Identification of dairy herds with animal welfare problems

Otten, Nina Dam

FACULTY OF HEALTH AND MEDICAL SCIENCES
UNIVERSITY OF COPENHAGEN

Identification of dairy herds with animal welfare problems

PhD Thesis 2014

Nina Dam Otten

Identification of dairy herds with animal welfare problems

PhD Thesis
Nina Dam Otten
2014

Department of Large Animal Sciences
Faculty of Health and Medical Sciences
University of Copenhagen

Principal Supervisor:
Professor Hans Houe
Department of Large Animal Sciences, Faculty of Health and Medical Sciences
University of Copenhagen

Co-Supervisors:
Professor Nils Toft
National Veterinary Institute, Technical University of Denmark

Senior Scientist Peter T. Thomsen
Department of Animal Science, Aarhus University

Professor Jan Tind Sørensen
Department of Animal Science, Aarhus University

Assessment Committee:
Associate Professor Ilka C. Klaas (Chairman)
Department of Large Animal Sciences
Faculty of Health and Medical Sciences, University of Copenhagen

Professor David C.J. Maine
School of Veterinary Sciences, University of Bristol

Professor Lene Munksgaard
Department of Animal Science, Aarhus University

The thesis has been submitted to the Graduate School of the Faculty of Health and Medical Sciences, University of Copenhagen

Submitted 8th of November 2013
Printed by: SL Grafik, Frederiksberg C, Denmark (www.slgrafik.dk)
ISBN: 978-87-7611-705-4

Contents

Preface

This project was funded from three different sources: i) Project "Dyrevelfærd I husdyrbesætninger: et samspil mellem husdyrbruger og myndigheder" funded by The Danish AgriFish Agency under the Danish Ministry of Food, Agriculture and Fisheries, ii) Faculty of Life Sciences, University of Copenhagen, and iii) the Research School for Animal Production and Health (RAPH). The project was carried out as collaboration between the University of Copenhagen and Aarhus University in 2009-2013.

During my enrolment in the Ph.D. curriculum I have had the pleasure to be working with very dedicated supervisors and co-workers both at the Department of Large Animal Sciences (University of Copenhagen) and the Department of Animal Science (Aarhus University).

First of all, I would like to express my sincere gratitude to my main supervisor, Professor Hans Houe, for continuously supporting, challenging and acknowledging my work. His expertise and guidance has brought me through these three years without any major crises. I am also very grateful to my co-supervisors. I wish to thank Professor Nils Toft for constructive dialogues and criticism, and especially for his patience and pedagogics when giving statistical first aid. I wish to thank Senior Researcher Peter Thomsen for his everlasting optimism and his great ability to keep focus on the important matters and to put matters into relevant perspectives. I am also grateful to Professor Jan Tind Sørensen for sharing his strategic advice and putting matters into a greater perspective.

Much of my Ph.D. work was done in collaboration with a number of dedicated people. First of all, my sincere gratitude goes to all the farmers, who participated in the project. Without their kind collaboration, this project would not have been possible. I would like to express my gratefulness to Senior Researcher Tine Rousing, Associate Professor Karsten Klint Jensen, Elke Burow and Kristian Knage-Rasmussen for fruitful discussions, good team spirit and great social times. Within the Group of Population Biology, sincere gratitude has to be sent to Jeanne – for always helping out with a positive and reassuring attitude; Professor Søren Saxmose Nielsen – "the Oracle" in any matter; and Professor Liza Rosenbaum Nielsen for valuable advice, help and cooperation throughout different projects. I would also like to thank Marta Bo Almskou from the Knowledge Centre for Agriculture, Cattle, for her patient help regarding data set extractions from Danish Cattle Database.

My fellow Ph.D. students, Sine, Karina, Iben, and Marshal, need to be thanked for sharing great laughs, good lunches, deep conversations and helping out when needed. Good luck to You all!

Special thanks to my family and the three great men in my life, my son, my husband and my father – You pick me up, inspire me, keep me going and make me laugh - this work is dedicated to You.

Fredriksberg, November 2013

Nina Dam Otten

List of abbreviations

AIC – Akaike information criterion

AUC – area under the curve

AWI – Animal welfare index

DCD – Danish Cattle Database

DPR – differential positive rate

ECM – energy corrected milk

OR – Odds ratio

PABAK – prevalence-adjusted bias-adjusted kappa

PPV – positive predictive value

ROC – receiver operating characteristics

SCC – somatic cell count

SD – standard deviation

Se – sensitivity

Sp – specificity

WHO – World Health Organisation

List of papers

Paper I: Otten, N.D., Toft, N., Thomsen, P.T., Houe, H., Sørensen, J.T. 2013. Evaluation of the performance of register data as predictors for dairy herds with high lameness prevalence. Resubmitted to *Animal*

Paper II: Otten, N.D., Toft, N., Thomsen, P.T., Houe, H. 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(3-4): 348-354.

Paper III: Otten, N.D., Rousing, T., Thomsen, P.T., Houe, H., Sørensen, J.T. 2013. Quantification of animal welfare in dairy herds using different sources of data. Submitted to *Animal Welfare*

Paper IV: Otten, N.D., Nielsen, L.R., Thomsen, P.T., Houe, H. 2013. Register-based predictors of violations of animal welfare legislation in Danish dairy herds. Submitted to *Animal*

Summary

Raised awareness and public concern for animal welfare has led to a variety of animal welfare assurance and assessment schemes for both commercial and legislative purposes. However, full scale on-farm welfare assessments are costly due to time-consumption. Hence, there is a need for cheap and feasible methods for identifying livestock herds with animal welfare problems.

The aim of the research presented in the present PhD thesis was to develop and evaluate a method for the identification of Danish dairy herds with high levels of welfare related disorders. The models were based on different levels of information available ranging from cheap secondary data to expensive on-farm collected data, trying to answer the hypotheses: i) cross-sectional findings (i.e. clinical observations) can be classified based on data from existing databases (**pilot study**, paper 1); ii) welfare indicators based on direct clinical observations can describe underlying welfare problems in a highly valid manner (**observer effect study**, paper 2); iii) dairy herds with animal welfare problems can be identified without visiting the farm (**index study**, paper 3 and **prediction model study**, paper 4). In order to answer the hypotheses a literature study on the definitions of animal welfare and welfare assessments both in general and for dairy cattle, in particular, was performed. The review showed, although by a limited number of publications, that register data could be used to identify herds with welfare problems. Furthermore, the aggregation of welfare measures into an overall score showed a need for a transparent, but yet valid method. This led to the methodological approaches presented in the research of the four papers included in this thesis.

The pilot study was performed as a cross-sectional study with register-based follow-up investigating the diagnostic potential of register data variables to classify herds with high lameness prevalences (≥ 16%) in 40 Danish dairy herds. Register variables significantly associated with high lameness prevalence used in the pilot study were mortality, bulk tank somatic cell count (SCC), proportion of lean cows at slaughter and the standard deviation (SD) of age at first calving. Variables were evaluated at an optimized and a pre-defined cut-off. Mortality and SD of age at first calving showed highest sensitivity (Se) at the optimized cut-off with Se=100% with a specificity (Sp) of 53% and 23%, respectively. SD of age at first calving also showed highest Se=80% at the pre-defined cut-off. However, the quantification of the diagnostic potential showed highest area under the curve (AUC) for the receiver operating characteristic (ROC) curve for mortality with an AUC=0.76; and adding more variables to the prediction model did not significantly improve the AUC of the given model. The pilot study showed the importance of using optimised cut-offs in order to enhance the accuracy of the identification of herds with high lameness prevalence.

Cross-sectional findings from clinical scorings in 80 Danish dairy herds were used in the second study to evaluate the validity of clinical scorings exemplified by lameness, as clinical scorings can be prone to misclassification bias, especially in unbalanced study designs. Four trained observers performed on-farm scorings in an unbalanced study design. Observers were trained at on-farm sessions and calibrated using 39 video sequences. Inter-observer agreement estimated based on the calibration material showed good agreement. Modelling observers' sensitivity (Se) and specificity (Sp) in a latent class model (LCA) to estimate the true lameness prevalence (TP) showed general underestimation of TP due to low Se (24-81%). A Bayesian risk factor model was established to evaluate grazing as a risk factor for both the apparent (AP) and true prevalence (TP). Grazing turned out as a risk factor for the AP due to the effect introduced by the observer characteristics (Se/Sp) in this unbalanced distribution of observers among grazing and non-grazing herds. Using the TP no effect of grazing on the lameness prevalence was found.

Paper 3 investigated the correlation of different animal welfare indices (AWI) based on different levels of information, i.e. register data, resource measure and animal-based variables. AWI's were calculated for 73 Danish dairy herds based on a weighted linear aggregation model, where measure weights were derived from expert opinion. Spearman rank correlation coefficients showed significant agreement in the ranking of herds for the AWI based on register data variables for a 180 day period and the AWI based on direct animal measures. Combining the AWI for register variables 180 day period and for resource measures also showed significant correlation with the direct animal-based AWI. However, no correlations were found between the animal-based and resource-based AWI's. The index study confirmed that the different measure groups (i.e. register-, resource- and animal-based measures) evaluate different aspects of animal welfare and hence, direct comparison of indices is not applicable. Thus, paper 4 used a different approach to investigate predictors for impaired animal welfare. Based on the current official animal welfare inspections the same 73 herds (index study) were assessed for violations of current animal welfare legislation (VoAWL). A total of 32% (N=23) of the included herds showed VoAWL. Register variables also included in the AWI in the index study were evaluated as predictors for herds having VoAWL. The final prediction model consisted of the standard deviation in average milk yield for second lactation cows, a high bulk tank somatic cell count (≥ 250,000 cells/mL), and low levels of recorded veterinary treatments (≤ 25 treatments/year).

In conclusion, the results from the PhD work show that selected register data variables can be used to identify herds with potential welfare problems due to high levels of welfare related disorders. However, the welfare of a given herd cannot be estimated based on register data alone as they only serve as indicators for certain aspects of animal welfare when aggregated into an overall welfare index. Hence, individual register data indicators do hold potential as screening tests enhancing the efficiency of risk-based sampling schemes for a targeted animal welfare control. Future research on optimizing the utilisation of these indicators e.g. surveillance systems is necessary to ensure a valid identification of problem herds.

Sammendrag

Øget opmærksomhed og bekymring blandt offentligheden omkring dyrevelfærd har ført til udviklingen af en bred vifte af både kommercielle såvel som lovmæssige velfærds garanti- og vurderingsprogrammer. Men omfattende velfærdsvurderinger i besætninger er yderst tidskrævende og dermed dyre at gennemføre, hvorfor der er et behov for mindre omkostningstunge og tilgængelige metoder, der kan udpege produktionsdyrsbesætninger med velfærdsproblemer.

Formålet med det nærværende ph.d. projekt var således at opstille og evaluere en operationel og valid metode til denne identifikation. De undersøgte modeller var baseret på forskellige informationsniveauer med data kilder som billige sekundære data, de såkaldte register data, og dyre velfærdsvurderinger udført ved besætningsbesøg. Data blev anvendt til at besvare følgende hypoteser: i) resultater af tværsnitsundersøgelser (f.eks. kliniske observationer) kan klassificeres på baggrund af data fra eksisterende databaser (**pilot studie**, manuskript 1); ii) velfærdsindikatorer baseret på direkte kliniske observationer reflekterer de underliggende velfærdsproblemer på valid manér (**observatør effekt studie**, manuskript 2); iii) malkekvægsbesætninger med dyrevelfærdsproblemer kan identificeres uden at besøge besætningen (**index studie**, manuskript 3 og **prædiktionsmodel studie**, manuskript 4). For at kunne besvare disse hypoteser, gennemførtes et indledende litteraturstudie omhandlende velfærdsdefinitioner og generelle aspekter vedrørende velfærdsvurderinger og de kvægspecifikke velfærdsvurderinger. Litteraturstudiet viste, at register data kunne bruges til at identificere kvægbesætninger med velfærdsproblemer, dog på baggrund af et begrænset antal undersøgelser. Aggregeringsmodeller for velfærdsindikatorer til én samlet score viste sig at kræve en gennemskuelig og valid tilgang. Dette resulterede i de metodiske tilgange præsenteret i de inkluderede studier i denne afhandling.

Pilot studiet var udformet som et tværsnitsstudie med follow-up af register data. Studiet undersøgte register datas diagnostiske potentiale til at klassificere besætninger med høj forekomst af svær halthed (≥ 16%) i 40 danske malkekvægsbesætninger. Der blev fundet signifikante sammenhæng mellem høj halthedsforekomst og register variablerne: dødelighed, tank celle tal, proportionen af magre dyr ved slagt og spredning i kælvningsalder. Variablerne blev evalueret ud fra hhv. en optimeret og en prædefineret tærskelværdi. Dødelighed og spredning i kælvningsalder havde højeste sensitivitet (Se) ved brug af den optimerede tærskelværdi med en Se=100% og en Sp på henholdsvis 53% og 23%. Spredning i kælvningsalder havde endvidere også højest Se (80%) ved brug af den prædefinerede tærskelværdi. For at kvantificere det diagnostiske potentiale af disse variable gennemførtes analyser af receiver operating characteristic (ROC) kurver og arealet under denne kurve (AUC). Dødelighed viste det største AUC=0.76, hvorimod ingen yderligere inkluderede variable kunne øge dette areal signifikant.

Pilot studiet illustrerede vigtigheden i at bruge optimerede tærskelværdier for at øge præcisionen af identifikationen af besætninger med høj halthedsforekomst.

Resultater for kliniske registreringer i en tværsnitsundersøgelse med 80 danske malkekvægsbesætninger blev undersøgt i det efterfølgende studie, for at evaluere validiteten af kliniske scoringer eksemplificeret ved halthedsprævalens. Kliniske scoringer kan være behæftet med misklassifikations bias, hvilket bliver yderst fremtrædende i studie designs, der har en skæv fordeling af observatører. I dette studie indgik fire trænede observatører, der havde modtaget træning i besætninger og deltaget i en kalibreringsøvelse baseret på 39 videosekvenser. Inter-observatør overensstemmelse udregnet på baggrund af kalibreringsdata viste god overensstemmelse. For at estimere den sande halthedsprævalens (TP), blev observatørernes Se og Sp modelleret i en latent klasse model (*LCA*), hvilket resulterede i en generel undervurdering af den sande prævalens på grund af de lave Se estimater (Se=24-81%) blandt observatørerne. For at evaluere effekten af afgræsning på hhv. den tilsyneladende (*apparent prevalence*=AP) og den sande prævalens (*true prevalence* =TP), blev en Bayesiansk risiko faktor model opstillet. Modellen viste afgræsning som risiko faktor for AP, som følge af den effekt, som observatørerne bidrog med i form af deres individuelle Se og Sp i den uens fordeling af observatører mellem afgræsnings- og ikke-afgræsningsbesætninger. Afgræsning havde ingen effekt på TP.

I manuskript 3 undersøgtes korrelationen mellem forskellige dyrevelfærdsindekser (Animal Welfare Index=AWI), baseret på forskellige informationsniveauer (register data, ressource mål, dyre-baserede mål) for i alt 73 danske malkekvægsbesætninger. Indeksene blev baseret på en vægtet additiv (lineær) aggregeringsmodel, hvor velfærdsmålene blev vægtet i henhold til ekspertvurderinger. Graden af overensstemmelse af rangeringen af besætninger indenfor indeksene blev vurderet med Spearmans rank koefficient. Der blev fundet signifikant overensstemmelse mellem indekset baseret på register data variabler for en 180 dages periode og indekset baseret på dyre-baserede variabler. En kombination af det førnævnte indeks og ressource-baserede mål viste også signifikant overensstemmelse med det dyre-baserede indeks. Der kunne ikke påvises korrelation mellem de ressource-baserede og det dyre-baserede indeks. Dette indeks studie understregede, at de forskellige informationsniveauer (register-, ressource- og dyre-baserede mål) hver især evaluerer forskellige aspekter af dyrevelfærd, hvilket umuliggør en direkte sammenligning af indeksene. Dette fund førte til en anden tilgang, beskrevet i manuskript 4, hvor register data blev undersøgt som prædiktorer for nedsat dyrevelfærd på besætningsniveau. Med udgangspunkt i det samme datasæt som i indeks studiet, blev de 73 besætninger undersøgt for brud på gældende dyrevelfærdsforeskrifter (Violations of Animal Welfare Legislation =VoAWL). I alt viste 32% (N=23) af besætningerne sådanne brud (VoAWL). Register variablerne, som var inkluderet i indeks studiet, blev i dette studie evalueret som prædiktorer for VoAWL. Den endelige prædiktionsmodel omfattede spredning i ydelse blandt anden laktationskøer, et højt tank celletal (≥ 250.000 celler/ml) og lavt antal af årligt registrerede dyrlægebehandlinger (≤ 25 behandlinger pr. år).

Således kunne dette ph.d. studie konkludere, at udvalgte register data variable kan bruges i identifikationen af besætninger med potentielle velfærdsproblemer som følge af høj forekomst af velfærdsrelaterede lidelser. Velfærdstilstanden i en given besætning kan dog ikke direkte bestemmes baseret på register data, da disse kun afdækker enkelte områder af velfærd, når de samles i et indeks. Register data må derfor anses som indikatorer, som individuelt kan agere som screening test og dermed øge effektiviteten af risiko-baserede udpegningssystemer for målrettet kontrol af dyrevelfærd. Fremtidig forskning burde derfor fokusere på at optimere udnyttelsen af disse register data indikatorer i f.eks. overvågningssystemer for at kunne retfærdiggøre en valid udpegning af problem besætninger.

Zusammenfassung

Eine verstärkte Aufmerksamkeit und Sorge der Öffentlichkeit um das Wohlbefinden der Tiere in der Nutztierhaltung hat zu der Entwicklung von zahlreichen kommerziellen und gesetzlichen Garantie- und Bewertungsprogrammen geführt. Jedoch sind diese umfassenden Bewertungen des Wohlbefindens mit einem immensen Zeitaufwand verbunden und daher äußerst kostspielig in ihrer Ausführung. Deshalb besteht ein Bedürfnis, günstigere und verfügbare Methoden zu erstellen, um Betriebe mit problematischen Zuständen identifizieren zu können.

