to improve, or governmental surveillance of animal welfare to penalize farmers that do not live up to specific requirements. There is a need to formulate the purpose prior to embarking on any self-assessment.
- Benchmarking is a powerful driver for change, both on individual farm level and internationally. It is very likely that one of the purposes of self-assessment of animal welfare is benchmarking.
- Four principles and twelve criteria are generally used for animal welfare monitoring in dairy cattle. They are measured using 31 animal-based measures. Timewise, it is often not possible to record these under auditing or self-assessment settings. Therefore, some can be selected. Here, four were selected: mobility, cleanliness, lesions and body condition. They cannot monitor all principles and criteria, but can be used for specific purposes e.g. raising awareness and comparison to external auditors.

Part II: Assessment of the quality of farmers' observations of animal-based measures

ABSTRACT

To apply self-assessment using animal-based measures for welfare monitoring, certification and benchmarking, a good agreement between the farmer's observation of these measures and what external persons (auditors) observe is necessary. The objective of this study was to estimate the agreement between farmers in the different animal-based measures used on individual (herd) and group (many herds) level. Eighty-three dairy herds in Denmark, Sweden, United Kingdom and Central Europe (Germany, the Netherlands, Belgium) where 4 animal-based measures were observed both by farmer and an auditor were used. The animal-based measures were mobility, cleanliness, lesions and body condition. Farmers and auditors should report the prevalence of abnormal animal-based measures after observation of all the cows in their herds. The data were analysed for agreement using Bland-Altman plot and Coverage Probability. Using an acceptability level of +/-5 percentage point (at a true prevalence of 10%) showed that individual agreement ranged from 6% to 28% of the farmers observations are within these limits. On average agreement, using an acceptability level of +/-2 percentage point at a 3% true prevalence demonstrated that the average agreement was acceptable for mobility and lesions. The individual Coverage Probabilities were deemed too low to be useful for the individual farmer to document the welfare performance of their own herd or to be useful for benchmarking on individual herd level at the specified level of acceptability. On average, agreement for mobility and lesion scoring were within the specified level of acceptability. Average agreement on cleanliness showed a large systematic bias, where farmers significantly underestimated the cleanliness and agreements were insufficient. More observations are needed to assess the average agreement on body condition, but the prevalence of abnormal body condition was so low that body condition most likely is not useful for self-assessment within the purposes described in this study. In general, the results suggest that far more training and follow up is needed to improve agreement.

INTRODUCTION

Application of self-assessment for animal welfare monitoring using animal-based measures requires good agreement between what the farmers observe in their own herd and what is observed by other (more trained) observers such as auditors. The question of our concern was: "Are the observations of the farmers and auditors close enough so that they can be used interchangeably?" This form of agreement is often called reproducibility (or reliability), because it is an assessment of how well different persons observe the same subject using the same method. This is in contrast to repeatability, where the same person observes the same subject several times (inter-observer-reliability). The variation arising from repeatability is often much smaller than the variation due to reproducibility.

Two different objectives and associated purposes can be identified when looking at self-assessment:

1) What is the agreement between one farmer and one auditor on the individual herd for the animal-based measures? If there is good agreement between farmer and auditor, it could be an opportunity for the farmer to document the welfare performance in their own herd. Quite similar to what is known for self-assessment schemes in the food production industry and restaurants. Benchmarking between farmers could also be an opportunity that could potentially improve animal welfare, because benchmarking has been shown to be a powerful driver for change.

2) What is the average agreement across multiple farmers and multiple auditors to describe a group of herd e.g. a region. This could be used for international benchmarking if there is good agreement on average, but it could also document changes in animal welfare nationally.

MATERIALS AND METHODS

Between December 10th and December 20th 2018, 83 dairy herds in Denmark, Sweden, United Kingdom and Central Europe (Germany, the Netherlands, Belgium) were assessed for animal welfare using animal-based measures. The farmers were chosen in a way ensuring that all regions within Arla Foods should be represented and their willingness to participate in the study. The farmers had done the animal welfare assessment themselves and auditors from the dairy Arla Foods subsequently assessed the animal welfare in the farm within two days. The animal-based measures observed were: mobility, cleanliness, lesions and body condition. Training of the farmers was limited to descriptions of different categories within each of the animal-based measure and small videos (around 4 minutes for each measure). The auditors had received more intensive training with a physical meeting and calibration.
All four animal-based measures were recorded on a scale including scores 0, 1, and 2. Score 0 described "normal", Score 1 was only a slight deviation from "normal" and Score 2 was severe deviation from "normal". Both farmer and auditor observed all the cows in the herd, including dry cows and cows in calving and sick pens. The total number of cows observed and the number of cows with Score1 and Score 2 were recorded. Based on these recordings, the prevalences of abnormal scores (Scores 1 and 2) for each of the animal-based measures were calculated within each herd.

The methodology to assess the agreement should support the data type and it must support the fact that the prevalences are pseudo-continuous by nature. The approach described by Bland and Altman (Altman and Bland, 1983) is commonly used for this purpose. Here the difference between two observations on the same subject is plotted against the average of these two observations. This is a graphical illustration, where important information about systematic differences between the two observers or changes in variation across the interval of observations can be detected, irrespective of the average recording. A central part in Bland-Altman plots is to access the degree of systematic bias. Systematic bias is the average difference between two methods applied to the same subject. The systematic bias is assessed with the Student's t-test (one-sample) to test if the systematic bias is different from zero using a significance level of 0.05. Based on the differences between the two methods, the limits of agreement were estimated. These express the 95% confidence interval on the individual observation. They should be interpreted as t a 2.5% chance of observing a difference between the two methods above the upper limit of agreement and a 2.5% chance of observing a difference below the lower limit of agreement, given that the actual true difference between the methods is zero. The limits of agreement then demonstrate how much variation between the two methods could be expected by chance.

There are numerous methods for estimating agreement; however, it is often more problematic to assess if a given agreement is acceptable. The question of acceptability should be based on *a priori* knowledge about what the results from the two methods should be used for or the purpose of observing the subject. Furthermore, this is usually a managerial issue, and should be defined by risk managers. Coverage Probability has been suggested to express acceptability numerically, because it is very simple and transparent (Barnhart et al., 2014). The Coverage Probability is defined as the probability that the absolute difference between two observations done on the same subject is less than a predefined difference of acceptability. The Coverage Probability then highlights the importance of having some knowledge of what we would expect to

acceptable, which is closely linked to the purpose of observing the subject. Central to the Coverage Probability is to define what differences are satisfactory depending on our purpose and is that defined by the manager.

RESULTS AND DISCUSSION

Descriptive statistics of prevalences of the abnormal scores in the four animal-based measures (mobility, cleanliness, lesions and body condition) can be seen in Table 1. The distributions of the differences in prevalence between auditor and farmer are also given. The differences are calculated as the auditor prevalence minus the farmer prevalence within each herd. The table shows that the median prevalence of abnormal scores in mobility was 3% regardless if the observations were carried out by the farmer or the auditor. This is also seen for the median of differences in prevalence between auditor and farmer, showing no median difference, but the 1st and 3rd quartile demonstrated some individual differences, where 25 pct. of the farmers were more than 2 pct. higher than the auditor and 25 pct. were more than 3 pct. lower than the auditor. Cleanliness showed a large systematic difference, where the median farmer prevalence was 6 pct. and the auditor prevalence was 27 pct.

Table 1. Descriptive statistics of the assessment prevalences from the 83 dairy herds. The numbers in the table describe the median and 1st and 3rd quartile in brackets of the individual herd prevalences.

	Herd Size	Mobility	Cleanliness	Lesions	Body Condition
Farmer	115 [64;182]	0.03 [0.01;0.06]	0.06 [0.03;0.17]	0.02 [0.00;0.04]	0.01 [0.00;0.04]
Auditor		0.03 [0.01;0.07]	0.27 [0.16;0.44]	0.04 [0.02;0.08]	0.01 [0.00;0.02]
Auditor/Farmer difference		0.00 [-0.02;0.03]	0.15 [0.05;0.25]	0.01 [-0.01;0.05]	0.00 [-0.02;0.01]

Body condition showed results similar to mobility with the same median prevalence and some large individual difference in prevalence between auditor and farmer. Lesions also demonstrated that the auditors observed a prevalence in relation to median of 4 pct. compared to 2 pct. observed by the farmers. Based on the initial descriptive statistics it seemed that farmers and auditors observed the same level of mobility, lesions and body condition, while cleanliness was severely underreported by the farmers compared to the auditors.

Bland –Altman plots

Figure 1 shows a classic Bland-Altman plot for mobility. In this plot, all observations were used to calculate the systematic bias, illustrated by the grey line and with the magenta lines delimiting the 95% confidence interval of the individual differences. An obvious outlier is observed, where there is a difference between the auditor and the farmer of more than 40 percent-points. An apparent distinct 'fanning pattern' is also observed where the observations deviates more and more from zero as the mean of the auditor and farmer observations increases. The grey horizontal line is the systematic bias between auditor and farmer

observations. However, this is only a valid description of the systematic bias, if the differences are randomly scattered. The other three animal-based measures demonstrated the same characteristics.
A modified Bland-Altman plot can be created by logit-transforming the auditor and farmer prevalence plotted on the y-axis. Then the results on the y-axis will be equal to the log of the odds ratio because:

$$logit\big(p(auditor)\big) - logit\big(p(farmer)\big) = \log\left(\frac{p(auditor)}{1-p(auditor)}\right) - \log\left(\frac{p(farmer)}{1-p(farmer)}\right)$$

$$= \log\big(odds(auditor)\big) - \log\big(odds(farmer)\big) = \log\left(\frac{odds(auditor)}{odds(farmer)}\right)$$

$$= \log(OddRatio)$$

In Figure 2, this modified Bland-Altman plot excluding the outlier can be seen. No apparent systematic pattern is seen, which indicates that the estimates of the systematic bias and limits of agreement are valid. The estimate of systematic bias is 0.188 on the log (odds ratio) scale, corresponding to an odds ratio of 1.2. The 95% confidence limit for the individual differences was +/- 1.9, which corresponding to an upper limit of agreement odds ratio of 6.8 and a lower limit odds ratio of 0.15.

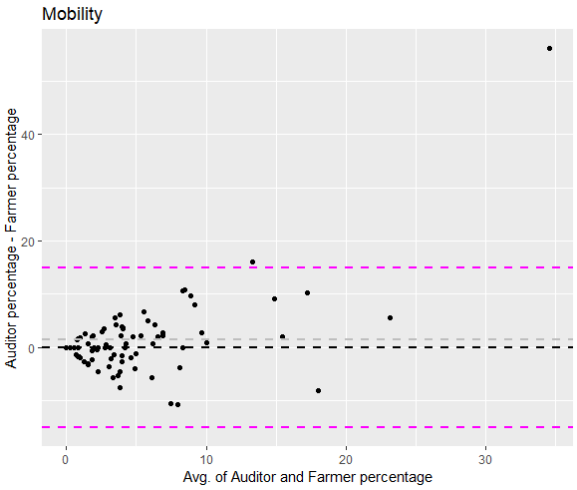

Figure 1: Classic Bland Altman plot with all observations of mobility

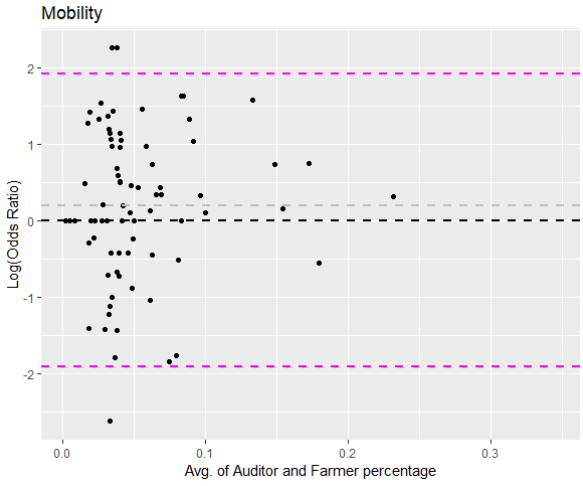

Figure 2: Modified Bland-Altman plot using logit-transform of the prevalences.