Das Ziel dieser Ph.D. Studie war daher die Erstellung und Bewertung von operationellen und validen Methoden zur Identifikation von Problembetrieben. Die bewerteten Modelle wurden jeweils auf verschiedene Informationsquellen aufgebaut – von billigen sekundären Daten, den sogenannten Registerdaten, zu den teuren, in Betriebsbesuchen eingesammelten Gesundheits- und Wohlbefindensdaten. Die Daten wurden zur Beantwortung folgender Hypothesen benutzt: i) Resultate von Querschnittsuntersuchungen (z.B. klinische Observationen) können auf Grund von Daten in existierenden Datenbasen klassifiziert werden (**Pilotstudie**, Manuskript 1); ii) Klinische Observationen als Indikatoren des Tierwohlbefindens beleuchten den zugrunde liegenden Allgemeinzustand des Wohlbefindens der Tiere auf eine valide Weise (**Beobachterstudie**, Manuskript 2); iii) Milchviehbetriebe mit problematischen Niveaus des Tierwohlbefindens können ohne Betriebsbesuch identifiziert werden (**Indexstudie**, Manuskript 3 samt **Prädiktionsmodelstudie**, Manuskript 4). Einleitend wurde ein Literaturstudium durchgeführt, um die Erhellung der Hypothesen zu ermöglichen. Hier wurden insbesondere generelle Aspekte der Definitionen für Tiergesundheit und Wohlbefinden und deren Bewertung, sowie die für Rinder spezifischen Beurteilungen untersucht. Ergebnisse wiesen darauf hin, wenn gleich in einer geringen Anzahl veröffentlichter Studien, dass Registerdaten zu der Identifizierung von Problembetrieben geeignet sind. Des Weiteren erwies sich, dass Aggregationsmodelle der Tierwohlbefindensindikatoren in einen gesamt Zusammenhang einen durchschaubaren und validen Zugang erfordern. Die Befunde des Literaturstudiums resultierten in den methodischen Erwägungen, die in dieser Abhandlung präsentiert werden.

Die **Pilotstudie** enthielt eine Querschnittsstudie mit einen Follow-up von Registerdaten. In dieser Studie wurde das diagnostische Potential von Registerdaten zur Klassifizierung von Betrieben mit hohen Lahmheitsfrequenzen (≥ 16%) in 40 Dänischen Milchbetrieben untersucht. Hohe Lahmheitsfrequenz war signifikant mit den folgenden Registervariablen assoziiert: Mortalität, Tank-Zellzahl, Anteil magerer Tiere beim Schlachten und die Standard- abweichung des Kälbungsalters. Die Variablen wurden an Hand von optimalen und prädefinierten Schwellenwerten evaluiert. Mortalität und Standard Abweichung des Kälbungsalters erzielten die höchsten Sensitivität Werte (*Se*) von 100% durch den optimierten Schwellenwert mit einer Spezifizität (*Sp*) von beziehungsweise 53% und 23%. Die Abweichung des Kälbungsalters erzielte die höchste Se (80%) durch den prädefinierten Schwellenwert. Um das diagnostische Potential zu quantifizieren, wurden Receiver Operating Characteristic (ROC) Kurven erstellt, sowie das Areal unter dieser Kurve (AUC) analysiert. Mortalität erreichte das größte AUC (0.76), und Erweiterungen des Models mit weiteren Variablen ergaben keine signifikanten Arealsteigerungen. Die Pilotstudie illustrierte die Bedeutung und Wichtigkeit von optimierten Schwellenwerten, um die Präzision der Identifikation von Betrieben mit hoher Lahmheitsfrequenz zu sichern.

Die Ergebnisse von klinischen Registrierungen in 80 dänischen Milchviehbetrieben wurden in einer Querschnittstudie erhoben und in der darauf folgenden Studie durch das Beispiel von Lahmheitsprävalenz auf ihre Validität geprüft. Klinische Registrierungen können durch Misklassifikation beeinflusst werden, welches besonders in Studien-Designs mit verschobener Verteilung von mehreren Beobachtern verstärkt zum Ausdruck kommt. In der **Beobachterstudie** wurden vier geübte Beobachter bei Betriebsbesuchen trainiert und anhand von 39 Videosequenzen kalibriert. Daten der Kalibrierungsübung zeigten eine gute Übereinstimmung zwischen den Beobachtern. Um die wahre Prävalenz (*true prevalence*=TP) zu ermitteln, wurden Se und Sp Werte der Beobachter in einer Latent-Klassen-Analyse (*latent class analysis*= LCA) modelliert. Diese Analyse wies auf eine generelle Unterschätzung der wahren Prävalenz hin, welches den niedrigen Se Werten der Beobachter (Se=24-81%) zuzuschreiben war. In einem Bayesianischen-Risiko-Faktor-Model wurden die Effekte von Weidegang auf die jeweilige wahre und scheinbare Prävalenz (*apparent prevalence*= AP) hin analysiert. Hier wurde Weidegang auf Grund der Beobachter Se und Sp Werte als Risiko für die scheinbare Lahmheitsprävalenz identifiziert, ein Effekt, der durch die ungleiche Verteilung von Beobachtern in Weide- und konventionellen Betrieben verstärkt wurde. Weidegang hatte keinen Effekt auf die wahre Prävalenz.

In der **Indexstudie** wurde die Korrelation zwischen verschiedenen Wohlbefindens-Indexen (*Animal Welfare Index*=AWI), basierend auf verschiedenen Informationsniveaus (Registerdaten, Systemdaten und klinische und Verhaltensdaten), in 73 Milchviehbetrieben untersucht. Die Indexe basierten auf einem gewichteten additiven (linearen) Aggregationsmodell, in dem die Wohlbefindens Indikatoren anhand von Expertpanel erstellter Gewichte evaluiert wurden. Der Übereinstimmungsgrad der Ranglistenplatzierung der Betriebsresultate wurde durch den Spearmans rank Koeffizient beurteilt. Eine signifikante Übereinstimmung wurde zwischen dem Registerdaten-Index für eine 180-Tage-Periode und dem Index aus klinischen Observationen und Verhaltensdaten gefunden. Eine Kombination zwischen dem vorher erwähnten Registerdaten-Index und dem System-Index wies auch eine signifikante Übereinstimmung auf. System-Index und Index aus klinischen Observationen und Verhaltensdaten zeigten keine Übereinstimmung. Diese Studie unterstreicht, dass die verschiedenen Informationsniveaus (Registerdaten, Systemdaten und klinische und Verhaltensdaten) jeweils abgegrenzte Bereiche des Tierwohlbefindens beschreiben und dadurch einen direkten Vergleich der Indexe nicht ermöglichen. Diese Befunde führten zu dem, in der **Prädiktionsmodellstudie** beschriebenen, Zugang, in welchem die Registerdaten auf ihr prädiktives Potenzial für einen auf Herdenniveau beeinträchtigten Wohlbefinden evaluiert wurden. Dieselben 73 Betriebe der Indexstudie wurden nun auf Verstoße gegen geltende Haltungs- und Tierrechtsvorschriften (*violations of animal welfare legislation*=VoAWL) überprüft. Insgesamt 32% (N=23) Betriebe erwiesen sich als positive VoAWL-Betriebe. Registerdaten der Indexstudie wurden nun erst individuell und danach in Kombination untersucht. Das Prädiktive Endmodell umfasste die Standardabweichung in der Milchleistung der Kühe in der zweiten Laktation, hohe Tank-Zellwerte (≥ 250.000 Zellen/mL) und eine niedrige Anzahl registrierter veterinärer Behandlungen (≤ 25 Behandlungen pro Jahr).

Diese Ph.D. Studie zeigte, dass auserwählte Registerdaten zur Identifikation von Betrieben mit potenziell problematischem Tierwohlbefinden, verursacht durch eine hohe Belastung von auf das Wohlbefinden bezogener Leiden, durchaus angewandt werden können. Das Wohlbefinden einer Herde kann jedoch nicht durch eine Aggregation dieser Registerdaten bestimmt werden, da ein solcher Index nur Teilaspekte beleuchtet. Daher müssen Registerdaten als Indikatoren angesehen werden die, individuell in einen Screening-Test angewandt, die Effektivität von risikobasierten Identifikationsmodellen für eine gezielte Tierwohlbefindenskontrolle erhöhen können. Zukünftige Forschung sollte daher auf die Optimierung der Anwendung dieser Registerdaten in z.B. Überwachungssystemen fokussieren, um eine valide Auserwählung von Problembetrieben zu rechtfertigen.

1. Introduction

1.1 Background and relevance

Increased public awareness in animal welfare has led to marked political initiatives not only in Denmark but globally over the past decades. On national level, the primary sector launched a number of initiatives to ensure animal welfare and label products. For example the dairy cooperations have introduced milk from grazing cows e.g. Arla Lærkevang, Naturmælk, and Thise or beef producers (Dansk Kalv) besides the organic label on eggs, meat and dairy products. Even in a global perspective, large cooperations like McDonald's Cooperation (2004) and Marks and Spencer Group plc (2010) initiated welfare assurance programmes to ensure consumers that producers delivering food products to these companies maintain a certain welfare standard on-farm (Blokhuis *et al.*, 2008). However, most of these commercial schemes operate with minimum standards for housing (e.g. space allowance) and management procedures (e.g. grazing, feeding), hence, welfare is defined by a limited or minimum set of criteria and the definition does not provide a full welfare assessment. This discrepancy becomes even more pronounced within the regulatory welfare inspection schemes. The official inspections are bound to evaluate farms on the basis of existing animal welfare legislation which also is mainly centred on minimum requirements on housing standards (feed and water supply) and proper treatment of animals. So both farm assurance and official inspections schemes assess animal welfare from a very limited perspective.

Animal welfare is often regarded as a multidimensional concept (Fraser *et al.*, 1997). The concept relates to three 'welfare schools', namely 'biological functioning' – reflecting the animals' attempt to cope with its' environment (Broom, 1996), 'emotional state' – defining welfare by animals' feelings (Duncan, 1996), and 'naturalness' – emphasizing the ability to live according to the animals' genetically encoded nature (Rollin, 1993). In order to assess animal welfare the animals' themselves have to be evaluated, as animal-based measures should be regarded more valid than resource-based measures as they more directly reflect the welfare consequences (Whay *et al.,* 2003; Webster *et al.,* 2004). However, the animal-based measures are time-consuming and can be prone to observer subjectivity decreasing the reliability of the given measure, while the more objective and feasible resource-measures present more efficient ways of assessing animal welfare. Many animal-based measures are associated not only with each other, but also with other more feasible measures like routine registrations held by national databases (deVries *et al.*, 2013). However, so far no studies have evaluated variables or models from these register data as risk factors or indicators for animal welfare as a complex entity e.g. given by an overall welfare index or score. Aggregating welfare measures into an overall score holds the advantages of being transparent, feasible and operational; making this approach attractive for e.g. national or industry screening of the welfare of dairy herds.

However, the question whether animal welfare can be assessed in a more cost-effective manner without visiting the farm remains open.

The Danish welfare inspections have tried to meet this challenge by using register data as risk parameters in their targeted or risk based sampling of livestock herds subjected to welfare inspections. In 2004, a five-percent risk-based sampling was implemented using risk parameters such as herd size, production system (beef or dairy), antimicrobial usage and previous violations of legislation, neglect of the compulsory ear-tagging, organic herds, percentage lean and thin cows at slaughter. Register data from accessible national databases are used for this purpose. But many of these recordings have varying sensitivities, challenging the accuracy of the risk-based identification and the coverage of the multi-complexity of animal welfare. Risk parameters have not been validated against a comprehensive on-farm welfare assessment, and consequently there is no substantial evidence, that the current risk-based sampling scheme is appropriate in this context. Results from the previous inspections in cattle showed non-compliance issues leading to warnings, enforcement notices or police reports in 25% of the inspected farms. This success rate is debatable as reasons for this low prevalence might be a general good standard amongst Danish herdsmen or an inefficient identification system. The risk parameters available are all obtained from official databases and are limited to mandatory recordings of location (GIS coordinates), herd size, breed, birth, culling, movements, production type (i.e. conventional/organic/dairy/beef) within the Central Husbandry Register (CHR); herd health agreement status within the National Veterinary Practioners Register (VetReg); antimicrobial consumption within the VetStat database; abattoir remarks within the meat inspection database; and the animal welfare database holding information on previous inspection results (Cleveland Nielsen, 2011). All the official databases are also linked to the privately industry and farmer owned Danish Cattle Database (DCD). This comprehensive database compiles data on milk yield and quality from the registration and yield control (RYK), veterinary treatments/farmers treatments, reproduction results from breeding schemes, hoof trimmers, and laboratory results (Houe *et al.*, 2004). In contrast to primary observational data collected for research purposes the data compiled in the DCD are considered as secondary data, as they are collected for other purposes than research. Hence, these data hold potential drawbacks as the validity and coverage of data may be compromised in regards to the specific research interest. This becomes evident in respect to animal welfare. Although, the register data within the DCD are covering the aspect of biological functioning to a great extent; and they can reflect some aspects of the emotional or affective state, by addressing the levels of disease and injuries, they have a deficit in the lack of their coverage of the emotional aspect concerning animal behaviour. Hence, potential synergies between register data and on-farm welfare assessment still need further investigations and clarifications.

1.2 Aim of thesis

The general aim of the present thesis is to develop and evaluate a method for the identification of herds at risk for animal welfare problems by investigating the potential of different levels of information for animal welfare assessment.

The objective of the study is to develop and evaluate models for the identification of animal herds with animal welfare problems on the basis of different levels of information available. Hence, the present thesis seeks to meet the objective by investigating information levels classified according to their feasibility and costs pursuing the hypotheses:

I. Cross-sectional findings such as clinical observations can be identified and classified based on existing data (such as: meat inspection data, mortality data, data on medicine use and milk recording data) in a feasible, valid and transparent manner.

II. Welfare indicators based on additional recordings such as clinical observations, animal behaviour observations can reveal the underlying nature of welfare problems with a high validity.

III. Livestock herds with animal welfare problems can be identified by existing data without visiting the farm.

Three objectives were established in order to answer the three hypotheses:

Objective I (Paper 1) – Evaluation of register-based measures

To investigate the predictive potential of incidence data collected over a longer period to predict cross-sectional findings exemplified by evaluating the diagnostic potential of register data to identify dairy herds with lameness prevalences above an acceptable threshold.

Objective II (Paper 2) - Evaluation of animal-based measures

To investigate the effect of misclassification bias of clinical scorings and to present a solution for alleviating the inaccuracy and uncertainty of these scorings in multiple-rater study designs.

Objective III (Paper 3 and Paper 4) – Evaluation of overall animal welfare

To investigate the level of agreement between different animal welfare indices/definitions based on different levels of information and identify potential risk factors for impaired animal welfare.

1.3 Outline of the thesis

The thesis contains 7 Chapters and 3 appendices. A general overview of animal welfare and animal welfare assessment is presented in a literature review in Chapter 2. The methodological aspects for the three objectives within this thesis are presented in Chapter 3 followed by the manuscripts in Chapter 4 and 5. A general discussion of issues encountered within the three objectives is presented in Chapter 6 providing a link to the conclusion and perspectives in Chapter 7. Appendix A gives an overview of the included register data variables; while Appendix B provides the recording sheets for the on-farm assessments, Appendix C provides definitions on the scoring of clinical and behavioural measures.

References

Blokhuis, H.J, Keeling, L.J, Gavinelli, A. and Serratosa, J. 2008. Animal welfare´s impact on the food chain. Trends in Food Science Technology 19: S79-S87.

Broom, D.M 1996. Animal Welfare Defined in Terms of Attempts to Cope with the Environment. Acta Agriculturae Scandinavica Section A- Animal Science Suppl. 27:22-28.

Cleveland Nielsen, A. 2011. Datawarehouse for assessing animal health, welfare, risk management and communication. Acta Veterinaria Scandinavica 53(S1):S3.

deVries, M., Bokkers, E.A., Dijkstra, T., van Schaik, G. and de Boer, I.J. 2011 Invited review: associations between variables of routine herd data and dairy cattle welfare indicators. Journal of Dairy Science 94: 3213-3228.

Duncan, I.J.H. 1996. Animal welfare defined in terms of feelings. Acta Agriculturae Scandinavica - Section A Animal science Suppl 27: 29-35.

Fraser, D., Weary, D.M, Pajor, E.A. and Milligan, B.N. 1997. A Scientific Conception of Animal Welfare that Reflects Ethical Concerns. Animal Welfare 6:187-205.

Houe, H., Ersbøll, A.K. and Toft, N. 2004. Introduction to Veterinary Epidemiology. Biofolia, Frederiksberg, Denmark.

Rollin, B. 1996. Ideology, "value-free science", and animal welfare. Acta Agriculturae Scandinavica - Section A Animal science. Suppl. 27:5-10.

Webster, A.J.F., Main, D.C.J. and Whay, H.R. 2004.Welfare assessment: indices from clinical observations. Animal Welfare 13(S):93-98.

Whay, H.R., Main, D.C.J, Green, L.E. and Webster, A.J.F. 2003. Animal-based measures for the assessment of welfare state of dairy cattle, pigs and laying hens: consensus of expert opinion. Animal Welfare 12: 205-217.

2. Literature review

2.1. The Five Freedoms – protection from cruelty to animal welfare

Animal welfare can be dated back to ancient civilisations and their religiously motivated concerns for animals, due believe in reincarnation (e.g. Indian religions, Hinduism, Buddhism) or embedded in ritual slaughter (Abrahamic religions, e.g. Jewish and Islamic religion). In Christianity, however, there are no direct concerns for animals unless virtues like kindness, mercy and love are transferred on to animals as well. In the book of Genesis man is given power to rule over animals, an inherent right that made animals the private property of man.

During the eighteenth and nineteenth century liberalism and human rights movements sparked of the ethical discussions not only on the ethical obligations man has towards others, e.g. racial domination, but also towards animals. A representative of Utilitarianism, believing that the ethically right act is the one, that maximizes the total welfare, the British philosopher Jeremy Bentham included animals into these ethical concerns based on their ability to 'feel': "*The question is not, Can they reason?, nor, Can they talk? But can they suffer?*" (Bentham, 1789). Later, the first law for the protection of animals was passed in British Parliament in 1822 in the "Act to Prevent Cruel and Improper Treatment of Cattle 22d July 1822". This bill described only "*wanton cruelty*" to four-legged farm animals by other persons. These limitations were set under the assumption that bad treatment of animals by their owners would be irrational in order to protect the value an animal presents (Sandøe & Christiansen, 2008).

The Second World War left Europe in a devastated and poor state with a desperate need for cheap food. The result was intensification of livestock production and with its subsequent consequences of increased efficiency and competition taking its toll on production animals. In 1964 Ruth Harrison published her book "*Animal Machines*" in order to expose the reality of intensive livestock production or "factory farming" as she called it (Harrison, 1964). She challenged the legislators and stakeholders by saying:

"*If one person is unkind to one animal it is considered as cruelty but where a lot of people are unkind to a lot of animals, especially in the name of commerce, the cruelty is condoned and, once large sums of money are at stake, will be defended to the last by otherwise intelligent people.*"

(Ruth Harrison, 1964)

Pictures and descriptions of the disgraceful lives of veal calves, battery hens and broilers, and pigs struck the British consumers with concern; and her well-informed criticism played a pivotal role in shifting the focus from *anti cruelty* to *animal welfare* in the legislative context. Hence, as a direct response the UK government commissioned the Brambell Committee to investigate the welfare of farmed animals. The final report published in December 1965 stated that farm animals are sentient beings and hence, setting out the minimum acceptable treatment of farm animals by requiring the freedom of animals "to stand up, lie down, turn around, groom themselves and stretch their limbs". The Brambell report led to the establishment of the Farm Animal Welfare Advisory Council (FAWAC) which was replaced by the Farm Animal Welfare Council (FAWC) in 1979; the responsible body for contextualising the Five Freedoms as they are known today:

1. **Freedom from Hunger and Thirst** - by ready access to fresh water and a diet to maintain full health and vigour.

2. **Freedom from Discomfort** - by providing an appropriate environment including shelter and a comfortable resting area.

3. **Freedom from Pain, Injury or Disease** - by prevention or rapid diagnosis and treatment.

4. **Freedom to Express Normal Behaviour** - by providing sufficient space, proper facilities and company of the animal's own kind.

5. **Freedom from Fear and Distress** - by ensuring conditions and treatment which avoid mental suffering.

According to a more recent statement by the FAWC the minimum treatment for farm animals is not only dependent on the Five Freedoms but also on having a life worth living; which implies a positive balance of experiences throughout the animals' entire lifespan (FAWC, 2009). This importance of positive experiences emphasizes the dramatic change and development of animal welfare as we define it today and as it is regulated in most countries.

Although animals were already recognised as sentient beings in the sixties, laying the foundation to create modern animal protection legislations, many countries did not implement Animal Welfare Acts until 20-30 years later. Although the European Union also supported the recognition of animals sentient in the annex to the EU Treaty of Amsterdam in 1997, it was not implemented as an Article until The Treaty of Lisbon in 2007, and further not ratified until 2009 in EU Treaty *Treaty on the Functioning of the European Union* (TFEU). Here Article 13 put animal welfare on equal terms with other key principles.

"In formulating and implementing the Union's agriculture, fisheries, transport, internal market, research and technological development and space policies, the Union and the Member States shall, since animals are sentient beings, pay full regard to the welfare requirements of animals, while respecting the legislative or administrative provisions and customs of the Member States relating in particular to religious rites, cultural traditions and regional heritage.". (Article 13, Title II, TFEU, 2009)

Within the past decades, EU directives have led to very species specific legislations and among the EU member states as many as 97 national animal welfare related legislations are listed by the European Enforcement Network of Animal Welfare lawyers and commissioners (http://lawyersforanimalprotection.eu). Even though, many developing countries are struggling with humanitarian issues animal welfare is also of importance. A recent review investigating animal welfare legislation in third world countries showed that out of 25 countries providing information 19 did have legislation on animal welfare or protection (Bracke, 2009). This global focus was put on the agenda by the OIE´s strategic plan from 2001-2005, where animal welfare was first identified as a priority. In 2005, animal welfare standards were included in the Terrestrial Animal Health Code, making animal welfare a prerequisite for bilateral trade within OIE membership countries. Although vast amounts of legislation concerning animal welfare exist, no definite definition of animal welfare exists.

2.2 Animal welfare as a multi-dimensional entity

"… animal welfare means how an animal is coping with the conditions in which it lives. An animal is in a good state of welfare if (as indicated by scientific evidence) it is healthy, comfortable, well nourished, safe, able to express innate behaviour, and if it is not suffering from unpleasant states such as pain, fear and distress. Good animal welfare requires disease prevention and veterinary treatment, appropriate shelter, management, nutrition, humane handling and humane slaughter/killing. Animal welfare refers to the state of the animal; the treatment that an animal receives is covered by other terms such as animal care, animal husbandry and humane treatment", OIE definition of animal welfare (2008).

Definitions of animal welfare or well-being are a matter of answering the general philosophical questions of *what constitutes a good life?* (Hurnik, 1993). The definition of animal welfare is highly influenced by ethical considerations as described in the preceding section. The political changes during the eighteenth century introducing liberal democracy and human rights also affected the normative ethics. Where, traditionally, the normative ethics had only concerned the rightness or wrongness of acts, the modern normative ethics were now adding more moral complexity to the matter.

Thinkers like Jeremy Bentham (1748-1832) and John Stuart Mill (1806-1873) now defined the rightness of acts by judging the amount of happiness the given act produces. This ethical theory of **Utilitarianism** relies on "the greatest happiness principle". Since happiness requires the ability to feel, i.e. being *sentient*, Bentham also included animals in to the moral concerns of man. Seeking to maximize happiness or welfare represents a **hedonistic theory** of maximizing pleasure and minimizing pain. Hence, the utilitarian and hedonistic views on animal welfare represent a subjective approach relating to the animals' *feelings*. In this view the quality of an animals' life is hence defined by the sum of positive and negative experiences (Simonsen, 1996). Another subjective approach to animal welfare is the **preference satisfaction**, where the fulfilment of desire will lead to increased pleasure and a positive mental state (Appleby & Sandøe, 2002). Parfit (1984) introduced a hybrid view of these two, **preference hedonism**, refining the amount of pleasure to one individual to be dependent on the individuals' own value of the pleasure. This hybrid view alleviates the problem of individuals having different thresholds for pain and pleasure sensation and makes well-being an individual trait.

A more objective animal welfare definition can be achieved by the **perfectionistic theory**. This theory defines welfare as the harmony between the animal and its surroundings in both physical and mental respect (Hurnik, 1993). Perfectionism goes back to Aristotle's' believe that only when man can realize his full potential in accordance to his nature, he will be content and complete. According to this definition, animal welfare is dependent on the nature of the animal as a point of reference. Hence, Rollin (1996) stated that welfare also had to cover the "*nurturing and fulfilment of the animals' natures*", their so-called *"telos"*. In other words, according to perfectionism animals should live natural lives that enable them to have their biological needs fulfilled; and hence they will produce and perform well.

The ethical views on animal welfare and animal well-being present different elements for addressing the contents of a good life. But how should the good life or quality of the animals' life then be assessed? There is no doubt, that the assessment method will be highly dependent on the overall notion of welfare; consequently, three different welfare schools have emerged emphasizing on either *biological functioning, feelings* or *naturalness*.

The spokesman for **biological functioning** and the fulfilment of the biological needs Donald M. Broom stated: "*The welfare of an individual is its state as regards its attempts to cope with its environment"* (Broom, 1986), hence failure to cope with the environment will lead to the following: reduced life expectancy, reduced ability to grow or breed, body damage, disease, immunosuppression, physiological attempts to cope, behavioural attempts to cope (stereotypies, aggressiveness, self-mutilation, and self-narcotisation) (Broom, 1993). A definition very close to the WHO definition of health: "*Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity*" (WHO 1948). This health-centred view on animal welfare is often of greatest concern to the people directly involved in the care of animals (e.g. farmers, veterinarians, animal scientists and advisors) (von Keyserlingk *et al.*, 2009; Veissier *et al.*, 2011). This approach offers more or less objective measures of welfare in terms of disease incidence, milk yield, and reproductive rates; and furthermore, as stated by Broom, welfare defined by biological functioning "*...can be measured in a scientific way that is independent of moral considerations*" (Broom, 1991).

Although, coping with the environment also may have an effect on behaviour its role within the definition of biological functioning lacks clear distinction. According to Duncan (1993) behaviour is motivated by "*feelings, emotions or affective states*", hence arguing that animal welfare is all about sentience and **feelings**. This hedonistic point of view is often stressed by animal scientists, as the prevention from pain and suffering has long been an integrated part of animal welfare. Assessing animals' feelings, however, is not straightforward, as valid measures are needed. For the adverse effects of negative emotional experiences like pain and fear, stress levels can be measured by evaluating serum cortisol levels (Duncan, 1996); while the long term adverse effects can be measured by evaluating stereotypies or agonistic behaviour. More recently research has focused on measuring the positive mental states identifying play and social behaviour (Boissy *et al*., 2007).

Finally, the most simplistic notion regards the **naturalness** of animals, emphasizing the ability of the animal to live according to its nature or telos (Rollin, 1993). This perfectionistic view is often strongly weighed by the consumers (Lassen *et al.,* 2006). However, in its original meaning, it constitutes a somewhat contradictory notion of welfare, as it can be argued that some aspects of animals nature e.g. as listed by (Spinka *et al.,* 2006) disease and infections, and predator attacks are resulting in poor welfare.

In order to cover all these aspects of animal welfare, an overlapping approach is needed (Fraser *et al*., 1997) to cover this multi-complex entity when assessing animal welfare. Hence, an integrated approach is needed to ensure the optimal coverage of all animal welfare aspects and to allow for overlap between definitions as depicted in Figure 2.1. As an example, lameness as a welfare measure can represent all three aspects. It is associated with the affective state of an animal due to its painful nature; as it is a symptom of impaired health it also belongs to the biological functioning; and finally it restricts the animal to perform its natural behaviour.

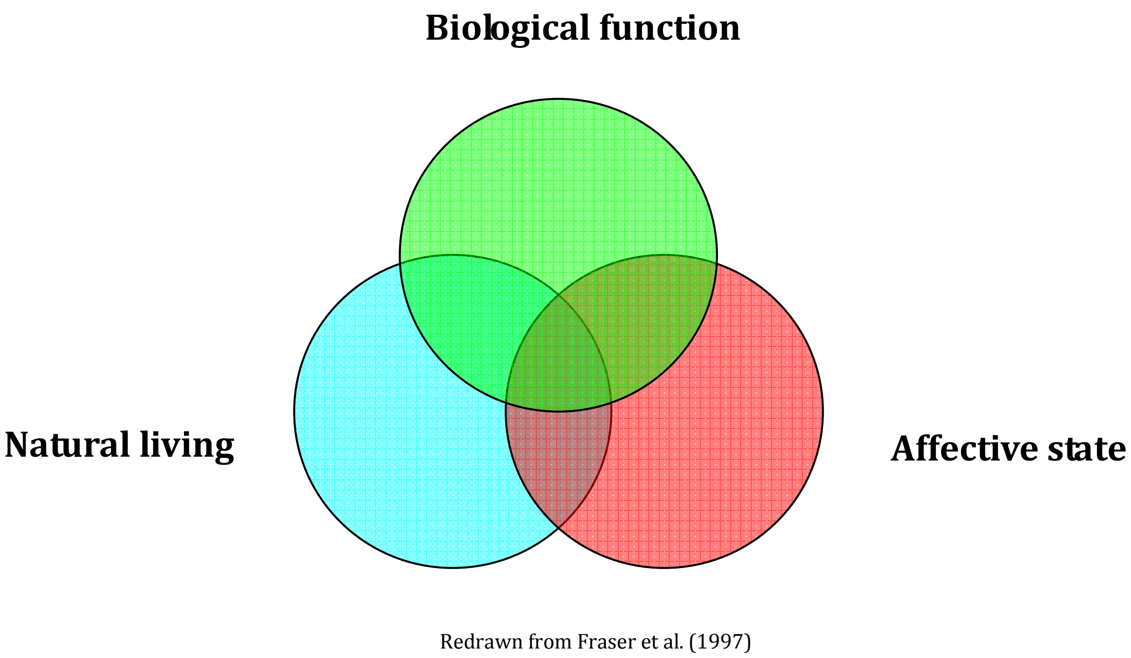


Figure 2.1 Conceptual animal welfare model adopted by Fraser et al. (1997) showing the three overlapping animal welfare concerns.

Furthermore, from a scientific point of view, the distinct animal welfare definition is of great importance for being able to move from the lexical (i.e. ethical theories) and explanatory definitions (i.e. the three welfare schools) to the operational level of animal welfare assessment (i.e. measures to be included and evaluation of results). The following section will introduce welfare assessment methods and the ethical, explanatory and operational issues related to the given method.

2.3 Animal welfare issues in dairy production

Although *the Freedom from pain* was part of the initial Five Freedoms, science did not fully acknowledge animal pain as a subjective state. Until the 1980´ies animal pain was regarded as a state of physiological stress and animal's consciousness and awareness was not considered legitimate. Dantzer & Mormede (1983) investigated the dependent relationship between the degree of physiological stress and the emotional reactions with particular emphasis on the long-term stress response as being the most common in farmed animals. However, chronically stressed animals showed little or no endocrine stress response, showing animals' impressive adaptive abilities. More recently, studies indicated that physiological adaptive changes come with an even more measurable cost on physiology when prolonged with impaired productivity, reduced appetite and immunosuppression as the most prominent costs (Webster, 2005).

Hence, looking at detrimental effects of modern dairy production these adaptations become clear in the cows responses in coping with their environment. Examples include integument alterations due to inadequate bedding and resting area conformation; lameness due to a genetic selection for high milk yield and the negative effects of these hereditary traits like poor hock and hind leg conformation (Brotherstone & Hill, 1991; Capion *et al.,* 2008) or due to infections (e.g. digital dermatitis, interdigital phlegmone, foot rot) caused by poor hygiene (Philipot *et al.,* 1994); and stereotypic behaviour like tongue rolling or licking of equipment due to stress or frustrations (Redbo *et al.*, 1996). Although, housing deficits exert a huge impact on behaviour and welfare in relation to the accommodation of cows' biological needs, scientist, farmers and veterinarians have primarily focused on the impact of painful conditions in relation to welfare. Kielland et al. (2010), however, made an interesting finding in a questionnaire, when asking farmers whether animals are able to feel physical pain like humans 70% agreed, while, surprisingly 15% disagreed (the remaining 13% were indifferent, and 2% did not answer). Furthermore, analysing the farmers' attitude towards pain showed better welfare on farms belonging to farmers showing a high level of empathy (i.e. high pain scores) towards the animals. Comparing the farmers' perception of pain intensity with the veterinarians' perception given in Table 2.1 shows a good agreement in alignment with Thomsen et al. (2012).

Locomotor disorders are ranked highly painful and are considered one of the major challenges in modern dairy production due the complexity associated with the multifactorial etiology. Lameness is recognized as one of the main factors responsible for economic losses in the British dairy industry contributing 27% to the total health costs (Kossaibati & Esslemont, 1997). The impact of a lameness incidence in a Danish dairy herd leads to an estimated loss of € 192 per first case (Ettema & Østergaard, 2006), due to a prolonged loss in milk yield (Green *et al.,*

2002) increased risk of premature culling (Thomsen *et al.*, 2004) and decreased conception rates (Sprecher *et al.,* 1997; Sogstad *et al.*, 2006). But, more importantly, lameness is also regarded as a prominent welfare issue due to the direct consequences for dairy cattle welfare in terms of pain (Oltenacu & Broom, 2010; Whay *et al.,* 1997). Another negative welfare aspect is the major contribution to cow mortality as 40% of euthanized Danish dairy cows are euthanized due to locomotor disorders (Thomsen *et al*., 2004). A message, clearly perceived and agreed upon by farmers, as they ranked “pain and suffering” as a more important consequence of lameness than “reduced profitability” (Leach *et al.*, 2010). Hence, lameness plays a pivotal role in welfare assessments.

Many other factors than merely clinical measures are associated with lameness either through a direct causal relationship or through other confounding effects as illustrated in Figure 2.2. These co-dependencies highlight the complexity presented by the major welfare issues in dairy production and should be regarded and reflected in a multi-dimensional welfare assessment in order to give a comprehensive picture of the welfare status.

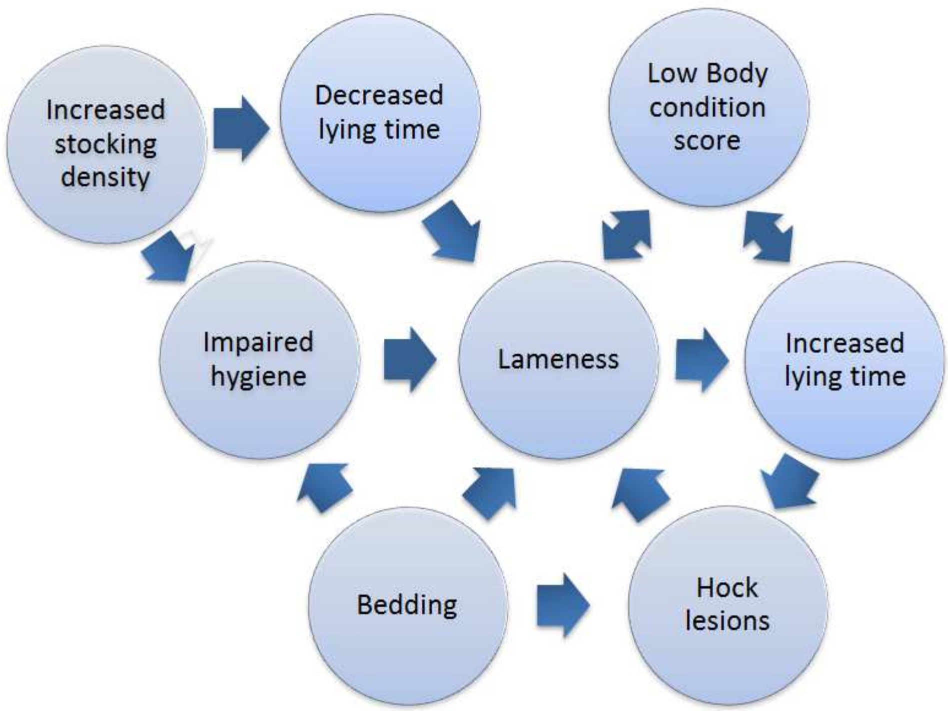


Figure 2.2 Causal diagram for lameness and co-dependencies between variables associated with lameness.

2.4 Measuring and assessing dairy cattle welfare

2.4.1 Welfare measures for assessment

As indicated by the preceding sections with the aim of covering the multi-complexity of animal welfare, both from and ethical and practical point view, a wide range of animal measures have to be incorporated in a welfare assessment. Welfare measures can be allocated into two groups: resource-based measures (e.g. management practices and system or housing related factors) or input measures; and animal-based measures (e.g. disease/injury incidence and behaviour, production results) the direct output of the animals attempt to cope with the given environment (Barnett & Hemsworth, 2009). The animal-based measures are further divided into direct measures from animal observations and indirect measures derived from routine registrations on e.g. mortality or milk production (EFSA, 2012). Traditionally, the resource-based measures have been used in farm assurance schemes and for regulatory purposes, as they provide minimum standards for the housing and handling of animals (Sørensen & Fraser, 2010), due to their feasibility of recording. However, the impact of the management factor presented by the caretaker on the actual welfare of the herd should not be underestimated, as good management can outweigh the negative consequences of bad housing systems (Whay *et al.*, 2003; Webster *et al.*, 2004; Rousing *et al.,* 2007). Hence, the most direct reflection of the animals' welfare state should be obtained by observing the unit of interest – the animal. As argued by Scott et al. (2003) the welfare should be assessed by means of objective assessment of signs – not symptoms. The underlying notion of this distinction derives from the definition of symptoms. Symptoms are considered as a subjective experience communicated by the patient, while signs are the objective results found by diagnostic tests. According to Saunders Comprehensive Veterinary Dictionary (2007) the term *symptom* is not applicable to animals. Observing animals, however, is very time consuming due to the training of observers and the actual data collection (Knierim & Winckler, 2009). Acknowledging that resource-based measures do play an important part as well, as they can comprise a risk to animal welfare in the long run and hence, combining both measure types would be beneficial for reaching highest feasibility and validity (Rousing *et al.*, 2001; Sørensen & Fraser, 2010). For dairy cattle a number of multi-dimensional assessment protocols have been developed, e.g. the British Freedom Food (RSPCA), the Dutch COWEL (Ursinus *et al.*, 2009), the Austrian *Tiergerechtheitsindex* TGI (Bartussek, 1999), the Danish Cattle Federation (DCF) protocol (Enemark & Rousing, 2007), and the comprehensive EU project Welfare Quality® (2009). The commonly used dairy cattle welfare measures are summarised in Table 2.2.