Expressing the systematic bias and limits of agreement in odds ratios has the consequence that the bias and limits of agreement increases proportionally with increasing auditor/farmer prevalence. The average of the auditor and farmer prevalences is an estimate of the unknown true prevalence. Since the odds ratio is a proportion between the auditor and farmer odds, there are multiple ways that farmer and auditor prevalences can result in the same odds ratio. What we really would like to know is the measurement error in terms of the original prevalence and, in this regard, the odds ratio is very difficult to interpret.

Instead of using the average of the auditor and farmer prevalence as an estimate of the true prevalence, we can plot the log (odds ratio) against the auditor prevalence. Then the estimate of bias and limits of agreement will remain unchanged, but we are able to estimate the bias on farmer prevalence and limits of agreement

based on the auditor prevalence. If the odds ratio (OR) is known and the auditor prevalence (p) is known then the bias as well as the limits of agreement can be estimated. In Table 2 bias and agreement limits are given for different auditor prevalences for mobility are given.

Table 2: Estimated farmer prevalence, bias and upper and lower agreement levels at different auditor prevalences for mobility

Auditor prevalence	**0.020**	**0.050**	**0.100**	**0.150**	**0.200**
Farmer prevalence	0.017	0.042	0.084	0.128	0.172
Bias	-0.003	-0.008	-0.016	-0.022	-0.038
Upper agreement level (95%)	0.122	0.263	0.430	0.545	0.630
Lower agreement level (95%)	0.003	0.008	0.016	0.025	0.036

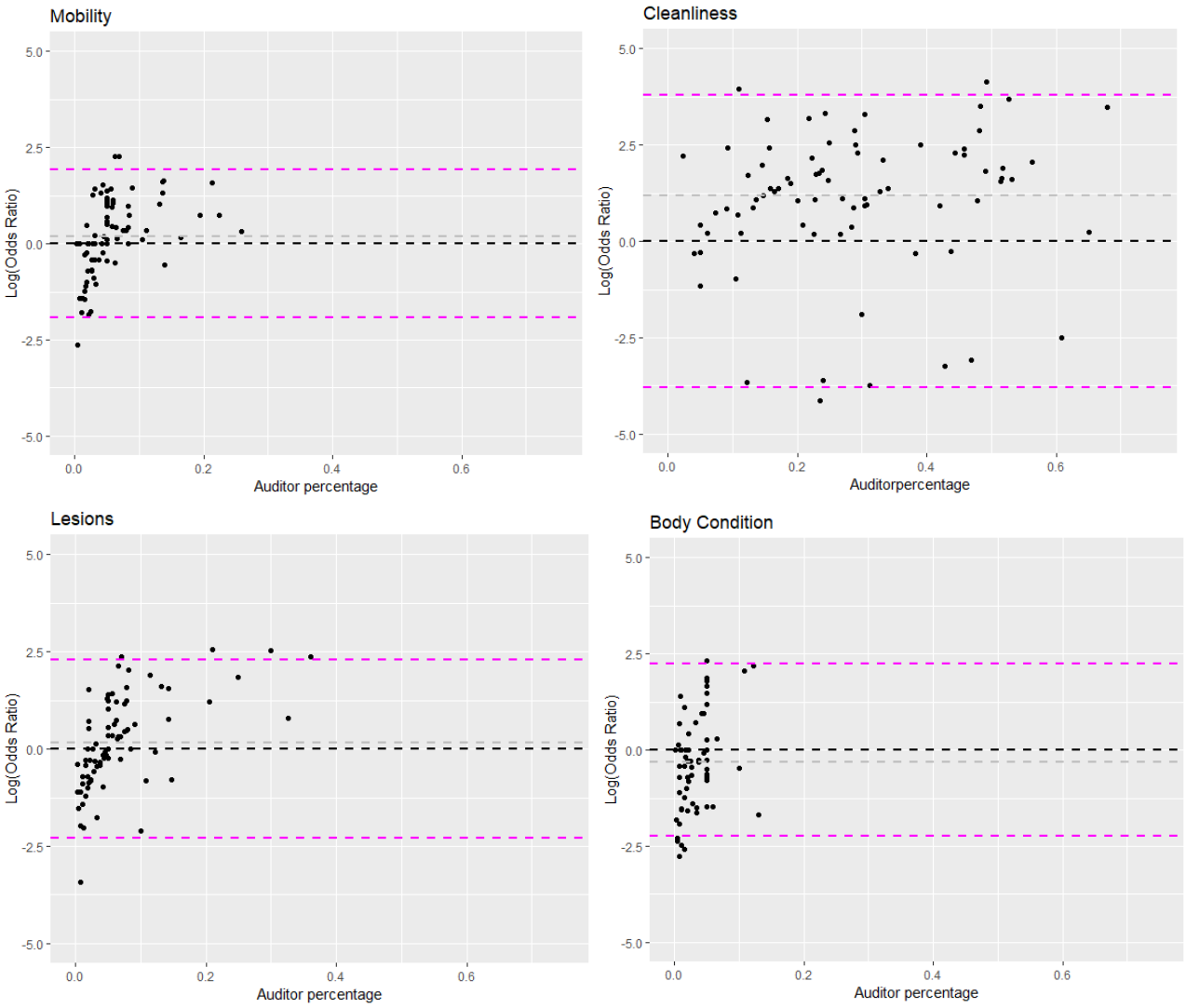

Figure 3: Bland-Altman plots using log(Odds Ratio) to assess the difference between auditor and farmer for mobility, cleanliness, lesions and body condition.

The results in Table 2 show that the bias is around +20% compared to the auditor prevalence. However, the limits of agreement are very wide. The interpretation of the limits of agreement would be that if the auditor prevalence is 10%, then the farmer prevalence is likely to be within the range of 2% and 43%.