Whereas the animal-based measures lead to a diagnosis of impaired animal welfare, the resource-based measures account for the potential risk for animal welfare. No matter, what approaches are being used, they will always be accompanied by specific issues concerning the validity and repeatability.

Firstly, looking at the animal-based approach concerned with diagnosing animal welfare, the etymology of the word *diagnosis* is as follows: *dia* = apart and *gignoskein*= to learn or distinguish. This definition implies a cognitive process of logically combining knowledge about observed signs and diagnostic tests to identify a possible disease or disorder, in other words only an accumulation of measures can provide the diagnosis of impaired welfare. Hence, measures should be appropriately associated with the given diagnosis of impaired animal welfare. Secondly, the scale of measurement is of importance – depending on the need to only discriminate between presence and absence of signs or on a more detailed description of the severity. Within animal welfare assessment, the relative order of categories found in an ordinal scale is the most frequently used (Scott *et al.,* 2003). This approach enables the assessments to capture the essential impact of intensity and to some extent also the duration of welfare related disorders. However, increasing the number of categories within such a scale will be associated with an increase in uncertainty as the observers' subjective interpretation of signals becomes more evident. In this light, Brenninkmeyer et al. (2007) reported better reliability of a reduced lameness scoring scale at early stages compared to a full scale. This matter of validity and reliability is crucial for choosing the appropriate measures to be included in a welfare protocol. However, there are many definitions of validity and what it encompasses. For clarity, in the present thesis validity will refer to the ability to detect the true state i.e. covering the sensitivity of measures or completeness of register data, specificity, and correctness or positive predictive value of data and measures. Reliability will be defined by the robustness of measures or indicators in terms of minimizing and quantifying the random variation within the given measure (e.g. inter observer agreement).

Traditionally, internal validity can be assessed by evaluating the correctness of the inference made based on a "gold standard test" given by the sensitivity and specificity – or sometimes as – the positive predictive value - of the test in question. In a traditional context, this is pretty straight forward, as most tests only have to distinguish between presence and absence of a disorder – usually based on serology results quantifying the presence of antibodies against a certain etiologic agent, where a given thresholds determines the outcome. However, these thresholds are highly specific to the given disease i.e. the case definition is dependent on a threshold, whereas for animal welfare requires an ethical decision on what is acceptable and what is unacceptable? For resource-based question the answer can be quite easy, any non-compliance with minimum standards or industry codices could classify the herd as a problem herd. But when it comes to the animal-based measures a whole new discussion arises.

2.4.2 Acceptable versus unacceptable welfare? – individual vs. herd?

Having established the means of assessing animal welfare lead on to a new arising issue: how should these measures be aggregated in order to depict the welfare state of a given herd? Should the individual or the herd be the element of focus? Again, this calls for ethical considerations and recalling the utilitarian view, where the sum of welfare counts, argues in favour of welfare aggregation to herd level. However, this view offers little regard to the individual animal and to the distinction between distributions of severity levels; hence, in the utilitarian light, a few individuals suffering for the greater good is not wrong as long as the net welfare is acceptable. The consequentialism of pooling welfare and judging the size stands in clear contrast to both the priority view and egalitarism, being centred on the individual animal. While the priority view puts an absolute weight on individuals proportional to the severity of their state, the egalitarian view judges the relative weight of the severity compared to other levels and individuals (Jensen, 2003).

At the end the goal or aim of the given welfare assessment should decide the aggregation procedure – for regulatory purposes minimum standards set the limits between acceptable and unacceptable welfare. However, when welfare is assessed by means of animal-based measures the choice is not that straightforward. Using the individual approach would be very time-consuming, as every single animal should be evaluated, which from a practical point of view hardly would be applied in large scale studies. Hence, evaluating welfare at herd level would be the more feasible approach. Next arises the issue of weighting severity levels e.g. weighting the welfare impairment of a severely impaired animal to a moderately impaired one – should this be a relative or an absolute weight and how should these be derived? Both Spoolder et al. (2003) and Rodenburg et al. (2008) agreed that in relations to weighting and aggregation of animal welfare measures a certain degree of subjectivity would be involved, for what reason expert opinion is needed. Finally, for determining the case definition "impaired animal welfare" – thresholds need to be defined either by a pre-set cut off or data-driven cut off e.g. percentiles. The following section will give examples on how these issues have been dealt with in different cattle welfare protocols.

2.4.3 Welfare assessment protocols for dairy cattle

The increasing awareness of consumers in animal welfare friendly products and legislators need to ensure animal welfare during the last decades lead to the development of several animal welfare assessments protocols. Although, all these protocols aim at measuring welfare, they differ strongly in their content. For this matter the real purpose or goal of the given welfare assessment protocol should be kept in mind. As pointed out by Johnsen et al. (2001) the European schemes mostly aim at certification of farms with respect to organic farming or welfare labelling, while only a small number of schemes serve as advisory tools or for benchmarking purposes. However, quite a number of new protocols have emerged since Johnsen et al. (2001) discussed the differences in methods, aims and goals. The continuous development of protocols emphasizes the necessity of meeting the conflicting interests of producers and the industry on one hand and consumers and legislators on the other. Hence, farm assurance schemes would primarily deal with compliance of minimum standards as their core issue (Mench, 2003) preferring resource-based measures (e.g. TGI), while welfare management schemes would rely more on animal-based measures (e.g. Freedom Foods, Bristol Welfare Assurance Programme) for the quantification of on-farm welfare (Webster *et al*., 2004). Ingemann et al. (2009) argued, that the Danish Cattle Federation (DCF) industry protocol was too narrow in its appointment of measures to meet all of the three overall policies of ensuring good animal welfare, good farmer profitability and to improve the dialogue with the public by meeting their concerns as well.

Therefore, measures included in the scheme are one challenge, while the aggregation of measures is another in order to process the vast information on single measures into an overall welfare interpretation. Aggregation methods vary from being based on simple discussions with farmers in advisory schemes (e.g. Sørensen *et al.,* 2001), direct comparison of benchmark components based on percentiles among the population results (e.g. Freedom Food, Fråga Kon), mean of ranks (Whay *et al.*, 2003) to sums or mean scores (e.g. TGI). Summing up scores by an additive linear approach is a very intuitive method of summing up all indicators of poor or good welfare in a transparent matter. However, it is also associated with issues of compensation between measures and the dependency between some measures (e.g. Figure 2) (Botreau *et al*., 2007). Accordingly, the very comprehensive EU funded Welfare Quality® project designed a hierarchical evaluation model where the aggregation runs from measures to criteria, to principle level and finally, to an overall assessment score (Veissier *et al.,* 2011). Along the aggregation process expert opinion is incorporated in the weighting of measure levels for the interpretation of scores, resulting in a non-linear model. The final construct of the Welfare Quality® assessment protocol for dairy cattle is illustrated in Table 2.3.

Table 2.3. The Welfare Quality® assessment protocol for dairy cattle described by measures, criteria and principles (Welfare Quality®, 2009).

Measure	Criteria	Principle
BCS	1. Absence of prolonged hunger	**Good Feeding**
Water provision & cleanliness, functioning of water points	2. Absence of prolonged thirst	**Good Feeding**
Lying-down and collisions, Animals lying partly/completely outside of lying area, Cleanliness (udder, flank, upper leg, lower leg)	3. Comfort around resting	**Good Housing**
No measures developed	4. Thermal comfort	**Good Housing**
Tethering (yes/no), Access to loafing area or pasture	5. Ease of movement	**Good Housing**
Lameness, integument alterations	6. Absence of injuries	**Good Health**
Coughing, Nasal/occular/vulvar discharge, Hampered respiration, Diarrhoea, SCC, Mortality, Downer cows	7. Absence of disease	**Good Health**
Disbudding/dehorning, tail docking	8. Absence of pain	**Good Health**
Agonistic behaviours	9. Expression of social behaviours	**Appropriate Behaviour**
Access to patsure	10. Expression of other behaviours	**Appropriate Behaviour**
Avoidance distance	11. Good human-animal relationship	**Appropriate Behaviour**
Qualitative behaviour assessment	12. Positive emotional state	**Appropriate Behaviour**

2.4.4 Welfare control versus welfare assessment

Official animal welfare control is in many instances limited to assessing resource-based measures as the legislative foundation for animal welfare is based on defining minimum requirements for housing and management procedures. As a result, welfare control cannot be seen as an actual welfare assessment, only as a control of compliance with current legislation. The context of this legislation varies quite a lot between countries and continents. In the United States, no national legislation regarding animal welfare exists and can only be regulated within the Humane Slaughter Act (1958) (Mench, 2003). In contrast, European countries have a longer tradition for animal welfare legislation, as mentioned in Section 2.1.1, and especially EU directives have spun off a number of national initiatives. In Denmark the Act on the Keeping of dairy Cattle and Offspring of Dairy Cattle was implemented in 2010, introducing a number of resource-based measures on space allowance, water provision, walking alleys, resting area, separation of animals etc. However, many paragraphs are not enforced until the expiration of given transition terms and finally completely implemented first of July 2034. At the present the Danish welfare inspections in cattle herds assess compliance with three overall acts (Animal Welfare Act, 2007; Act on keeping of Dairy Cattle and Offspring from dairy Cattle, 2010; and Act on Prohibited Slaughter of and Euthanasia of Foetuses from Production Animals and Horses in the last Tenth of their Pregnancy, 2004), and additional nine executive orders concerning farmed animals, euthanasia, protection of calves, tail-budding and castration, disbudding/dehorning, the use of electrical aggregates, ear tagging and livestock owners use of pharmaceuticals. At the present, five percent of all cattle herds with more than ten animals are sampled in a risk-based scheme including varying risk parameters on an annual basis. Results from previous inspections showed issues with non-compliance in a quarter of all inspected cattle herds. However, this does not directly imply a poor overall welfare status of a given herd, as a great part of non-compliance issues are regarding lack of proper housing for single sick animals and presence of animals in a state requiring euthanasia. Only an on-farm welfare assessment can currently reveal the true nature of the animal welfare status.

References

Appleby, M.C. and Sandøe, P.T. 2002. Philosophical Debate on the Nature of Well-being: Implications for Animal Welfare. Animal Welfare 11: 283-294.

Barnett, J.L and Hemsworth, P.H. 2009. Welfare Monitoring Schemes: Using Research to safeguard Welfare of Animals on the Farm. Journal of Applied Animal Welfare Science 12(2): 114-131.

Bartussek, H. 1999. A review of the animal needs index (ANI) for the assessment of animals' well-being in the housing systems for Austrian proprietary products and legislation. Livestock Production Science 61(2-3): 179-192.

Bentham, J. 1789. A utilitarian view. In: Animal Rights and Human Obligations (eds. T.Regan & P.Singer)(1989), pp. 25-26. Prentice Hall, Englewood Cliffs, NJ.

Boissy, A., Manteuffel, G., Jensen,M.B., Moe, R.O., Spruijt, B., Keeling, L. J., Winckler, C., Forkman, B., Dimitrov, I., Langbein, J., Bakken, M., Veissier, I. and Aubert, A. 2007. Assessment of positive emotions in animals to improve their welfare. Physiology & Behaviour 92(3):375-397.

Botreau, R., Bonde, M., Butterworth, A., Perny, P., Bracke, M.B.M., Capdeville, J. and Veissier, I. 2007. Aggregation of measures to produce an overall assessment of animal welfare. Part 1: a review of existing methods. Animal 1(8): 1179-1187.

Bracke, M.B.M. (Editor) 2009. Animal Welfare in a Global Perspective – A Survey of Foreign Agricultural Services and case studies on poultry, aquaculture and wildlife. Wageningen UR Livestock Research, Lelystad, Netherlands.

Brenninkmeyer, C., S. Dippel, S. March, J. Brinkmann, C. Winckler,U. Knierim. 2007. Reliability of a subjective lameness scoring system for dairy cows. Anim. Welfare.16: 127-129.

Broom, D.M 1986. Indicators of poor welfare. British Veterinary Journal 142:524-526.

Broom, D.M. 1991. Animal welfare: concepts and measurement. Journal of animal Science 69:4167-4175.

Broom, D.M 1996. Animal Welfare Defined in Terms of Attempts to Cope with the Environment. Acta Agriculturae Scandinavica Section A- Animal Science Suppl. 27:22-28.

Brotherstone, S. and Hill, W.G. 1991. Dairy-herd life in relation to linear type traits and production. 2. Genetic analyses for pedigree and non-pedigree cows. Animal Production 53: 289-297.

Capion, N., Thamsborg , S. M. and Enevoldsen.,C. 2008. Conformation of hind legs and lameness in Danish Holstein heifers. Journal of Dairy Science 91: 2089-2097.

Duncan, I.J.H. 1996. Animal welfare defined in terms of feelings. . Acta Agriculturae Scandinavica - Section A Animal science Suppl 27: 29-35.

EFSA Panel on Animal Health and Welfare (AHAW).2012. Statement on the use of animal-based measures to assess the welfare of animals. EFSA Journal 10(6): 2767-2796.

Enemark, P.S. and Rousing, T. 2007. Værktøj til måling og vurdering af dyrevelfærd - Oplæg til danske mælkeproducenter og deres rådgivere. KvægInfo 1720. Available online:https://www.landbrugsinfo.dk/Kvaeg/dyrevelfaerddoedelighed/Dyrevelfaerd/Sider/Vaerktoej_til_maaling_og_vurdering_af_dy.aspx. Accessed 01-10-2013.

Ettema, J.F. and Østergaard, S. 2006. Economic decision making on prevention and control of clinical lameness in Danish dairy herds. Livestock Science 102(1-2): 92-106.

Farm Animal Welfare Council 2009. Farm Animal Welfare in Britain: Past, Present and Future.

Fraser, D., Weary, D.M, Pajor, E.A. and Milligan, B.N. 1997. A Scientific Conception of Animal Welfare that Reflects Ethical Concerns. Animal Welfare 6:187-205.

Green, L.E., Hedges, V.J., Schukken, Y.H., Blowey, R.W. and Packington, A.J. 2002. The impact of clinical lameness on the milk yield of dairy cows. Journal of Dairy Science 85(9): 2250-2256.

Harrison, R. 1964. Animal Machines. The New Factory Farming Industry. Vincent Stuart, London.

Huxley, J.N. and Whay, H.R. 2006. Current attitudes of cattle practioners to pain and the use of analgesics in cattle. Veterinary Record 159:662-668.

Ingemann, R., Sandøe, P., Enemark, P. and Forkman, B. 2009. Conflicting goals of welfare assessment schemes: a case study.Animal Welfare 18(4): 487-495.

Jensen, K.K. 2003. What is the difference between (moderate) egalitarianism and prioritarianism? Economics and Philosophy 19: 89-110.

Kielland, C., Skjerve, E, Østerås, O., and Zanella A.J. 2010. Dairy farmer attitudes and empathy towards animals are associated with animal welfare indicators. Journal of Dairy Science 93: 2998-3006.

Knierim, U. and Winckler, C. 2009. On-farm welfare assessment in cattle: validity, reliability and feasibility issues and future perspectives with special regard to the Welfare Quality® approach. Animal Welfare 18(4): 451-458.

Kossaibati, M.A. and Esslemont, R.J. 1997. The costs of production diseases in dairy herds in England. Veterinary Journal 154(1): 41-51.

Laven, R.A., Huxley, J.N, Whay, H.R., and Stafford, K.J. 2009. Results of a survey of attitudes of dairy veterinarians in New Zealand regarding painful procedures and conditions in cattle. New Zealand Veterinary Journal 57(4): 215-220.

Leach, K.A, Whay, H.R., Maggs, C.M., Barker, Z.E., Paul, E.S., Bell, A.K. and Main, D.C.J. 2010. Working towards a reduction in cattle lameness: 2. Understanding dairy framers' motivations. Research in Veterinary Science 89:318-323.

Mench, J.A. 2003. Assessing animal welfare at the farm and group level: A United States Perspective. Animal Welfare 12: 493-503.

Oltenacu, P. A., and Broom, D.M. 2010. The impact of genetic selection for increased milk yield on the welfare of dairy cows. Animal Welfare 19:39-49.

Parfit D 1984. Reasons and Persons. Oxford University Press, Oxford, UK.

Philipot, J.M., Pluvinage, P., Cimarosti, I, Sulpice, P. and Bugnard. 1994. Risk factors of dairy cow lameness associated with housing conditions. Veterinary Research 25: 244-248.

Redbo, I., Emanuelson, U., Lundberg, K. and Oredsson, N. 1996. Feeding level and oral stereotypies in dairy cows. Animal Science 62: 199-206.

Rodenburg, T.B., Tuyttens, F.A.M, de Reu, K., Herman, L., Zoons, J. and Sonck, B. 2008. Welfare assessment of laying hens in furnished cages and non-cage systems: assimilating expert opinion. Animal Welfare, 17: 355-361.

Rollin, B. 1996. Ideology, "value-free science", and animal welfare. Acta Agriculturae Scandinavica - Section A Animal science. Suppl. 27:5-10.

Rousing, T., Bonde, M. and Sørensen, J.T. 2001. Aggregating welfare indicators into an operational welfare assessment scheme: A bottom-up approach. Acta Agriculturae Scandinavica Section A-Animal Science 51(30):53-57.

Rousing, T., Jakobsen, I.A., Hindhede, J., Klaas, I.C, Bonde, M. and Sørensen, J.T. 2007. Evaluation of a welfare indicator protocol for assessing animal welfare in AMS herds: researcher, production advisor and veterinary practioner opinion. Animal Welfare 16: 213-216.

Sandøe, P. & Christiansen, S.B. 2008. Ethics of animal use. Blackwell Publishing, Chichester, West Sussex. United Kingdom.

Saunders Comprehensive Veterinary Dictionary, 3 ed. © 2007 Elsevier, Inc.

Scott, E.M., Fitzpatrick, J.L., Nolan, A.M., Reid, J. and Wiseman, M.L. 2003. Evaluation of welfare state based on interpretation of multiple indices. Animal Welfare 12: 457-468.

Simonsen, H.B. 1996. Assessment of animal welfare by a holistic approach: Behaviour, health and measured opinion. Acta Agriculturae Scandinavica - Section A Animal science. Suppl. 27:91-96.

Sogstad, A.M., Østerås, O. and Fjeldaas, T. 2006. Bovine claw and limb disorders related to reproductive performance and production diseases. Journal of Dairy Science 89(7): 2519-2528.

Spinka, M. 2006. How important is natural behaviour in animal faming systems? Applied Animal Behaviour Science 100:117-128.

Spoolder, H., de Rosa, G., Horning, B., Waiblinger, S. and Wemelsfelder, F. 2003. Integrating parameters to assess on-farm welfare. Animal Welfare 12: 529-534.

Sprecher, D.J., D.E. Hostetler, J.B. Kaneene & X.W.o.S.S.l.c.d. Carre (1997): A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance. Theriogenology. Vol. 47, no. 6, pp. 1179-1187.

Sørensen, J.T., Sandøe, P. and Halberg, N. 2001. Animal Welfare as one among several values considered at farm level: The idea of an ethical account for livestock farming. Acta Agriculturae Scandinavica Section A-Animal Science 51(30):11-16.