Figure 3 shows that the bias for mobility, lesions and body conditions are close to 0, which indicate that on average the auditors and farmers were able to give comparable results. Farmers underestimate the prevalence of mobility and lesions and overestimate the prevalence of abnormal body condition. For cleanliness, the bias is more substantial, with 15 percent-point underestimated at an auditor prevalence of 10 pct. Table 3 shows the bias and upper and lower agreement levels for different auditor prevalence of cleanliness, lesion and body condition at relevant ranges. The systematic bias for mobility, cleanliness, lesions and body condition were estimated to 0.19, 1.18, 1.15, -0.31. Student's T-tests demonstrated that cleanliness ($P<0.0001$) and body condition ($P=0.014$) were significantly different from 0, which indicate that the auditors' and farmers' observations of these two animal-based measures on average differed beyond chance (95% confidence) and a systematic bias exists.

Table 3: Estimated farmer prevalence, bias and upper and lower agreement levels for different auditor prevalences for cleanliness, lesions and body condition

	Auditor prevalence	Farmer prevalence	Bias	Upper agreement level (95%)	Lower agreement level (95%)
Cleanliness	0.1	0.03	-0.07	0.83	0.003
	0.25	0.09	-0.16	0.93	0.007
	0.4	0.17	-0.23	0.97	0.01
	0.5	0.23	-0.27	0.98	0.02
	0.6	0.31	-0.29	0.99	0.03
Lesions	0.02	0.017	-0.003	0.17	0.002
	0.05	0.043	-0.007	0.34	0.005
	0.10	0.087	-0.013	0.52	0.01
	0.20	0.176	-0.024	0.71	0.02
	0.40	0.363	-0.037	0.87	0.06
Body Condition	0.01	0.014	0.004	0.09	0.001
	0.025	0.034	0.009	0.19	0.003
	0.05	0.067	0.017	0.33	0.006
	0.075	0.100	0.025	0.43	0.009
	0.10	0.132	0.032	0.51	0.012

Coverage Probability

To use the Coverage Probability we need to define what we consider an acceptable difference between farmer and auditor on one subject. As an example, we can state that for at auditor prevalence of 10 pct. we have acceptable agreement if the farmer is between 5 pct. and 15 pct. These limits are now converted into

the log (odds ratio) scale of -0.2 and 0.3. Table 4 demonstrates that the acceptability limits increase with increasing auditor prevalence.

Table 4: Acceptability limit for different auditor prevalences after defining the level based on +/- 5% at an auditor prevalence of 10%

Auditor prevalence	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
Upper acceptability prevalence	0.02	*0.05*	0.08	0.11	0.14	0.17	0.20	0.24	0.28	0.32
Lover acceptability prevalence	0.08	*0.15*	0.22	0.28	0.35	0.41	0.46	0.51	0.57	0.61

We can now estimate the Coverage Probability as the proportion of farmer and auditor odds ratios that are within this limit. On the original observed data the Coverage Probability is calculated, which can be visualized by adding the acceptability limits of -0.2 and 0.3 to the plots in figure 3. Then the Coverage probability will be the proportion of odds ratios inside this limit. Using an acceptability limit of +/- 5 percentage points at an auditor prevalence of 10% gives a Coverage Probability for mobility of 25%, cleanliness 6%, lesions 17% and body condition 28% when assessed by the farmers.

Another question is if the farmers and auditors agree on average. This can be useful if we want to provide a description of a group of herds. This is an evaluation of the systematic bias. In the traditional Bland-Altman plot the systematic bias is evaluated based on the confidence interval of the systematic bias. If the confidence interval include zero, then we would say that the two observers were equal on average. This is based on a Student's t-test, where the precision of the systematic bias estimate will increase with increasing sample size. If sample size is large enough, then it will always be possible to demonstrate a systematic bias with statistical significance. Systematic bias is however not necessarily a problem, because the assessment of agreement should be based on *a priori* understanding acceptability (Giavarina, 2015) equivalent to the Coverage Probability above. If the systematic bias is less than what is believed to be acceptable, then it is of minor importance, similar to setting the acceptability limits for the Coverage Probability. From Table 1 we can see the median prevalences are low, except for cleanliness. We then assume that we want an acceptability of +/- 2% at an auditor prevalence of 3%. This will be acceptability limit of -0.24 to 0.49 on the log(odds ratio) scale. In Table 5 the mean and upper and lower boundaries of the 95% confidence interval is given for mobility, cleanliness, lesions and body condition. The table shows that the confidence intervals for mobility and lesions are within the acceptability limits of -0.24 and 0.49 using 83 observations.

Table 5: Mean, upper and lower 95% confidence interval for the 4 animal-based measures calculated based on the 83 herds. Figures outside the acceptability limits of -0.24 and 0.49 are written in bold.

	Mobility	**Cleanliness**	**Lesions**	**Body Condition**
Upper confidence limit of mean	0.40	**1.60**	0.41	-0.06
Mean log(odds ratio)	0.19	**1.18**	0.16	**-0.31**
Lower confidence limit of mean	-0.03	**0.75**	-0.10	**-0.57**

Cleanliness is clearly outside the acceptability limits with the entire confidence limit and body condition is outside the acceptability limits for the mean and one confidence boundary. Based on this it will be possible

to use farmer observations of prevalence of mobility and lesions to get an average estimate that is within the acceptability limits using a sample of approximately 80 herds. For body condition it would perhaps be possible to get estimates that are within the acceptability level, but this would both require a new comparison study that demonstrate that the mean body condition is actually within the acceptability limits and a far larger sample of farmers to estimate the mean prevalence. Farmer prevalence of cleanliness is not likely to be useful for because of the large systematic bias, which cannot be accounted for.