Sørensen, J.T. and Fraser, D. 2010. On-farm welfare assessment for regulatory purposes: Issues and possible solutions. Livestock Science 131(1) 1-7.

The Treaty of Amsterdam. Protocol on Protection and Welfare of Animals. Official Journal C 340, 10 November 1997.

Thomsen, P.T., Kjeldsen, A.M, Sørensen, J.T. and Houe, H. 2004. Mortality (including euthanasia) among Danish dairy cows (1990-2001). Preventive Veterinary Medicine 62(1): 19-33.

Thomsen, P.T., Anneberg, I. and Herskin, M.S. 2012. Differences in attitudes of farmers and veterinarians towards pain in dairy cows. The Veterinary Journal 194:94-97.

Ursinus, W.W., Schepers, F., de Mol, R.M., Bracke, M.B.M, Metz, J.H.M. and Groot Koerkamp, P.W.G. 2009. COWEL: a decision support system to assess welfare of husbandry systems for dairy cattle. Animal Welfare 18: 545-552.

Veissier, I., Jensen, K.K, Botreau, R. and Sandøe, P. 2011. Highlighting the ethical choices underlying the scoring of animal welfare. Animal Welfare 20(1):89-101.

Weber, C.N. and Müller, K.E. 2012. Schmerztherapie beim Rind – eine Umfrage unter in der Rinderpraxis tätigen Tierärzten in Deutschland. Teil 2: die Einschätzung von Schmerzen bei Rindern. Tierärztliche Umschau 67: 462-469.

Webster, A.J.F., Main, D.C.J. and Whay, H.R. 2004.Welfare assessment: indices from clinical observations. Animal Welfare 13(S):93-98.

Webster, J., 2005. The assessment and implementation of animal welfare: Theory into practice. Rev. Sci. Tech. Int. Epiz., 24: 723-734.

Welfare Quality®. 2009. Welfare Quality® assessment protocol for cattle. Welfare Quality® Consortium, Lelystad, Netherlands; ISBN/EAN 978-90-78240-04-4.

Whay, H. R., Waterman, A.E. and Webster, A.J.F. 1997. Associations between locomotion, claw lesions and nociceptive threshold in dairy heifers during the peri-partum period. The Veterinary Journal 154(2):155-161.

Whay, H.R., Main, D.C.J, Green, L.E. and Webster, A.J.F. 2003. Animal-based measures for the assessment of welfare state of dairy cattle, pigs and laying hens: consensus of expert opinion. Animal Welfare 12: 205-217.

3. Material and methods

The investigations performed in the pursuit on all three objectives were based on empirical data derived from cross-sectional studies (papers 1, 2, 3, and 4) with an additional register-based follow-up for papers 1, 3, and 4. In the following section the origin of data and data management will be described for each dataset and its' relation to given objective. The fundamental methodological considerations for each study will be given in this section as the specific methods are described in each manuscript.

3.1 Sampling considerations

Selection of herds

The Danish dairy herds have undergone major structural changes over the last decades with a marked decline in the number of herds and a subsequent increase in average herd size. Hence, the number of dairy herds went from 9,800 dairy herds in year 2000 down to 4,062 in 2011; accordingly, herd size doubled in the same period from 65,9 to 148,4 cows (RYK, 2011; Danish Cattle Federation, 2006). In 2011, 2,482 herds had more than 100 cows representing 61% of the Danish dairy herds and more than 80% of the Danish dairy cow population (Kristensen *et al.,* 2010). Consequently, the target population was defined as dairy herds with more than 100 cows. Furthermore, 93% of the dairy herds are assumed to be loose-housing systems of which 93% have cubicles (Kristensen, 2010). In order to meet the characteristics of the target population, the study herds had to meet the following criteria:

1. Herd size greater than 100 cows
2. Loose-housing system
3. Cubicles in the resting area

Sampling of study herds for Objective I (paper 1) was done for previous study with a different study purpose investigating the prevalence of so called "loser cows", defined by cows not being able to keep up with the rest of the herd. Hence, sampling considerations are described by Thomsen et al. (2007). Sample size was calculated based on the formula given for estimating the proportion or prevalence of loser cows:

$$n= \frac{Z^2_{1-\alpha/2}\, p(1-p)}{L^2} \qquad \text{Eq. 1}$$

where n is the required sample size

$Z^2_{1-\alpha/2}$ = 1.96 corresponding to the 95% confidence interval of a two-sided standard normal distribution p is an estimate of the prevalence of interest (estimated loser cow p= 0.05)

L is the maximum allowable error (0.01)

The remaining Objectives II and III and the accompanying papers 2, 3, and 4 were based on data material from a pool of study herds within two collaborating Ph.D. projects investigating dairy cattle welfare. The sampled study herds originated from the same initial pool of herds. These were drawn as random sample of 812 herds amongst the 2,349 registered dairy herds with a herd size larger than 100 cows in the DCD back in 2009 (Kristensen, 2010). These herds were invited to a questionnaire on grazing strategies with 401 herds responding positively. Hence, the first Ph.D. project (grazing-project) focusing on welfare aspects in grazing herds included 42 out of the 131 grazing herds form the initial respondent pool (Burow *et al.*, 2013b). The second, and present, Ph.D. project (welfare-project) drew a random sample of 90 herds from the same initial respondent list (Figure 3.1).

Sample size calculations were made according to the above mentioned formula to estimate the proportion of herds with welfare problems. The proportion of herds with welfare problems was assumed to be p=0.24, according to the results from the official welfare inspections in 2010 (Anonymous, 2011).

Thus, setting p at 0.24, $Z^2_{1-\alpha/2}$ at 1.96 and L at 0.01 yields a sample size:

$$n= \frac{1.96^2 * 0.24(1-0.24)}{0.01^2} = 7{,}007 \text{ cows} \approx 70 \text{ herds (herd size 100 cows)}$$

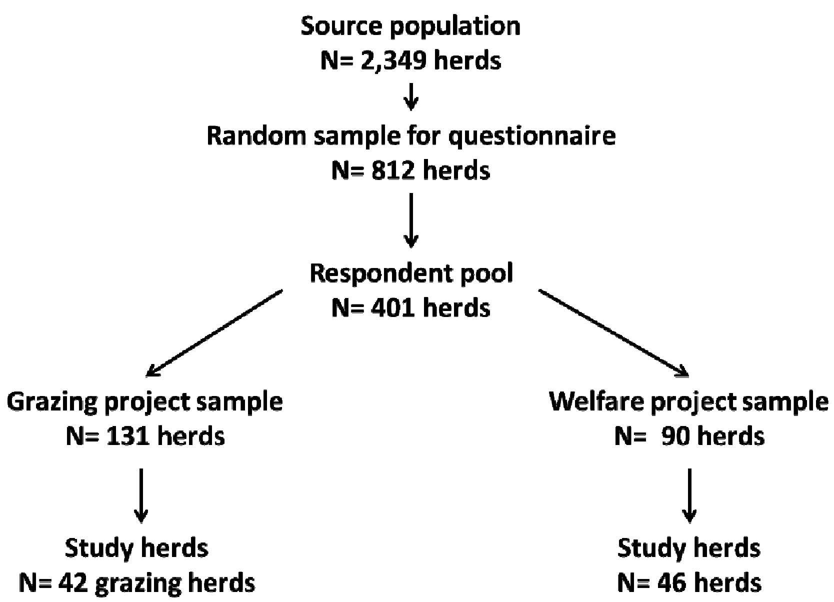


Figure 3.1. Overview of the sampling of herds.

The 90 herds were contacted first by letter in April 2009 followed by a telephone call four weeks later for clearance of participation. Among the 90 herds there was an overlap of 14 herds with the grazing-project, which had already been visited. In total 60 herds accepted to participate and hence, 46 herds were visited by the welfare-project besides the 42 grazing-herds. Additionally, the grazing-project used 20 of the non-grazing herds from the present welfare-project for two studies (Burow *et al.*, 2013a; Burow *et al.*, 2013b). The pool of study herds is illustrated in Figure 3.2.

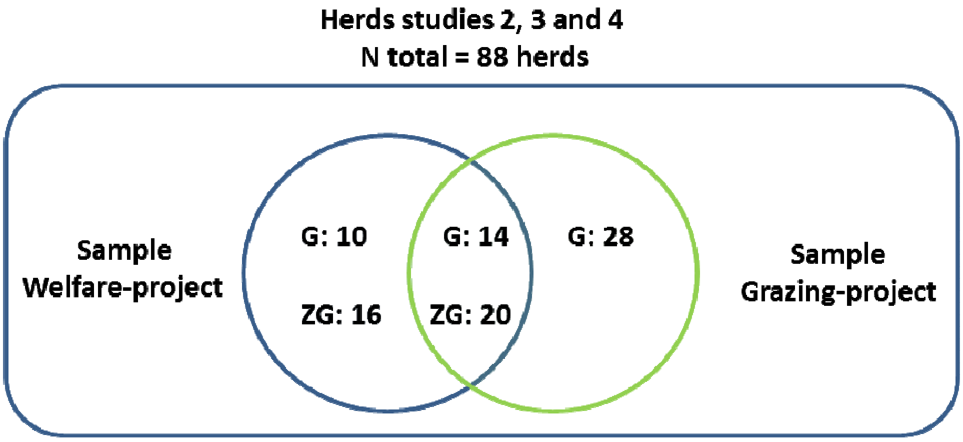


Figure 3.2. The pool of 88 study herds and their origin from two separate samples (Welfare-project and Grazing-project) given as grazing herds (G) and zero-grazing herds (ZG).

Herds sampled within the grazing-project were distributed across Jutland, while herds within the welfare-project were distributed across Jutland, Funen, and Sealand. Study herds for each of the papers 2, 3 and 4 herds were included depending on fulfilment of different criteria for each study (e.g. compliance with overall inclusion criteria, completeness of on-farm registrations and register data). Further details on the selection of herds are described in the individual manuscripts.

Selection of animals

Within-herd sample sizes for papers 2 and 3 were calculated according to Equation 1 with the following assumptions given by the Welfare Quality® protocol: estimated proportion p=0.5, $Z^2_{1-\alpha/2}$ = 1.96 and L=0.1. The sample size was the adjusted for herd size by the formula for sampling from a known population size:

$$n_a = \frac{n}{1+n/N} \qquad \text{Eq. 2}$$

where n_a is the adjusted sample size for population size

n is the required sample size of 96

N is population size

3.2 Data sources and data collection

For the present thesis a number of different datasets were collected, because Objective 1 was pursued in a pilot study prior to the investigations performed in the pursuit of Objectives 2 and 3. Empirical data from on-farm data collection was used in all four studies, however, with different content. Clinical observations were used for papers 1, 2, and 3; behavioural observations were used in paper 3; and system and management data were used in papers 3 and 4. Existing data from routine registrations gathered for other purposes were used for papers 1, 3 and 4. Furthermore, an online questionnaire study was performed for paper 3 in order to obtain expert opinion on welfare indicators for dairy cattle. The following section will provide an overview of specific data sources and the different data collection procedures for the given studies and all studies are summarized in Table 3.6.

3.2.1 Register data within the National Danish Cattle Database

The Danish Dairy Cattle Federation (DCF) manages the overarching database The National Danish Cattle Database (DCD) compiling both mandatory registrations from official databases (public databases e.g. CHR and VetStat) and voluntary registrations by farmers, hoof trimmers, inseminators, breeding organisations, milk recording scheme, slaughterhouses, and laboratories. While the public data are freely accessible, all other data within the DCD are owned by the farmers and the DCF. An overview of data input, data flow is given in Table 3.1.

Table 3.1. Information gathered in the Danish Cattle Database (DCD) from other primary sources including the Central Husbandry Register (CHR), the Registration and Yield Control (RYK), veterinarians, slaughter houses and the person responsible for the registration. Modified from Bundgaard (2005).

Information	Responsible for recording	Primary database	Database Administrator
Birth of animal* (breed, sex, parents) Death* Movement of animals* Production type (conventional/organic)	Farmer	CHR	Danish Veterinary and Food Administration
Salmonella dublin status	Laboratory	DCD	Danish Cattle Federation
Assisted/unass. calving and calf status* Service by bull/insemination Drying off date Body Condition Score Weight recordings (heifers and adults)	Farmer	DCD	Danish Cattle Federation
Artificial Insemination Pregnancy check	Inseminator/Farmer	DCD	Danish Cattle Federation
Disease* Treatments* Death/Euthanasia	Veterinarian/Farmer/Hoof trimmer	Veterinarians administrative system & DCD	Veterinarian/Danish Cattle Federation
Slaughter results	Slaughter houses	Slaughter houses	Kødbranchens Fællesråd
Laboratory results (milk and blood samples)	Laboratories	DCD	Danish Cattle Federation
Breeding and show results	Officials from breeding organisation	DCD	Danish Cattle Federation

** Mandatory recordings required by legislation*

3.2.2 Data collection of register data variables

Three out of the four papers, namely papers 1, 3 and 4, involved data extracted from the DCD. An initial list of required variables was established based on a literature review (Table 3.2) on register data and their associations with animal welfare issues (Table 3.3). The initial list of variables ordered from the DCD was identical for both the pilot study (study 1) and the following studies 2, 3, and 4 (Appendix A). Data were extracted from the DCF for a number of specified herds for the following periods for paper 1 from 1st of July 2003 to 1st of December 2004 (40 herds) and for paper 3 and 4 for the period from the 1st of January 2008 to 1st of January 2012 (80 herds, see papers 2 and 3 for details). Data were received in ten separate data sets according to their origin as milk yield, milk quality, reproduction, treatments, calving, culling/movements, and abattoir data (Appendix A).

Table 3.2 Overview of proposed welfare indicators from register databases found in literature.

Study	**Country**	**Indicators**
deVries *et al.* (2009)	NL	Lameness prevalence Milk yield Fat/protein content Calving-to-service interval Non-return rate Number of services Average age of herd Culling rate
Sandgren *et al.* (2009)	S	Percentage cows with late on going AI (> 120 days) Percentage heifers without mating/AI > 17 month of age Calf mortality 2-6 month
Welfare Quality® (2009)	EU	Somatic cell count Dystocia cases Downer cows Mortality rate Culling rate Life expectancy
Anonymous (2009) ”Arbejdsgrupperapport om hold af malkekvæg”	DK	Mortality (cows/calves) Premature culling rate Drug consumption Painful conditions (traumatic reticuloperitonitis, toxic mastitis, lameness) Hoof related disorders Abattoir findings (emaciation, fractures, chronic infections, liver cirrhosis) Mastitis cases, new Milk fat content
Anonymous (2003) “Kvægproduktion 2010”	DK	Mortality Premature culling Painful conditions Mastitis cases, new
Vaarst & Nissen (2006) “Koliv 100”	DK	Mortality Life expectancy (no of lactations) Percentage abattoir remarks Percentage beef carcase classification 1 Average interval from calving to slaughter Percentage acutely and chronically increased SCC
Main *et al.* (2003)	UK	Annual average milk yield Calving-to-service interval Annual mastitis cases per 100 cows/year Annual assisted calvings
Whay *et al.* (2003)	UK	Mastitis cases per 100 cows/year Lameness cases per 100 cows/year

3.2.3 On-farm recordings

Animal observations

In order to meet Objective I by evaluating the potential of register data to predict cross-sectional findings an existing data set with clinical scorings of 40 Danish dairy herds was used in the pilot study (paper 1). On-farm recordings were performed by one observer during September 2003 to October 2004. The complete clinical protocol is described in Thomsen et al. (2007), however, only lameness scores were included in the present study 1. Lameness was initially scored on a five-point ordinal scale (Sprecher *et al.*, 1997) and recorded for 95% of the cows at three visits with a 120 day interval.

For Objectives II and III a clinical and behavioural observation protocol was developed based on a modified version of the Welfare Quality® assessment protocol for dairy cattle (Welfare Quality®, 2009). Modifications were made due to seasonality and/or low prevalences for vaginal, ocular and nasal discharge, and hampered respiration were excluded; while behavioural measures of social and antagonistic behaviour and the Qualitative Behaviour Assessment (QBA) were excluded due to time constraints. Additionally, measures of hair coat and rising-behaviour were added due to better feasibility and their suggested potential as welfare indicators in literature in order to ensure coverage of as many WQ® criteria (Table 2.3) as possible. In total six criteria of the twelve WQ® were covered. The on-farm assessments were performed during April 2010 to July 2011 by four trained observers with varying levels of experience in clinical and behavioural scoring of cattle (2-6 years). Observers were introduced to the measures on two on-farm training sessions and calibration tests were done during one photo and video session prior to data collection and upon completion of data collection.

Resource measures

Resource measures were assessed according to the CORE organic ANIPLAN (2011) manual. This manual provides a very comprehensive assessment of system characteristics. The final protocol was supplemented to fit Danish production settings in alignment with Danish animal welfare legislation (Appendix B). The final protocol consisted of 127 measures regarding feed and water supply, resting and walking area, additional barn equipment (e.g. brushes, harmful or damaged equipment, hoof trimming box), and separation of sick/weak or calving animals. Resource measures were covering six criteria as listed in Table 3.3. In total nine out of the twelve WQ® criteria were covered except for the criteria concerning *thermal comfort*, *absence of pain* (not applicable in Denmark), and *positive emotional state*.

Table 3.3. Welfare measures assessed on-farm in Danish dairy cattle herds during 2010-2011.

	Welfare measure	WQ® criteria
Clinical	Hygiene – leg, hind, udder	Comfort around resting
	Integument alterations – hock, carpal, body	Absence of injuries
	Body condition score	Absence of prolonged hunger
	Overgrown claws	Absence of injuries
	Faeces score	Absence of prolonged thirst
	Lameness	Absence of disease
	Hair coat	Absence of disease
Behaviour	Rising behaviour	Ease of movement
	Avoidance distance	Good human-animal relationship
	Lying-down – duration and collisions	Comfort around resting
Resources	Water supply	Absence of prolonged thirst
	Water cleanliness	Absence of prolonged thirst
	Feeding slots – dimensions	Absence of prolonged hunger
	Occupancy rate bed stalls	Comfort around resting
	Bed stall length	Comfort around resting
	Bed stall width	Comfort around resting
	Passage ways	Ease of movement
	Passage way – Width	Ease of movement
	Passage way – Skid resistance	Ease of movement
	Passage way – Flooring	Absence of injuries
Resources	Dead ends	Ease of movement
	Calvin pen size	Expression of other behaviours
	Separation of animals	Expression of other behaviours
	Sick animals not in sick bay	Expression of other behaviours
	Harmful/damaged equipment	Absence of injuries
	Brushes	Expression of other behaviours
	Scraping system	Absence of injuries
	Grazing	Expression of other behaviours

3.3 Data management and specification of variables

3.3.1 Register-based welfare indicators

The pilot study was investigating the annual means for the 35 variables listed in Table 3.4 for the year 2004. First, eleven new sub-sets of the initial datasets (Appendix A) were created containing only registrations from the reference year 2004. Secondly, variables were created within each of the respective eleven sub-sets and finally, merged into a single data set based on the herds CHR number, yielding a dataset with 40 observations for each of the 35 variables (Figure 3.3). The following variables were excluded due to very low prevalences: stillborn, abattoir remarks on fresh fractures, joint luxation and bruises/beating marks and due to high uncertainty concerning registration validity: acute/chronically increased SCC, calving-to-service interval, leaving 28 variables for further investigation within papers 1 and 3. Paper 4 only investigated 17 variables based on the previous descriptive results in paper 3. Variables were calculated as within-herd means for a specified time period within paper 3 (365, 180 and 90 days prior to on-farm data collection) and study 4 (365 days prior to on-farm data collection).