CONCLUSION / PERSPECTIVES

The current sample of 83 dairy herds from in Denmark, Sweden, United Kingdom and Central Europe (Germany, the Netherlands, Belgium) that were observed both by farmer and auditor on 4 animal-based measures were analyses for agreement using Bland-Altman plot and Coverage Probability. The Coverage Probability on individual agreement range from 6% to 28%, when requiring a level of acceptability +/- 5 percentage point when the auditor observed 10%. These Coverage Probabilities are deemed too low to be useful for the individual farmer to document the welfare performance of their own herd or to be useful for benchmarking on individual herd level using the selected levels of acceptability. Cleanliness showed a large systematic bias (median difference 16 percentage points), where farmers severely underestimate the cleanliness. On average, using an acceptability level of +/-2 percentage point at a 3% auditor prevalence, demonstrated that the average agreement was acceptable for mobility and lesions. Not acceptable for cleanliness and perhaps not acceptable for body condition. The prevalence for abnormal body condition is so low that this animal-based measure most likely not is useful for self-assessment within the purposes described in this study.

The low coverage probabilities of the farmers' assessments suggests that systematic training and follow up is needed for the farmers to provide more precise estimates of the prevalences of the animal-based measures. The fact that the farmers assessment of mobility and lesions are on average close to the auditors' point to farmers having problems both with under-reporting and over-reporting these animal-based measures.

Part III: Benchmarking farmers' cow assessment data

ABSTRACT

Farmers' observations of animal-based measures of mobility, lesions and possibly body conditions are found to be within an acceptable level on average (regional/group) compared to external auditors. Cleanliness was not found to be acceptable both on herd and regional level. The objective of this study was to use the farmers' observations to monitor developments over time within region and to compare regions. Seven quarters of observations were included in this analysis from November 2017 to June 2019, with a total of 55,384 recordings on herd level. Herds were included from Sweden, Denmark, United Kingdom, Belgium, Netherlands, Luxembourg and Germany. Belgium, Netherlands, Luxembourg and Germany were grouped together as Central Europe. The first three assessment periods were removed from the analyses due to a systematic decrease in all three animal-based measures. Significant differences between regions were found, where mobility was lowest in Sweden and Denmark and both were lower than United Kingdom and Central Europe; lesions were less prevalent in United Kingdom compared to the other regions and for body condition, the prevalence of thin cows was lower in Central Europe compared to the other regions. Further analysis is needed to find possible reasons for the regional difference found, that can be associated with both management and production systems. Based on these results we find that results are best used to monitor developments within region until precision in the individual farmers' observations is dramatically improved.

INTRODUCTION

Benchmarking or simple comparisons of welfare performance between groups or individuals are stated as the major benefits of doing animal welfare assessments (Whay, 2007). As concluded in Part II of this report none of the four animal-based measures were assessed sufficiently precise by the farmers to be useful on the individual (herd) level given the acceptability levels of +/- 5 percentage points at a true prevalence of 10%. Though benchmarking has been demonstrated to be very successful in motivating farmers (Nir, 2003), the imprecision on the prevalence of the individual herds makes it problematic to do individual herd benchmarking. For group level benchmarking the evaluation in Part II concluded that mobility and lesions could be used for group level benchmarking, body condition scoring perhaps could be included, whereas cleanliness could not due to large systematic bias.

The objective of this study was to use the farmers' observations to monitor developments over time within region and to compare regions.

MATERIALS AND METHODS

Within Arlagaarden®Plus, which is a supplement to the mandatory milk quality assurance scheme, four animal-based measures should be observed quarterly on all cows in the herd. The animal-based measured observed were: mobility, cleanliness, lesions and body condition. Training of the farmers were limited to descriptions of different categories within each of the animal-based measure and small videos (around 4 minutes for each measure). All four animal-based measures were recorded on a 0, 1, 2 scale. Score 0 was characterized as "normal", Score 1 was only a slight deviation from "normal" and Score 2 was severe deviation from "normal". The total number of cows observed and the number of cows with Scores 1 and 2 were recorded in a central database. Based on these recordings, the prevalence of abnormal scores (Scores 1 and 2) for each of the animal-based measures were calculated within each herd. Seven quarters of observations were included in this study from November 2017 to June 2019, with a total of 56,099 recordings on herd level. Herds were included from Sweden, Denmark, United Kingdom, Belgium, Netherlands,

Luxembourg and Germany. Belgium, Netherlands, Luxembourg and Germany were grouped together as Central Europe. For the included assessment periods no external audit system were yet in place, so assure some kind of quality control of the assessment data. Also no validation to took place at submission to the database. Hence it was possible to achieve invalid registrations (like prevalences above 1 and herd sizes above 10000 cows). After initial data cleaning 55384 observation were kept for further analysis.

RESULTS AND DISCUSION

There was a slight increase in number of herds included from 7,412 in November 2017 to 8,196 included in June 2019. The proportions of observation in each region were 23% in United Kingdom, 24% in Central Europe, 25% in Sweden and 28% in Denmark. The average of the individual herd prevalences across the regions and time period can be seen in Table 6. The average prevalence of mobility ranged from 3.2% in Sweden to 7.0% in United Kingdom; Lesions from 3.7% in United Kingdom to 5.6% in Denmark. Body Condition ranged from 3.5% in Sweden and United Kingdom to 5.3% in Central Europe.

Table 1: Average prevalences of mobility, lesions and body conditions in the four Arla Foods regions between November 2017 to June 2019

Region	Mobility	Lesions	Body Condition
United Kingdom	0.070	0.037	0.035
Central Europe	0.062	0.054	0.053
Denmark	0.041	0.056	0.036
Sweden	0.032	0.055	0.035

Figure 1 shows the average prevalence for mobility, lesion and body condition in each region. For all three animal-based measures the prevalence decreased in the first 3 sampling periods and then appeared to be more stable. In the results we focus on the last 5 assessment periods, because we assume that the first 3 assessment periods were influenced by introduction to Arlagaarden@Plus. The vertical bars on each observation indicate the 95% confidence interval of that average. If the confidence intervals are separated, there is also a statistical difference between them. So based on statistics of the farmers observations of mobility then the prevalence on mobility in Sweden is significantly lower than in Denmark and both regions are significantly lower than United Kingdom and Central Europe in the last 5 assessment periods. For lesions, United Kingdom is significantly lower than the other three regions and for body condition Central Europe is significantly higher than the other three regions.