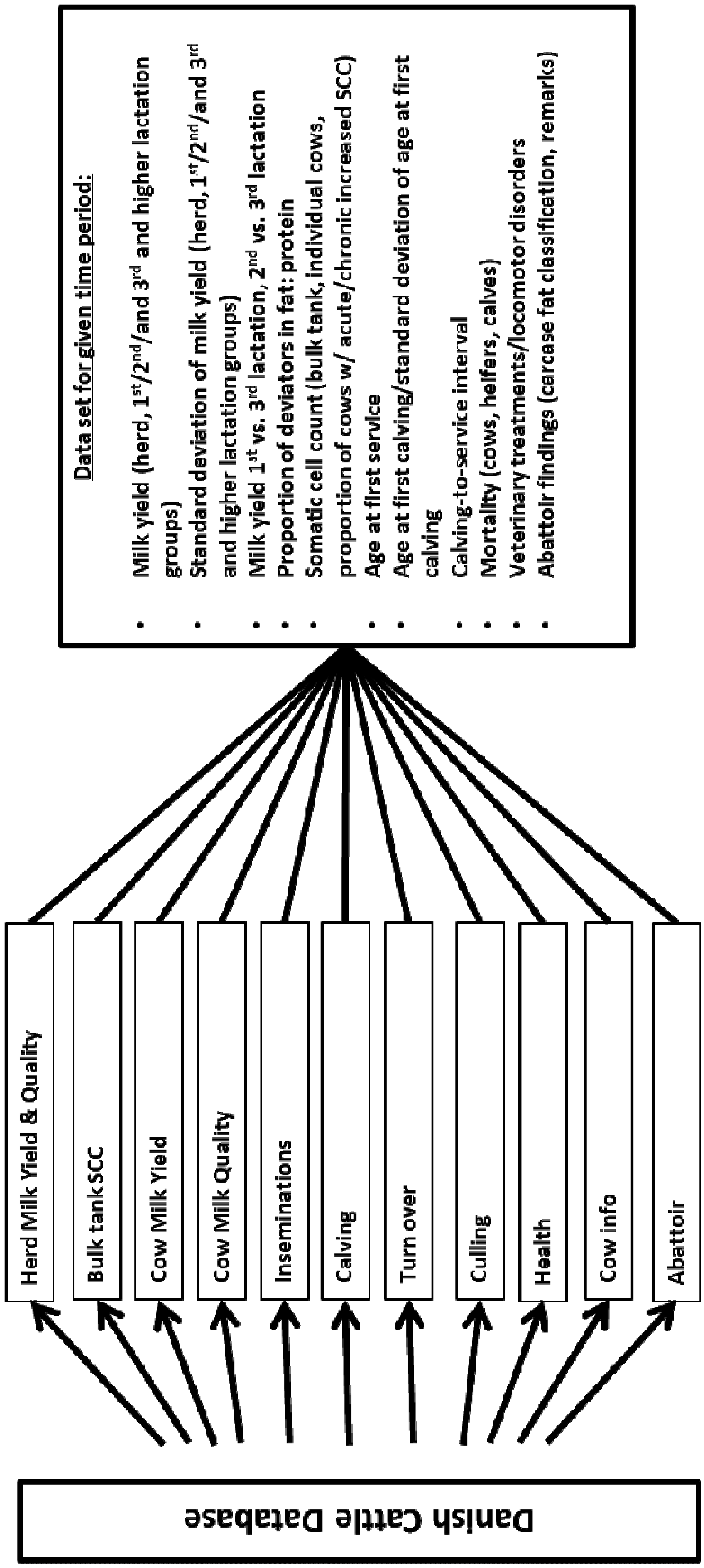


Figure 3.3 Illustration of data-flow from the central Danish Cattle Database to study data set.

3.3.2 Resource-based indicators

A total of 127 resource-based measures were assigned to 16 overall welfare criteria and level scores were assessed according to Danish legislation concerning housing aspects of dairy cattle (Act on the Keeping of dairy Cattle and Offspring of Dairy Cattle, 2010) and the general obligations of good care of animals in the Animal Welfare Act (2007) (Table 3.5). The subsequent assessment involved coding of 16 overall resource criteria into based compliance with the given legislation ranging from 0 (full compliance) to 1 (absolute non-compliance) for the individual measures. This assessment was done using a range rather than a binary system to avoid the use of thresholds e.g. in cases where barn or equipment designs and dimensions differ within sections of the same barn. Hence, a herd with old cubicles not fulfilling the latest minimum standards in one part of the barn and a section with new compliant cubicles would be assessed according to the relative non-compliance, e.g. if one third of cubicles were old and non-compliant the herd would receive a score 0.33 for cubicle length/width. Although, the legislative context referred to in the Act on the keeping of dairy Cattle is subject to different transition periods until full implementation in 2034, and the current assessment operates under the assumption of fulfilling the legislation regardless of exceptions during transition periods.

3.3.3 Clinical and behavioural observations

The clinical protocol consisted of 11 measures evaluated on a qualitative scale (Table 3.6, Appendix C). For the ten graded measures according to either a normal level (score 0), a moderate (score 1) or a severe (score 2) impairment of the given sign. Claw conformation was assessed as presence of overgrown claws or normal claws. Qualitative measures or measures with more than three levels, i.e. lying-down duration and collisions, rising behaviour and BCS, were transformed into the same three categories of normal, moderate or severe impairment as described in Table 3.5. Finally, all within-prevalences of the given levels for all measures were calculated using the proc freq procedure in SAS© 9.3 (SAS Institute Inc., SAS Campus Drive, Cary, North Carolina). These within-herd prevalences of moderate and severe levels were used to determine the level of animal welfare impairment due to high levels of welfare related disorders. Due to a large number of missing values for lying-down, the measure was excluded from further analyses.

3.4 Data analysis

3.4.1 From measuring to assessing welfare – aggregation models

In order to evaluate the overall animal welfare status of dairy herds as specified in Objective III, an overall quantification model of animal welfare was needed. Although, the Welfare Quality® protocol holds both resource- and animal-based measures, this aggregation approach was not suitable for the study purposes within this Ph.D. project. Mainly because the present objective was to evaluate the potential of welfare models with different information sources, an index providing the possibility of comparing ranks of herds based on the different information sources was needed. This also implied applying a more separate aggregation process instead of the integrated Welfare Quality® approach, combining resource- and animal-based measures into one index. Secondly, a transparent linear model was preferred over the complicated non-linear approach. These considerations lead to the development of the additive and linearly weighted index described in paper 3. However, since the different measures do not have an equal effect on welfare impairment, individual measure weights were essential.

Measure weights from expert opinion

Measure weights could have been derived from literature studies on risk factors for single clinical or behavioural measures; however, this approach would be troublesome due to a couple of reasons. Firstly, due contradictory findings in different studies making interpretation difficult (Rushen, 1991; deVries *et al.*, 2011) either due to differences in case definitions (as described in sections 2.3.2 and 2.3.3) or due to differences in data collection. Another and more feasible approach is the use of expert opinion (Collins, 2012), an approach also used within the Welfare Quality® work (Bonde *et al.*, 2009). Hence, an online expert survey was performed during December 2011 to January 2012, initially inviting 32 experts within dairy cattle production and welfare as described in paper 3. The survey entailed two separate rounds; the first round concerned the direct animal-based measures, while the second round concerned the register- and resource-based measures. The response rate in the first round was 63% (20 out of 32), while only respondents from the first round also were invited to the second round, yielding a response rate of 75% (15 out of 20). In order to gain as many different views on animal welfare a wide range of experts were invited coming from the industry, animal welfare scientists, practioners, welfare control officers, agricultural advisors and animal protection organisations. All groups were represented among the respondents in both rounds. Basically, the overall content of the surveys was structured identically with closed questions regarding the relative weighting of measures in the context of a welfare assessment to obtain the measure weights (W). For the animal-based measures, additional semi-open questions were used to obtain the level weights for moderate (MW) and severe (SW) levels.

Aggregation models

Hence, the aggregation model for the animal welfare index (AWI) based on the non-graded register- and resource-based measures (N) became the sum of all weighted measures:

$$AWI = \sum_{j=1}^{k} \mathrm{N_j W_j}$$

Measures within the index for register-data were categorized and scored according to percentiles within the study population; score 0 for values among the 25% best herds, category 1 for values between the 25-75%, score 2 for values between the worst 10–25%, and score 3 for values among the worst 10%. The resource-based measures were used as continuous numbers reflecting the range of compliance (0 to 1).

The slightly more complex aggregation model for the animal-based welfare index was based on measure prevalences and included the relative measure level weight for each graded measure level (moderate and severe) before adding on the measure weight and summing these up together with the non-graded measure for overgrown claws:

$$AWI = \sum_{i=1}^{k} (\mathrm{M_i\, MW_i} + \mathrm{S_i})\mathrm{W_i} + \sum_{j=1}^{k} \mathrm{N_j W_j}$$

In conclusion, the higher the index value of a given herd is, the higher levels of welfare impairment are present either due to production results from register data relative to the study population, potentially higher impairment due to negative housing consequences reflected in the resource-based index, or due to high levels of welfare related disorders in the animal-based index.

3.4.2 Statistical analyses

Regression models

Table 3.7 provides an overview of the statistical approaches used in the pursuit of the three objectives. Basically, all associations between explanatory variables and outcome were assessed by either uni- and multivariable models based on linear and logistic regression depending on the outcome definition (papers 1, 3, and 4). The multivariable models were built in stages based on the strategy described by Hosmer & Lemeshow (1989). First, all associations between explanatory and the respective outcome variables in the univariable analyses were evaluated and only explanatory variables within the significance level of $p < 0.2$ were included in the next step. Finally, a backwards elimination process was used to only regard variables adding significance at a 0.05-level in the model. Model fit was evaluated by assessing residuals and goodness-of-fit test (Shapiro Wilk-test) for fulfilling the assumptions of normal distribution, variance homogeneity and independence of observations. Model selection was based on the Aikaike Information Criterion (AIC), where models were selected for the lowest AIC.

Receiver Operating Characteristic (ROC) curves

Quantification of diagnostic or predictive potential in paper 1 was assessed by evaluating and analysing Receiver operating characteristics (ROC) curve and the area under this curve (AUC). The ROC curve plots the sensitivity against 1-specificity (i.e. the false positive rate) at varying cut offs. This approach can be used to identify optimal cut offs maximizing either sensitivity, specificity or both, i.e. the differential positive rate (DPR=(Se + Sp)-1).

Reliability

Inter-observer agreement in paper 2 was evaluated by the prevalence-adjusted bias-adjusted kappa (PABAK) (Byrt *et al.,* 1993) taking the categories of the given measure into account: PABAK= [((k*p)-1) / (k-1)], where k is the number of categories and p the proportions of matches.

Latent class analysis (LCA) and Bayesian modelling

In paper 2 the LCA approach was used to estimate the true lameness prevalence in different populations by implementing a Bayesian model. The approach was chosen, as four observers were responsible for the lameness scoring introducing misclassification bias. Hence, none of the four observers could be regarded as a 'perfect test' and therefore the LCA offers the possibility to account for observers' sensitivity and specificity and estimate the true prevalence based on these conditions (Hui and Walter, 1980). However, the LCA model presents three underlying assumptions: (i) tests (in this case observer) are regarded as conditionally independent; (ii) the test characteristics (Se and Sp) are constant across populations; and (iii) prevalences for the differing populations should also be different. The true prevalences were estimated in a Bayesian model regarding Se and Sp estimates for each observer based on data from calibration study. Finally, a Bayesian risk factor model was established to investigate the differences in true prevalences between two populations (i.e. grazing and non-grazing herds).

Correlation analysis

The non-causal relationship between the rankings of herds based on their animal welfare score within three different indices in paper 3 was assessed by the non-parametric Spearman Rank Correlation Coefficient, quantifying the degree of linear association between the ranks of herds within each index.

References

Anonymous. 2003. Projektplan 2003 - Kvægproduktion 2010. Available online: https://www.landbrugsinfo.dk/Kvaeg/Filer/projektplan.pdf . Accessed 01-10-2013.

Anonymous. 2009. Arbejdsgrupperapport om hold af malkekvæg. Justitisministeriet. Available online: http://www.dyrenesbeskyttelse.dk/sites/default/files/rapporter%20o.lign./Rapport_om_hold_af_malkekvaeg.pdf. Accessed 01-10-2013.

Anonymous. 2011. Kontrol af dyrevelfærd 2010. ©Fødevarestyrelsen, 1. Edition, GP-Tryk, Copenhagen, Denmark.

Bonde, M., Boutreau, R., Bracke, M.B.M., Butterworth, A., Capdeville, J., De Rosa, G., Engel, B., Keeling, L.J., Perny, P., Rousing, T., Sørensen, J.T., Van Reenen, K. and Veissier, I. 2009. Final model of multicriterion evaluation of animal welfare. Deliverable 2.28, subtask 2.3.1.3, EU Food-CT-2004-506508.

Bundgaard, E. 2005. Registreringernes vej fra gården til kvægdatabasen (In Danish). Pages 26-27 in Dansk Kvæg kongres.

Burow, E., Thomsen, P.T., Rousing, T. and Sørensen, J.T. 2013a. Daily grazing time as a risk factor for alterations at the hock joint integument in dairy cows. Animal 7(1): 160-166.

Burow, E., Thomsen, T., Rousing, T., Otten, N.D. and Sørensen, J.T. 2013b. Effect of grazing on the cow welfare of dairy herds evaluated by a multidimensional welfare index. Animal 7(5): 834-842.

Byrt, T., J. Bishop, J.B. Carlin. 1993. Bias, prevalence and kappa. Journal of Clinical Epidemiology 46(5):423-429.

Chaplin, S. and Munksgaard, L. 2001. Evaluation of a simple method for assessment of rising behaviour in tethered dairy cows. Animal Science 72: 191-197.

Corazzin, M., Piasentier, E., Covier, S., and Bovolenta, S. 2010. Effect of summer grazing on welfare of dairy cows reared in mountain tie-stalls barns. Italian Journal of Animal Science 9(3): 304-312.

CORE organic ANIPLAN. 2011. CORE organic ANIPLAN on-farm manual for loose housed dairy cows. ANIPLAN project no. 1903. Accessed October 2013. http://www.coreorganic.org/research/projects/aniplan/index.html.

Danish Cattle Federation 2006. Figures on Danish Cattle. Available online: https://www.landbrugsinfo.dk/Kvaeg/Tal-om -kvaeg/Sider/kvbrug_ital2006ukweb.pdf?download=true. Accessed 10-02-2013.

deVries, M., Bokkers, E.A., Dijkstra, T., van Schaik, G. and de Boer, I.J. 2009. Potential of routinely collected farm data to monitor dairy cattle welfare. In: Book of abstracts (no 15) of the 60th Annual Meeting of the European Association for Animal Production, August 24th-27th 2009, Barcelona, Spain.

deVries, M., Bokkers, E.A., Dijkstra, T., van Schaik, G. and de Boer, I.J. 2011 Invited review: associations between variables of routine herd data and dairy cattle welfare indicators. Journal of Dairy Science 94: 3213-3228.

Hosmer, D.W. & S. Lemeshow (1989): Applied Logistic Regression. Wiley, New York, p.301.

Hui S.L. and Walter, S.D. 1980. Estimating the error rates of diagnostic tests. Biometrics. 36: 167-171.

Kristensen, T., Madsen, M.L. and Noe, E. 2010. The use of grazing in intensive dairy production. Grassland Sciences in Europe 15: 964-966.

Kristensen, T. 2010. Produktionssystemer i de danske malkekvægbedrifter. Landbrugsinfo 2117. Available online: https://www.landbrugsinfo.dk/Kvaeg/Foder/Grovfoder/Afgraesning/Sider/Produktionssystemeridanskemalkekvaegbedrifter.aspx. Accessed 10-02-2013.

Main, D.C.J., Whay, H.R., Green, L.E. and Webster, A.J.F., 2003. Preliminary investigation into the use of expert opinion to compare the overall welfare of dairy cattle farms in different farm assurance schemes. Animal Welfare, 12, 565-569.

Rushen, J. 1991. Problems associated with the interpretation of physiological data in the assessment of animal welfare. Applied Animal Behaviour Science 28: 381-386.

RYK Ydelseskontrollens Årsberetning 2011. Available online: https://www.landbrugsinfo.dk/Kvaeg/RYK/Sider/Aarsopgoerelse-fra-Ydelseskontrollen-2011-12.pdf?download=true. Accessed 10-02-2013.

Sandgren, C.H., Lindberg, A. and Keeling, L.J. 2009. Using a national dairy database to identify herds with poor welfare. Animal Welfare 18: 523-532.

Sprecher, D.J., Hoesteler, D.E., Kaneene, J.B., 1997. A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance. Theriogenology 47, 1179–1187.

Thomsen, P.T., Østergaard, S., Sørensen, J.T: and Houe, H. 2007. Loser cows in Danish dairy herds: Definition, prevalence and consequences. Preventive Veterinary Medicine 79: 116-135.

Vaarst, M. & T.B. Nissen. 2006. Manual til Koliv 100 – Vurdering af køernes velfærd i økologiske besætninger. Økologisk Landsforening og Danmarks Jordbrugs Forskning. Trykkestedet, Århus N, Denmark.

Welfare Quality®. 2009. Welfare Quality® assessment protocol for cattle. Welfare Quality® Consortium, Lelystad, Netherlands; ISBN/EAN 978-90-78240-04-4.

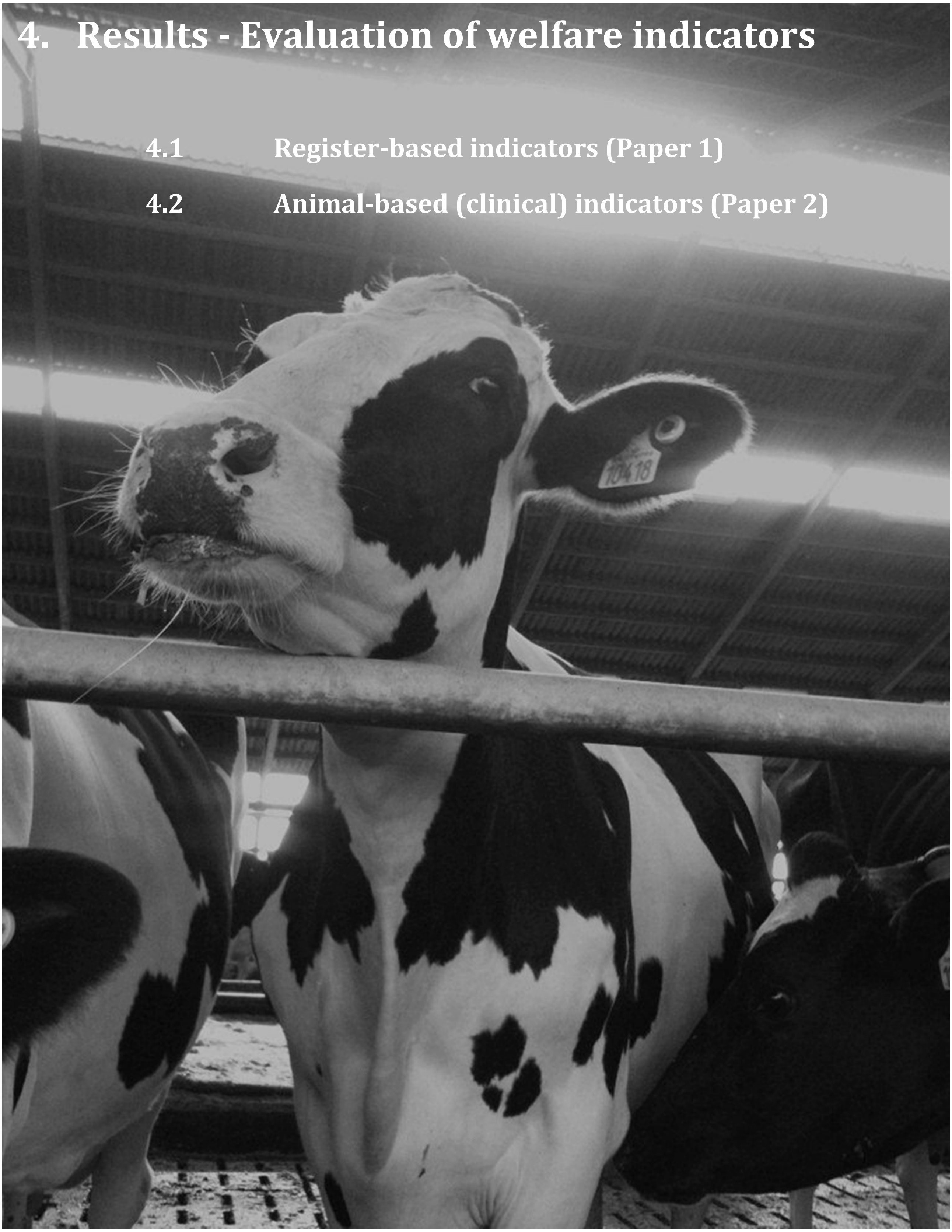
Whay, H.R., Main, D.C.J, Green, L.E. and Webster, A.J.F. 2003. Animal-based measures for the assessment of welfare state of dairy cattle, pigs and laying hens: consensus of expert opinion. Animal Welfare 12: 205-217.

Wildman, E.E., Jones, G.M., Wagner, P.E., Boman, R.L.,Trout, H.F. and Lesch, T.N. 1982. A dairy cow body conditioning scoring system and its relationship to selected production variables in high producing Holstein dairy cattle. Journal of Dairy Science 65:495.

4. Results - Evaluation of welfare indicators

4.1 Register-based indicators (Paper 1)

4.2 Animal-based (clinical) indicators (Paper 2)



Paper 1 – Register-based indicators

Evaluation of the performance of register data as predictors for dairy herds with high lameness prevalence

Nina Dam Otten[1], Nils Toft[1], Peter T. Thomsen[2], Hans Houe[1]

[1] Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Groennegaardsvej 8, DK-1870 Frederiksberg C, Denmark

[2] Department of Animal Science, Aarhus University, P.O. Box 50, DK-8830 Tjele, Denmark

The manuscript has been resubmitted in this revised form to Animal on December 2nd 2013

Evaluation of the performance of register data as predictors for dairy herds with high lameness prevalence

Nina Dam Otten[1], Nils Toft[1], Peter T. Thomsen [2], Hans Houe [1]

[1] *Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Groennegaardsvej 8, DK-1870 Frederiksberg C, Denmark*

[2] *Department of Animal Science, Aarhus University, P. O. Box 50, DK-8830 Tjele, Denmark*

Short title: Register data as predictors for dairy cow lameness

Submitted to *Animal*

Abstract

The present study evaluated the diagnostic performance of four routinely registered predictors from the central Danish Cattle Database (DCD) for identifying herds with high lameness prevalence in 40 Danish dairy herds. The predictors were extracted as within-herd annual means for a one-year period for mortality, bulk tank somatic cell count, proportion of lean cows at slaughter and the standard deviation (SD) of age at first calving. The target condition "high lameness prevalence" was defined by a within-herd prevalence ≥ 16% (third quartile). Diagnostic performance was evaluated by constructing and analysing Receiver operating characteristic (ROC) curves and their area under the curve (AUC) for single predictors and predictor combinations. Sensitivity (Se) and specificity (Sp) of the explanatory variables (i.e. predictors) were assessed at the optimal cut-off based on data and compared to a set of pre-defined cut-off levels (national annual means or 90-percentile). Highest AUC for single predictors was obtained by mortality (AUC 0.76), while adding on the other three predictors did not yield a significant increase in AUC. Mortality and SD of age at first calving yielded highest Se (100%, 95% CI: 72-100), while highest Sp was found for the proportion of lean cows at slaughter (83%, 95% CI: 66-93). The highest differential positive rate (DPR=0.53) optimizing both Se and Sp was found for mortality. Optimal cut-off points were lower than the pre-defined cut-offs, emphasizing the importance of a clear distinction of the purpose in the use of these data: whether register data are intended as predictors for a given problem or as a problem definition on its own i.e. in risk-based surveillance. Hence, optimized cut-offs should be preferred in official surveillance and control schemes in order to enhance the accuracy of the identification of herds with problematic levels of clinical findings.

Implications

The modern dairy industry routinely generates data on production and disease. In Denmark, the dairy industry owns the Danish Cattle Database acting as an overarching centralised database compiling both official and privately owned data. These data may have great potential for predicting certain states of welfare and health of livestock herds, because they are much cheaper to obtain than many other health and welfare indicators. Therefore, the use of these cheap and at times even “free” data to predict a given state of welfare in a cost-effective manner is evaluated in the present study. Hence, optimising the use of these register data could be used in the pursuit on identifying herds at risk of having animal welfare problems.

Introduction

Estimating the national welfare level of e.g. dairy herds would require assessments of all herds as on-farm assessments measuring the direct clinical and behavioural state of animals are still perceived to be closer to the true state of welfare. However, they are very costly and time-consuming due to investigator training and calibration and the actual time needed on location. Hence, a more targeted approach is needed. This approach could target herds at risk of having welfare problems, subsequently reducing the amount of herds visited. For this purpose, the potential of register data could be utilised, as direct consequences of clinical manifestations are reflected in e.g. milk production and reproductive results within register data. The Danish welfare control programme uses register-based predictors to identify livestock herds at risk of welfare problems based on a set of risk parameters from the national databases. This initial screening is followed by a control visit by the authorities in selected herds. The initial screening is based on certain cut-offs for the given parameters, but there is a need to investigate how sensitive these cut-offs are and how optimized cut-offs would perform instead.

Within modern livestock production vast amounts of data are generated and routinely recorded in databases. Data like disease recordings and production results are of great value for epidemiological research and have traditionally been used in e.g. investigating risk factors. Over the past decade non-specific routine registrations (i.e. secondary data) have also become of interest in so-called syndromic surveillance schemes (Perrin *et al*., 2012; Elbers *et al*., 2009; Brouwer *et al*., 2012). Various register-based indicators such as treatment records have previously been used to predict clinical manifestations in dairy cattle. Milk production data, e.g. milk yield, bulk tank somatic cell count (SCC) and fat/protein content have been studied intensely for associations with several outcome measures like mortality (Smith *et al*., 2000; Thomsen *et al*., 2006), lameness related diseases (Alban *et al*., 1996; Green *et al*., 2002) and metabolic disorders (Duffield *et al*., 1997; Heuer *et al*., 1999). A recent review by deVries et al. (2011) investigating associations between register-based variables and welfare indicators from the Welfare Quality® assessment protocol found contradicting reports on the associations between variables related to productivity and the welfare indicators. This lead to the conclusion, that even though numerous associations between register-based predictors and direct clinical and mental conditions exist, their potential of estimating levels of animal welfare is not fully understood and limited to only a few aspects of the multi-dimensional complex of animal welfare.

In Denmark, all livestock herds are obliged to register birth, death and movement of animals to the Central Husbandry Register (CHR). All treatments with prescription drugs performed by either a veterinarian or the farmers must be reported to the national database VetStat. Additionally, the industry database, the Danish Cattle Database (DCD), compiles data from the official databases, the milk recording scheme, breeding associations, laboratory findings and abattoirs. In order to explore the opportunities of predicting a direct physiological state on a given day, we tried to combine information from different origin in order to answer the question, whether register data are able to predict clinical observations, in our case high lameness prevalences. Therefore, the objective of the present study was to investigate the predictive potential of different register data predictors to identify dairy herds with high lameness prevalence and compare predefined cut-offs with optimized cut-offs of the given predictors.

Materials and methods

In this study, data from a previous study was used to establish the target condition of high within-herd lameness prevalence in order to evaluate the potential of the register based predictors. In the previous study, Thomsen et al. (2007) observed lameness (among other clinical signs) to calculate the prevalence of so called loser cows in Danish dairy herds.

Herds

The study population was based on a list extracted from the DCD of all herds meeting the following criteria: loose-housing system, more than 100 cows of which more than 95% of cows were Danish Holstein, herds enrolled in conformation scoring of the cows by a breeding association and enrolled in a milk recording scheme, a minimum of 0.1 cases of mastitis recorded per cow per year and furthermore a maximum distance to the Research Centre Foulum of 150 km. A random sample of 40 herds was drawn among the 274 herds fulfilling these inclusion criteria. All farms were visited over a one year period from September 2003 to October 2004 and cows were clinically examined three times with a three to four month interval (for more details, see Thomsen *et al*., 2007).

Clinical protocol – target condition

The lameness score used was a five point ordinal scale described by Sprecher et al. (1997) ranging from a score 1 for normally walking cows to a score 5 for severely lame cows. Cows with a score 4 and 5 were classified as "lame". Lameness of all cows was evaluated at the three herd visits yielding an overall mean within-herd lameness prevalence of 12.9% ±9.88 (SD) and a median of 11%. In order to evaluate the predictors in a surveillance scheme, the obtained mean herd level prevalence was dichotomized using the third quartile as a cut-off. This lead to a classification of herds into either having a low lameness level for herds having a mean prevalence of lame cows below 16%; or a high lameness level (the target condition) for herds with a prevalence ≥ 16%.

Register-based predictors

The predictors were extracted from the DCD for the year 2004 as annual means per 100 cow years (the sum of feeding days of all cows per herd for the corresponding 365 days). Initial uni- and multi-variable screening of register based predictors identified eight predictors significantly associated with lameness. Mortality, bulk tank SCC, proportion of lean cows at slaughter (cows with fat score 1 according to the EU Beef Carcase Classification Scheme) and standard deviation (SD) of age at first calving were chosen for further analysis.

Predictors were assessed in two different models: a *data-driven model* evaluating predictors measured as continuous variables and a *predefined cut-off model* based on dichotomization of predictors based on predefined cut-offs. These predefined cut-offs were set reflecting national means for the predictors mortality, SCC and SD of age at first calving (Table 1) as stated by the Danish Knowledge Centre for Agriculture and the dairy industry. The cut-off for the predictor lean cows at slaughter was not based on the national mean, as this was as low as 15% compared to the variable mean of 24% in the current sample. Hence, the cut-off was chosen to reflect the national 90th-percentile instead.

Table 1. The national means used as cut-offs at herd-level for dichotomizing continuous explanatory variables associated with high lameness prevalence in 40 Danish dairy herds.

Variables	Cut-off	
	Pre-defined	Optimised
Mortality [a] (%)	5.7	3.6
Bulk tank SCC [b] (x1000 cells/mL)	245	214
Lean cows at slaughter [c] (%)	40	44
Standard deviation of age at first calving [d] (month)	2.1	2.0

[a] Mortality = annual mean mortality per 100 cow years; [b] Bulk tank SCC= annual mean bulk tank somatic cell count based on monthly or bimonthly recordings; [c] Lean cows = 90th-percentile of cows with fat score 1 according to the EU Beef Carcase Classification; [d] SD age at first calving = annual mean standard deviation

Statistical analyses

Data-driven model

The optimal cut-offs maximizing the differential positive rates (DPR = Se + Sp - 1) of each predictor were identified by analysing Receiver operating characteristic curves (ROC). The predictive potential was quantified by assessing the area under the curve (AUC) (Hanley & McNeil, 1982). All ROC analyses were made in *R* (R Development Core Team, 2012) using the

R-package Epi() (Carstensen *et al*., 2012). All predictors were assessed individually followed by the assessment of different predictor combinations. Model selection was based on comparing the Aikaike Information Criterion (AIC) (significance level of $p < 0.05$) of the given models.

Sensitivity and specificity estimates and their 95% confidence intervals (95% CI) at the optimal cut-off points were determined. Sensitivity (Se) was defined as the fraction of herds with a predictor level above the given cut-off among the truly identified herds with high lameness level, i.e. with a prevalence of lame cows ≥ 16%. Specificity (Sp) was defined as the fraction of herds with a predictor level below cut-off among the truly identified herds with low lameness level (< 16%).

Predefined cut-off model

AUC, Se and Sp estimates for combinations of dichotomized predictors were assessed equal to the data-driven model.

Results

A descriptive summary for the predictors is given in Table 2.

Data-driven model

Individual evaluation of the predictor variables resulted in mortality having the highest AUC of 0.76, followed by SCC with an AUC of 0.73 (Table 3). Furthermore, only the AUC´s for mortality and SCC were significantly different from 0.5 (the random AUC of a test with no information). Combining the variables yielded only a slight and non-significant improvement of the AUC (0.79) (Table 4, Fig.1). Cut-offs maximizing DPR are given in Table 4. Highest Se was found for mortality (100%, 95% CI: 72-100) and SD of age at first calving (100%, 95% CI: 72-100), but with fairly low corresponding Sp (mortality 53%, 95% CI: 36-70; SD age at first calving 23%, 95% CI: 12-44) (Table 5). Highest Sp was found for the proportion of lean cows at slaughter (83%, 95% CI: 66-93), but with the lowest corresponding Se (40%, 95% CI: 17-69). The cut-offs for the continuous variables identified by the optimization approach were consequently lower than the predefined cut-off except for the proportion of lean cows at slaughter (44%).

Predefined cut-off model

When models combining predictors based on predefined cut-offs were evaluated, the model containing mortality, SCC and lean cows gave the highest AUC (0.71). Figure 2 illustrates the consequences for correct classification of herds by the predefined cut-offs compared to the data-driven cut-offs for maximizing the AUC, Se and Sp.

Table 3. The area under the curve (AUC) and the p-value for differences from the area under a random and non-informative receiver operating characteristic (ROC) curve where AUC=0.5.

Predictor	AUC	*p*-value
Mortality	0.76	0.001
Somatic cell count	0.73	0.01
Lean cows	0.53	0.82
SD age at first calving	0.50	0.95

Discussion

The present study showed that register based secondary data hold a predictive potential for discriminating between high and low prevalences of a clinical direct measure, such as lameness The best combinations of sensitivity and specificity were found for mortality (Se 100%, Sp 53%) and SCC (Se 80%, Sp 63%). For the two other predictors Sp was high but at the expense of a low Se. The data-driven cut-offs with maximized DPR were lower than the predefined cut-offs. This shows that optimal cut-offs are dependent on the sample population and hence extrapolation to the general population should only be made with great care.

The general loss in AUC and DPR values for the predefined cut-offs illustrates the pit-falls of using means or norms. Furthermore, it became obvious how essential it is to establish whether the chosen predictor should be used as a predictor and hence be used with optimal cut-offs rather than being investigated as a risk factor with predefined thresholds. If the latter is the case, the trade-off between Se and Sp should be evaluated and match the purpose of the surveillance or identification scheme. When using register based predictors, it should be decided whether the variable actually only is used as a predictor or whether it is assessed as a problematic condition itself. In case of mortality, this would mean that the herd specific mortality could either be used to predict high lameness in herd based on optimised cut-offs. On the other hand, if predefined cut-offs are assigned to the herd specific mortality, it could be assessed as a problematic condition itself and not as a predictor. The question, whether the given predictor is a risk factor or hazard for the outcome of interest or just poses another problem on its own should therefore be answered first.

Traditionally, risk-based surveillance or targeted surveillance focus on risk-factors for given diseases leading to a more focused sampling of “high-risk populations” (Salman, 2003) within the given qualitative risk factors. This conversion of quantitative measures leads to a general loss of information and to an unwanted loss in test sensitivity, but is essential in order to develop the first step in the risk-based surveillance scheme i.e. the identification of hazards and to stratify the population into subgroups. As the risk-based surveillance schemes act like initial screening tests a high Se is needed (Stärk *et al.*, 2006), at least from the investigators (authorities) point of view. The subjects (herds/farm managers) being investigated would benefit from a highly specific model – avoiding false incrimination of herds with truly low lameness prevalence, although a subsequent control visit would elicit any doubts.

Currently, the Danish authorities only use the public accessible register data (i.e. culling, movement, abattoir and antimicrobial consumption data) for the risk-based identification of dairy herds for welfare control visits based on national means or norms as alarming thresholds. Within the past years the risk parameters have been: mortality, herd size, production system (beef/dairy), or previous sanctions (Anonymous, 2010). In order to identify herds at risk of having a given state, i.e. acceptable/unacceptable welfare or high/low lameness prevalence on a given day, certain cut-offs are established, which is rather challenging by the means of secondary incidence data. The choice of time period for our incidence data was restricted to cover the year 2004. This caused differences in time periods before and after herd visits between herds. Having a more systematic collection of register data with a fixed period before and after herd visits would probably have improved the predictive performance of our register data predictors. If the accuracy of the current surveillance system should be improved, optimal cut-offs should be used for the risk-based sampling. However, in order to enhance the accuracy of the predictive values the general prevalence of impaired welfare should be investigated in a large scale cross-sectional study. Furthermore, optimised cut-offs could also result in a higher number farms at risk challenging the implementation of such a surveillance due to operational constraints such as limited personnel and time.

Screening for multi-factorial syndromes assessed in cross-sectional studies, like welfare or just lameness, should regard more aspects of a dairy cow's life like health, productivity and management in order to cover as many associations as possible. The Danish risk-based welfare surveillance scheme only regards the five percent worst livestock herds *per annuum* based on a very limited set of predictors. The findings of the present study are in alignment with the current identification scheme, but emphasize the challenge of choosing the right threshold as the optimal cut-off for mortality was markedly lower than even the national means. Furthermore, the combinations of predictors only yielded a significant difference from the random ROC curve when put in combination with mortality. Adding more information to the predictor mortality did not result in any significant increase in AUC making this predictor the most capable predictor for lameness. The association between mortality and lameness under Danish settings was expected, since locomotor disorders are the primary reason behind 40% of all cases of euthanasia in Danish dairy cows (Thomsen *et al*., 2004).