The average prevalence of abnormal mobility is here below 10% on herd level. In a study from England and Wales, Barker et al. (2010) found an average herd level prevalence of 35.8% (range from 0 to 79%) on a scale similar to the one used here. Lameness prevalence across 101 Swedish herds were found to be 5.1% (Manske et al., 2002). Recent prevalence studies of lameness are however also limited especially on herd level.

There appears to be a cyclical pattern in lesions with more lesions winter/early spring compare to summer and fall. For lesions, it is well known (i.e. Regula et al. (2004)) that the prevalence of hock lesion are lower in loose housing systems compared to tie stalls. Also in loose housing systems there is a positive (lower prevalence) with increased grazing time (Burow et al., 2012). This is also reflected in the lesions scores where United Kingdom has the lowest average prevalence, which corresponds to the very low number of tethered

cows in United Kingdom. A review of hock lesion prevalence demonstrate large differences in prevalence depended on region, type of hock lesion (hair loss, ulceration and swelling) but in one of the studies included they found cow-level prevalences below 8% (Kester et al., 2014).

Low body condition was found to be at an average herd level prevalence of 2% for United Kingdom, Sweden and Denmark, whereas the prevalence in Central Europe 2-3% points higher. There is very limited literature available to compare these prevalences to.

Comparing the prevalences from this study to the prevalences in the validation study (Part II) shows that the prevalences here were 1-2% higher than in the validation study.

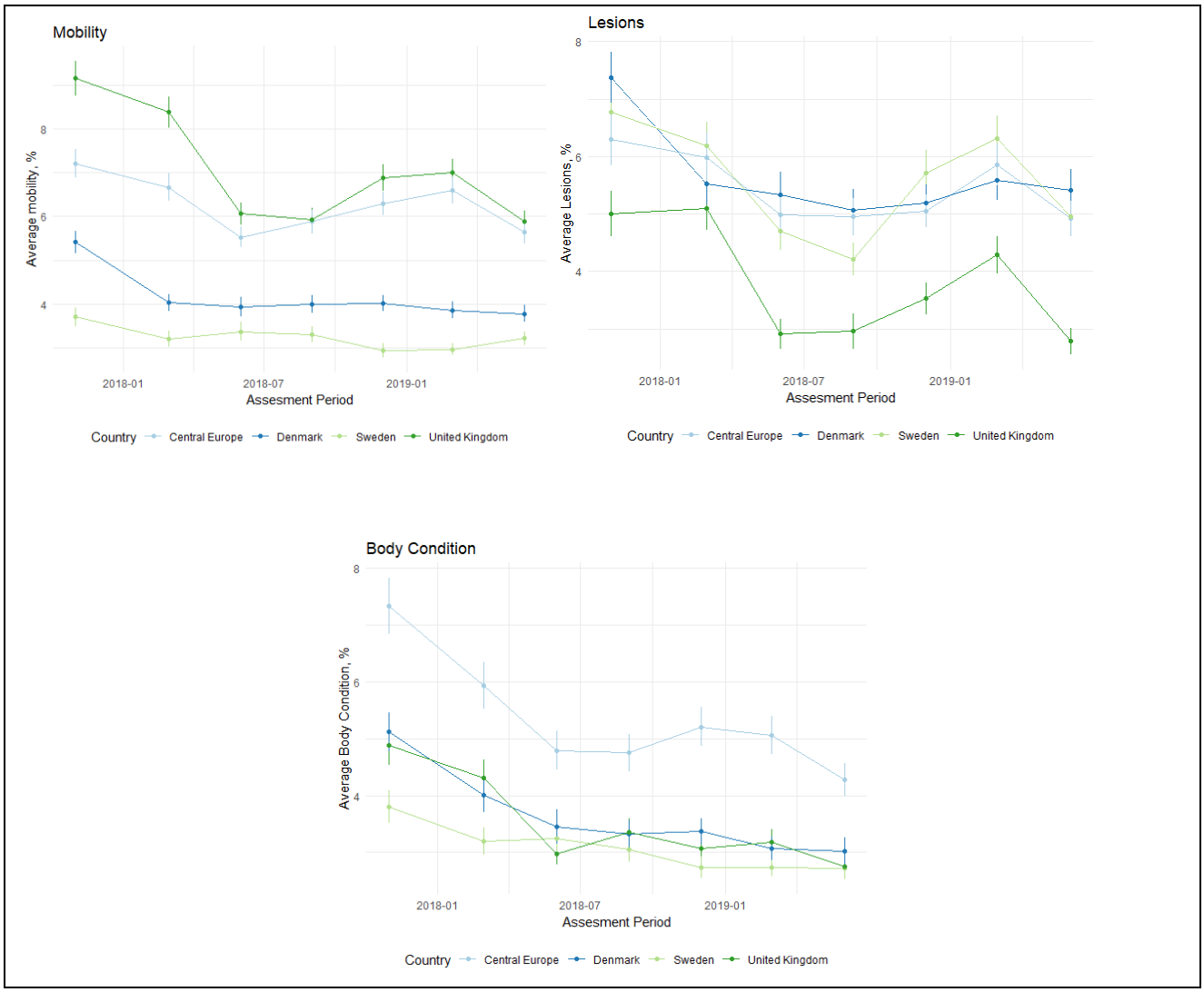

Figure 1: Average prevalence at the 7 assessment periods in the four regions for mobility, lesions and body condition. The vertical bars represent the 95% confidence interval for the average prevalence.

CONCLUSION / PERSPECTIVES

Regional differences in the herd level average prevalences of the animal-based exist. These can be used to focus efforts on future improvement in the three animal-based measures. However, deeper analysis is needed to understand the possible association to regional differences in management and production

systems that can influence the prevalences. Comparison of the prevalences to literature is difficult due to differences in scales and limited number of recent estimates of prevalences at herd level. The differences in prevalences between this and the previous study of 1-2 % on average, adds to the uncertainty about the population included in validations study and possibly biased results. At the current state of Arlagaarden@Plus, we suggest that the results are most useful to monitor developments within regions.

Conclusions

- Variation in perceptions of animal welfare challenges the monitoring dairy cow welfare, which should be fit-for-purpose. When the perceptions vary, the purpose(s) may vary as well.

- Collection of data for use in self-assessment should address specific purposes. These could include local appraisal of animal welfare with the intent from a farmer to improve, or governmental surveillance of animal welfare to penalize farmers that do not live up to specific requirements. There is a need to formulate the purpose prior to embarking on any self-assessment.

- Animal-based measures are important in assessment of animal welfare. Using Welfare Quality as a standard the four animal-based measures partly cover three out of four principles, with focus on disease/injury by two animal-based measures and no animal-based measures that describes the welfare principle related to behavior.

- The four animal-based measures chosen were mobility, cleanliness, lesions and body condition are important in evaluation of animal welfare and can be used for specific purposes e.g. raising awareness and comparison to external auditors.

- Limits on reliability in self-assessment (the farmers ability to observer an animal-based measure just as well as some external auditor) should be set by the one(s) that should use the results given a specific purpose. We suggest using the Coverage Probability as a very transparent measure of agreement, which can easily be understood.

- Estimating the Coverage Probability with a level of acceptability of +/- 5%-point at a true prevalence of 10% in the herd were between 6% and 28%. It is interpreted as: Given a true prevalence in the herd of 10% of cows with low body condition only 28% of the farmers are able to observe a prevalence between 5% and 15%.

- The observed Coverage Probabilities of all four animal-based measures are so low, that it is questionable for which purpose self-assessment observations can be useful on herd level. We find it unrealistic to do certification, targeted audits or herd-level benchmarking (comparing herds) without a considerable increase in reliability of the self-assessment.

- It is likely that intensive training, communication and external audits can improve the quality of the self-assessment data. Further studies on agreement in the future are needed to demonstrate better agreement until the self-assessment data can be used on herd level.

- Group level or regional benchmarking is however possible for the three animal-based measures of Mobility, Lesions and Body Condition. Cleanliness demonstrated large systematic bias (consistent underreporting by the farmers) that invalidate the use of that animal-based measure.

- Regional differences in average herd level prevalences were found in the three animal-based measures of mobility, lesion and body condition. These can be used to direct focus to specific regions and find risk factors that can be associate with the differences. We think that these regional estimates can be used to

monitor regional development, but still more validation studies and/or external audits are needed and would be highly beneficial.

REFERENCES

Altman, D.G., Bland, J.M., 1983. Measurement in medicine: the analysis of method comparison studies. Journal of the Royal Statistical Society: Series D 32, 307-317.

Bareille, N., Beaudeau, F., Billon, S., Robert, A., Faverdin, P., 2003. Effects of health disorders on feed intake and milk production in dairy cows. Livestock Production Science 83, 53-62.

Barkema, H., Schukken, Y., Lam, T., Beiboer, M., Benedictus, G., Brand, A., 1998. Management practices associated with low, medium, and high somatic cell counts in bulk milk. J Dairy Sci 81, 1917-1927.

Barkema, H., Van der Ploeg, J., Schukken, Y., Lam, T., Benedictus, G., Brand, A., 1999. Management style and its association with bulk milk somatic cell count and incidence rate of clinical mastitis. J Dairy Sci 82, 1655-1663.

Barker, Z.E., Leach, K.A., Whay, H.R., Bell, N.J., Main, D.C.J., 2010. Assessment of lameness prevalence and associated risk factors in dairy herds in England and Wales. J Dairy Sci 93, 932-941.

Barnhart, H.X., Yow, E., Crowley, A.L., Daubert, M.A., Rabineau, D., Bigelow, R., Pencina, M., Douglas, P.S., 2014. Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. Statistical Methods in Medical Research 25, 2939-2958.

Broom, D.M., 1996. Animal welfare defined in terms of attempts to cope with the environment. Acta Agr Scand a-An, 22-28.

Burow, E., Thomsen, P.T., Rousing, T., Sørensen, J.T., 2012. Daily grazing time as a risk factor for alterations at the hock joint integument in dairy cows. Animal : an international journal of animal bioscience 7, 160-166.

Duncan, I.J.H., 1993. Welfare is to do with what animals feel. Journal of Agricultural and Environmental Ethics 6, 8-14.

FAWC, 1979. Farm Animal Welfare Council Press Statement.

Ferguson, J.D., Galligan, D.T., Thomsen, N., 1994. Principal descriptors of body condition score in Holstein cows. J Dairy Sci 77, 2695-2703.

Fraser, D., Weary, D.M., Pajor, E.A., Milligan, B.N., 1997. A scientific conception of animal welfare that reflects ethical concerns. Anim Welfare 6, 187-205.

Giavarina, D., 2015. Understanding Bland Altman analysis. Biochem Med (Zagreb) 25, 141-151.

Hansson, H., Lagerkvist, C.J.J.F.P., 2015. Identifying use and non-use values of animal welfare: Evidence from Swedish dairy agriculture. 50, 35-42.

Hultgren, J., Bergsten, C., 2001. Effects of a rubber-slatted flooring system on cleanliness and foot health in tied dairy cows. Preventive veterinary medicine 52, 75-89.

Huxley, J.N., 2013. Impact of lameness and claw lesions in cows on health and production. Livest Sci 156, 64-70.

Kester, E., Holzhauer, M., Frankena, K.J.T.V.J., 2014. A descriptive review of the prevalence and risk factors of hock lesions in dairy cows. 202, 222-228.

Knierim, U., Winckler, C., 2009. On-farm welfare assessment in cattle: validity, reliability and feasibility issues and future perspectives with special regard to the Welfare Quality (R) approach. Anim Welfare 18, 451-458.

Kostoulas, P., Nielsen, S.S., Branscum, A.J., Johnson, W.O., Dendukuri, N., Dhand, N.K., Toft, N., Gardner, I.A., 2017. STARD-BLCM: Standards for the Reporting of Diagnostic accuracy studies that use Bayesian Latent Class Models. Preventive veterinary medicine 138, 37-47.

Kristensen, E., Dueholm, L., Vink, D., Andersen, J., Jakobsen, E., Illum-Nielsen, S., Petersen, F., Enevoldsen, C., 2006. Within-and across-person uniformity of body condition scoring in Danish Holstein cattle. J Dairy Sci 89, 3721-3728.

Lassen, J., Jensen, K.K., Thorslund, C., 2012. Egenkontrol af dyrevelfærd. Rapport nr. 215. Department of Food and Resource Economics, University of Copenhagen.

Leach, K., Knierim, U., Whay, H., 2009a. Cleanliness Scoring for Dairy and Beef Cattle and Veal Calves. Welfare quality report 11, 25-30.
Leach, K., Whay, H., Knierim, U., 2009b. Condition Scoring for dairy and beef Cattle and veal Calves. Welfare quality report 11, 1-6.
Manske, T., Hultgren, J., Bergsten, C., 2002. Prevalence and interrelationships of hoof lesions and lameness in Swedish dairy cows. Preventive veterinary medicine 54, 247-263.
McInerney, J., 2004. Animal welfare, economics and policy. Report on a study undertaken for the Farm Animal Health Economics Division of Defra 68.
Nir, O., 2003. What are production diseases, and how do we manage them? Acta Vet Scand, 21-32.
Otten, N.D., Rousing, T., Houe, H., Thomsen, P.T., Sørensen, J.T.J.A.W., 2016. Comparison of animal welfare indices in dairy herds based on different sources of data. 25, 207-215.
Otten, N.D., Toft, N., Houe, H., Thomsen, P.T., Sørensen, J.T., 2013. Adjusting for multiple clinical observers in an unbalanced study design using latent class models of true within-herd lameness prevalence in Danish dairy herds. Preventive veterinary medicine 112, 348-354.
Potterton, S.L., Green, M.J., Millar, K.M., Brignell, C.J., Harris, J., Whay, H.R., Huxley, J.N., 2011. Prevalence and characterisation of, and producers' attitudes, towards hock lesions in UK dairy cattle. Veterinary Record.
Welfare Quality®, 2009. Welfare Quality® assessment protocol for cattle. Welfare Quality® Consortium, Lelystad, Netherlands.
Regula, G., Danuser, J., Spycher, B., Wechsler, B.J.P.v.m., 2004. Health and welfare of dairy cows in different husbandry systems in Switzerland. 66, 247-264.
Relun, A., Lehebel, A., Bruggink, M., Bareille, N., Guatteo, R., 2013. Estimation of the relative impact of treatment and herd management practices on prevention of digital dermatitis in French dairy herds. Preventive veterinary medicine 110, 558-562.
Reneau, J.K., Seykora, A.J., Heins, B.J., Endres, M.I., Farnsworth, R.J., F. Bey, R., 2005. Association between hygiene scores and somatic cell scores in dairy cattle. Journal of the American veterinary medical association 227, 1297-1301.
Roche, J.R., Friggens, N.C., Kay, J.K., Fisher, M.W., Stafford, K.J., Berry, D.P., 2009. Invited review: Body condition score and its association with dairy cow productivity, health, and welfare. J Dairy Sci 92, 5769-5801.
Rollin, B.E., 1993. Animal welfare, science, and value. Journal of Agricultural and Environmental Ethics 1993.
Sandøe, P., Christensen, T., Forkman, B., Lassen, J., 2011. Definitioner af og holdninger til dyrevelfærd. Dyrevelfærd I Danmark 2010. Ministeriet for Fødevarer, Landbrug og Fiskeri, 14-25.
Sandøe, P., Simonsen, H., 1992. Assessing animal welfare: where does science end and philosophy begin? J Animal welfare 1, 257-267.
Schreiner, D., Ruegg, P., 2003. Relationship between udder and leg hygiene scores and subclinical mastitis. J Dairy Sci 86, 3460-3465.
Thomsen, P.T., Munksgaard, L., Sørensen, J.T., 2012. Locomotion scores and lying behaviour are indicators of hoof lesions in dairy cows. The Veterinary Journal 193, 644-647.
Thomsen, P.T., Munksgaard, L., Tøgersen, F.A., 2008. Evaluation of a Lameness Scoring System for Dairy Cows. J Dairy Sci 91, 119-126.
Trillo, Y., Quintela, L.A., Barrio, M., Becerra, J.J., Peña, A.I., Vigo, M., Garcia Herradon, P., 2017. Benchmarking welfare indicators in 73 free-stall dairy farms in north-western Spain. Veterinary Record Open 4.
Vanhonacker, F., Verbeke, W., Van Poucke, E., Tuyttens, F.A.J.L.s., 2008. Do citizens and farmers interpret the concept of farm animal welfare differently? 116, 126-136.
Von Keyserlingk, M., Rushen, J., de Passillé, A.M., Weary, D.M., 2009. Invited review: The welfare of dairy cattle—Key concepts and the role of science. J Dairy Sci 92, 4101-4111.
Whay, H., 2007. The journey to animal welfare improvement. J ANIMAL WELFARE 16, 117.

Winckler, C., Brinkmann, J., Glatz, J., 2007. Long-term consistency of selected animal-related welfare parameters in dairy farms. Anim Welfare 16, 197-199.
Winckler, C., Capdeville, J., Gebresenbet, G., Hörning, B., Roiha, U., Tosi, M., Waiblinger, S., 2003. Selection of parameters for on-farm welfare-assessment protocols in cattle and buffalo. Anim Welfare 12, 619-624.
Winckler, C., Willen, S., 2001. The Reliability and Repeatability of a Lameness Scoring System for Use as an Indicator of Welfare in Dairy Cattle. Acta Agriculturae Scandinavica, Section A — Animal Science 51, 103-107.
Yeates, J.J.A., 2018. Naturalness and animal welfare. 8, 53.
Zapf, R., Schultheiß, U., Knierim, U., Brinkmann, J., Schrader, L., 2017. Tierwohl messen im Nutztierbestand–Leitfäden für die betriebliche Eigenkontrolle. LANDTECHNIK–Agricultural Engineering 72.