The present study used a very small sample size of 40 herds yielding large confidence intervals for the estimates and making our ROC curves jagged, but still predictors showed reasonable predictive performance. However, the wide confidence intervals imply a larger error margin, which should be sought decreased by increasing the sample size.

Register-based predictors have been investigated in other surveillance settings, e.g. when evaluated as naïve Bayesian classifiers to give updated probabilities of a given outcome of interest, e.g. of emerging diseases in animal populations (Elbers *et al.*, 2002; Shephard *et al*., 2006; Geenen *et al*., 2011 But before considering distinct models based on no-gold standard methods like latent class analysis or Bayesian methods further investigations should be done in evaluating the effects of different time periods on the predictor performance.

Conclusion

The present study has shown that the quantitative assessment of indirect incidence register data can be used as an identification tool for direct cross-sectional measures, but it is essential to evaluate these predictors and predictive models as diagnostic tests for the given case definition in order to determine their predictive performance.

References

Alban, L. 1996. Lameness in tied Danish dairy cattle: the possible influence of housing systems, management, milk yield, and prior incidents of lameness. Preventive Veterinary Medicine 29: 135-149.

Anonymous. 2010.Kontrol af dyrevelfærd 2008-2009. Ministeriet for Fødevarer, Landbrug & Fiskeri, Fødevarestyrelsen. Mørkhøj, Copenhagen.

Brouwer, H., Bartels, C.J.M., van Wuijckhuise, L., and van Schaik, G. 2012. Syndromic surveillance on blood samples of aborting cows for early detection of (emerging) infectious diseases using different statistical methods. Proceedings of the Society for Veterinary Epidemiology and Preventive Medicine, March 28-30 2012, Glasgow, United Kingdom, pp. 233-244.

Carstensen, B., Plummer, M., Laara, E. and Hills, M. 2013. Epi: A Package for Statistical Analysis in Epidemiology. R package version 1.1.49. URL http://CRAN.R-project.org/package=Epi

de Vries, M., Bokkers, E.A., Dijkstra, T., van Schaik, G., and de Boer, I.J. 2011. Invited review: associations between variables of routine herd data and dairy cattle welfare indicators. Journal of Dairy Science 94: 3213-3228.

Duffield, T.F., Kelton, D.F., Leslie, K.E., Lissemore, K.D. and Lumsden, J.H. 1997. Use of test day milk fat and milk protein to detect subclinical ketosis in dairy cattle in Ontario. Canadian Veterinary Journal 38: 713-718.

Elbers, A.R., Bouma, A. and Stegeman, J.A. 2002. Quantitative assessment of clinical signs for the detection of classical swine fever outbreaks during an epidemic. Veterinary Microbiology 84: 323-332.

Geenen, P.L., van der Gaag. L.C., Loeffen, W.L.A. and Elbers, A.R.W. 2011. Constructing naïve Bayesian classifiers for veterinary medicine: A case study in the clinical diagnosis of classical swine fever. Research in Veterinary Science 91: 64-70.

Green, L.E., Hedges, V.J., Schukken, Y.H., Blowey, R.W. and Packington, A.J. 2002. The impact of clinical lameness on the milk yield of dairy cows. Journal of Dairy Science 85(9): 2250-2256.

Hanley, J.A. and McNeil, B.J. 1982. The Meaning and Use of the Area Under A Receiver Operating Characteristic (Roc) Curve. Radiology 143: 29-36.

Hanley, J.A. and McNeil, B.J. 1983. A Method of Comparing the Areas Under Receiver Operating Characteristic Curves Derived from the Same Cases. Radiology 148: 839-843.

Heuer, C., Schukken, Y.H. and Dobbelbar, P. 1999. Postpartum body conditio score and results from the first test day milk as predictors of disease, fertility, yield, and culling in commercial dairy herds. Journal of Dairy Science 82: 295-304.

Perrin, J.P., Ducrot, C., Vinard, J.L., Morignat, E., Calavas, D. and Hendrikx, P. 2012. Assessment of the utility of routinely collected cattle census and disposal data for syndromic surveillance. Preventive Veterinary Medicine 105:244-252.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Salman, M.D., Stärk, K.D.C. and Zepeda, C. 2003. Quality assurance applied to animal disease surveillance systems. Revue scientifique et technique OIE, 22,689-696.

Shephard, R.W., Toribio, J.A., Thomsen, A.R., Thomson, P. and Baldock, F.C. 2006. Incorporating the Bovine Syndromic Surveillance System (BOSSS) within an animal health surveillance network. Proceedings of the 11th Symposium of the International Society for Veterinary Epidemiology and Economics, Cairns, Aug. 2006, available online: http://www.sciquest.org.nz/elibrary/edition/5053

Smith, J.W., Ely, L.O. and Chapa, A.M. 2000. Effect of region, herd size, and milk production on reasons cows leave the herd. Journal of Dairy Science 83: 2980–2987.

Stärk, K.D.C, Regula, G., Hernandez, J., Knopf, L., Fuchs, K., Morris, R. and Davies, P. 2006. Concepts for risk-based surveillance in the field of veterinary medicine and veterinary public health: Review of current approaches. BMC Health Services Research 6: 20.

Sprecher, D.J., Hostetler, D.E. and Kaneene, J.B. 1997. A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance. Theriogenology 47: 1179-1187.

Thomsen, P.T., Østergård, S., Sørensen, J.T. and Houe, H. 2007. Loser cows in Danish dairy herds: Definition, prevalence and consequences. Preventive Veterinary Medicine 79: 116-135.

Thomsen, P.T. and Houe, H. 2006. Dairy cow mortality. A review. Veterinary Quarterly 28: 122-129.

Thomsen, P.T., Kjeldsen, A.M., Sørensen, J.T. and Houe, H. 2004. Mortality (including euthanasia) among Danish dairy cows (1990-2001). Preventive Veterinary Medicine 62: 19-33.

Paper 2 – Animal-based indicators

Evaluation of the performance of register data as predictors for dairy herds with high lameness prevalence

Nina Dam Otten[1], Nils Toft[1], Peter T. Thomsen[2], Hans Houe[1]

[1] Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Groennegaardsvej 8, DK-1870 Frederiksberg C, Denmark

[2] Department of Animal Science, Aarhus University, P.O. Box 50, DK-8830 Tjele, Denmark

This paper has been published in Preventive Veterinary Medicine, 2013, Volume 112, pages 348-354. DOI information: 10.1016/j.prevetmed.2013.09.002

Adjusting for multiple clinical observers in an unbalanced study design using latent class models of true within-herd lameness prevalence in Danish dairy herds

Nina Dam Otten[1], Nils Toft [1], Peter T. Thomsen [2], Hans Houe [1]

[1] *Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Groennegaardsvej 8, DK-1870 Frederiksberg C, Denmark*

[2] *Department of Animal Science, Aarhus University, P. O. Box 50, DK-8830 Tjele, Denmark*

Published in *Preventive Veterinary Medicine 112(3-4): 348-354*

Abstract

The elimination of misclassification bias introduced by multiple observers was evaluated and discussed based on an illustrative example using lameness prevalence in 80 Danish dairy herds. Data from 5,073 cows from loose-housed cubicle herds larger than 100 cows were included in the analysis. Four trained observers performed clinical scoring on cow level and undertook a calibration test with 39 video sequences. The calibration test served both the purpose of estimating inter-observer agreement (=0.69) in accordance with previous results and to estimate the sensitivity (Se) and specificity (Sp) for each observer. In the absence of a gold standard for the clinical observations, a latent class analysis (LCA) evaluating the true within-herd lameness prevalence was used. Sensitivity amongst observers was fairly low (0.24-0.81) inducing a general underestimation of the true prevalence. Comparative analyses were made to assess the effect of grazing on the lameness prevalence in order to demonstrate the consequences of using unadjusted apparent prevalences (AP) compared to the true prevalences (TP). Lameness prevalence was higher in grazing herds using AP estimates (19.0% zero-grazing, 20.2% grazing); while the TP estimates showed the expected higher lameness prevalence in zero-grazing herds (42.3% vs. 35.9%). Hence, this study emphasizes the importance of adjusting for observer Se and Sp to obtain true prevalence and avoid false interpretation.

Introduction

On-farm assessment of health and welfare on animal level requires the use of clinical and behavioural observations on individual animals. It is well known that such data are prone to misclassification and that differences between individual observers occur (Baadsgaard & Jørgensen, 2003). To alleviate the observer effects, training and calibration is seen as an essential part of studies involving multiple observers. Evaluation of the training and calibration of observers before, during and after study completion can be assessed as inter-observer agreement (IOA). Several studies rely only on kappa or prevalence adjusted bias adjusted kappa (PABAK) values as measures of agreement validating the given clinical condition measured. As an example, the clinical measures and assessment schemes used within the global welfare assessment protocol Welfare Quality® (WQ) were partly selected based on their validity in terms of inter-observer agreement. Within the overall WQ assessment of the clinical measures, the IOA for e.g. lameness was evaluated at four successive training sessions, improving the mean PABAK values from 0.6 to 0.7 over time (Brenninkmeyer *et al.,* 2007), which were considered as sufficient levels of agreement. Kappa and PABAK hold one major disadvantage: they only yield information about the agreement and cannot tell whether disagreement between observers is systematic. Hence, they do not eliminate the problem of observer subjectivity. Another challenge concerning the reliability of the included animal-based measures is that no real consensus on the limits for discriminating between acceptable and unacceptable agreement exists, although several limits are proposed in literature (Knierim & Winckler, 2009).

Consider as an example a study aimed at investigating the effect of grazing on the occurrence of lameness at herd level. This study could be performed using a number of herds with and without grazing in which a sample of cows were scored according to their lameness status (lame/not lame). In order for such a study to be conducted in a reasonably short time span, multiple observers each visiting separate herds are needed. Hence, data from such a study will consist of observed (or apparent) prevalences for a number of grazing and non-grazing herds. The problem with apparent prevalences is that they do not easily compare across populations as they represent the joint effect of observer bias and the true underlying lameness prevalence, thus resulting in potentially biased and misleading results and conclusions. In particular if the observers are not randomly allocated to the grazing and non-grazing herds, confounding bias is likely to occur. However, the true lameness prevalence may be derived from the observed lameness prevalence given that information about the sensitivity (Se, probability that a truly lame cow is classified as lame) and specificity (Sp, probability that a truly not lame cow is classified as not lame) is available for each observer. Unfortunately, getting reliable estimates of properties of the tests involved, i.e. in our case, estimates of Se and Sp for the individual observers of lameness remains a challenge. While literature and pilot-studies might be relevant for Se and Sp of e.g., serological, bacteriological or histopathological tests, the notion that a diagnostic test/mechanism ideally must be evaluated in the population where it is intended to be used, holds probably even stronger for clinical or observational data (Greiner & Gardner, 2000). Furthermore, the true underlying condition for most welfare related issues is generally unobservable. This requires models which do not rely on a perfect test for comparison. Latent class models (Hui & Walter, 1980) provide a tool for estimating Se and Sp of diagnostic tests in the absence of a perfect reference test given certain assumptions about the tests and the test subjects. Hence, models which allow for adjustment of the essentially unknown misclassification caused by different observers should be utilized to obtain reliable and unbiased results of general clinical data regarding health and welfare of animals.

Lameness is an animal welfare problem and almost 40% of all Danish dairy cows are lame to some degree (Jørgensen *et al.,* 2010). Furthermore, Thomsen et al. (2004) reported locomotor disorders being the reason for euthanasia in 40% of the euthanized dairy cows in Denmark. Numerous risk factor studies have been performed in the past with different lameness scoring systems being used for prevalence estimation, but without stating observer Se and Sp.

The overall objective of this study was to provide and discuss a framework for an unbalanced design with multiple observers using latent class models to estimate the true prevalence, illustrated and motivated by the example of the effect of grazing on lameness at herd level.

Materials and methods

Study design and model

To meet the challenges outlined above, the framework that we propose must have two distinct features. Firstly, comparison between groups must be done using the true prevalence in order to adjust for the misclassification imposed by the use of different observers. Secondly, data must allow inference about the Se and Sp of the observers. The procedures and dataflow involved in the modelling are summarized in Fig.1.

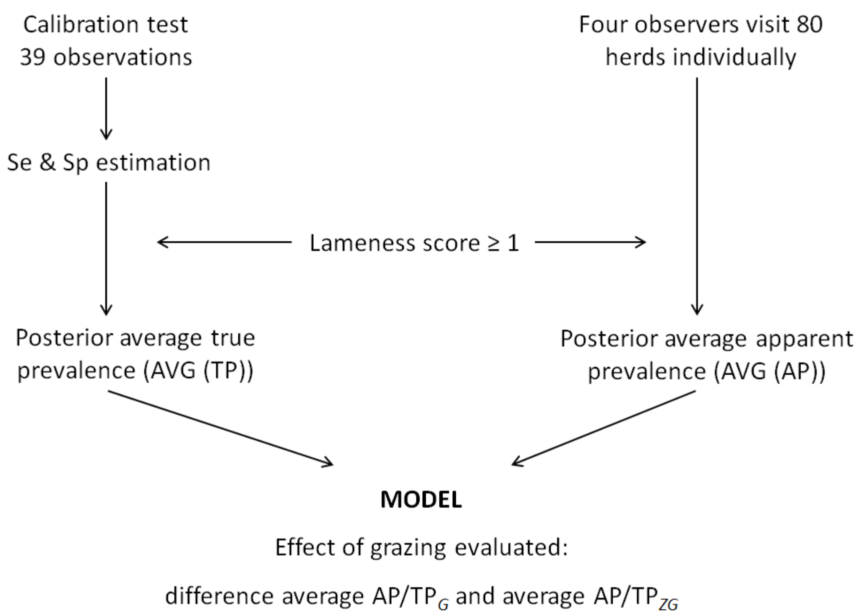


Figure 1. Flow-diagram of the statistical models and data.

Inference about Se and Sp of the observers can be achieved in several ways. The predominant problem is the lack of a perfect reference observer or so called gold standard. As stated above, this problem can be circumvented by applying latent class models (Hui & Walter, 1980). The general assumptions underlying these models are sometimes referred to as the Hui-Walter paradigm: I) the data must be from two or more populations with differing prevalence; II) the Se and Sp of the two or more individual observers must be constant across these populations; III) the observers are considered to be conditionally independent given the underlying disease/welfare condition. However, in the original paper it is noted that also 3 or more conditionally independent observers and just one population will yield sufficient information (i.e. degrees of freedom) to allow estimation of the required parameters. In our example, we will make use of this setup, but we will address more general designs in the discussion.

For the example that will be elaborated further in the next section, data from a reliability study were available to calculate Se and Sp of the 4 observers, using the following Bayesian model:

$$O_{ij} \sim \text{Bernoulli}(OP_{ij}),\ i=1,2,3,4,\ j=1,\ldots,39$$

$$OP_{ij} = Se_i \times TC_j + (1-Sp_i) \times (1-TC_j),\ i=1,2,3,4,\ j=1,\ldots,39$$

$$TC_j \sim \text{Bernoulli}(P),\ j=1,..,39$$

$$Se_i \sim \text{Beta}(1,1),\ i=1,2,3,4$$

$$Sp_i \sim \text{Beta}(1,1),\ i=1,2,3,4$$

$$P \sim \text{Beta}(1,1)$$

Where O_{ij} is the observation of the *i*th observer on the *j*th cow (in the reliability study), this observation follows a Bernoulli distribution, i.e. is either 0 (not lame) or 1 (lame); OP_{ij} is probability of observing a 1; TC_j is the true condition of the *j*th cow, which is following a Bernoulli distribution, with probability P; Se_i and Sp_i are the Se and Sp of the *i*th observer, respectively. As we do not have any useful prior information on the Se, Sp and P, they are assumed to follow Beta(1,1) distributions, which are essentially uniform on the interval [0;1]. To explore the effect of the choice prior distributions for Se and Sp on the posterior inference, we ran the analysis using Jeffreys prior (Beta(0.5,0.5)) as an alternative non-informative prior for Se and Sp.

Adjusting for misclassification of a dichotomous outcome in a risk factor analysis, can also be achieved in several ways. However, the underlying mechanism is essentially the same: observed data are modelled using apparent prevalence and the apparent and true prevalence are linked through the Se and Sp of the observer. The true prevalence is then modelled in a manner that allows for comparison between groups.

The risk factor model was also formulated as a Bayesian model:

$L_f \sim \text{Binomial}(AP_f, N_f)$, f=1,…80

$AP_f = Se_{fi} \times TP_f + (1-Sp_{fi}) \times (1-TP_f)$, f=1,…,80, and fi = 1,2,3 or 4

$TP_f \sim \text{Beta}(1,1)$, f=1,…80

Where L_f was the observed number of lame cows (out of N_f) on the *f*th farm; AP_f was the apparent prevalence of lameness on the farm; TP_f the true prevalence modeled as a Beta(1,1), since no prior relevant information was available; Se_{fi} and Sp_{fi} are the Se and Sp of the observer visiting the *f*th farm. To evaluate the potential effect of grazing, the average AP (AVG(AP)) and the average TP (AVG(TP)) were calculated for the 34 zero-grazing and 46 grazing farms and the hypotheses: H_0: $AVG(AP_{grazing}) > AVG(AP_{zero\text{-}grazing})$ and H_0: $AVG(TP_{grazing}) < AVG(TP_{zero\text{-}grazing})$ were evaluated using posterior probabilities (POPR) obtained using the step-function in OpenBUGS. Initially a model allowing for herd random effects of true prevalence was used, but due to convergence issues caused by the small sample size in the reliability study it was replaced by the current model (using independent Beta(1,1) distributions) instead.

To summarize the model framework: we have combined a latent class model for diagnostic test evaluation with a Bayesian regression model with misclassification of the outcome. The benefit of an integrated model rather than a two-step approach where the latent class posterior estimates serve as priors for the regression model lies in the additional information about observer Se and Sp which can be inferred from the risk factor data. The true prevalence in a herd must be above 0 and below 1; this affects the Se and Sp of the observers through the bounds imposed by the TP-AP relationships in the individual herds.

The model was implemented in OpenBUGS (Thomas *et al.*, 2006) which uses a Markov Chain Monte Carlo (MCMC) sampling algorithm to obtain samples from the posterior distribution. The model was initially run for 10,000 iterations to ensure convergence of the model before obtaining 100,000 samples for posterior inference. Convergence diagnostics were assessed by checking history plots, Gelman-Rubin plots (using three chains with different starting values) and autocorrelation was reduced by thinning to every 10^{th} sample as proposed by Toft et al. (2007). Posterior inferences for AP, TP, Se and Sp were assessed by posterior means and credibility intervals (95% PCI). The POPR was assessed as the proportion of iterations for which the hypotheses were true using the step-function in OpenBUGS. The code is given in Appendix 1.

Model application

The model described above was applied to a study evaluating the effect of grazing on lameness. The data provide potential sources of bias because the data were collected with an unbalanced distribution of observers between the two risk factor groups (Fig.2a).

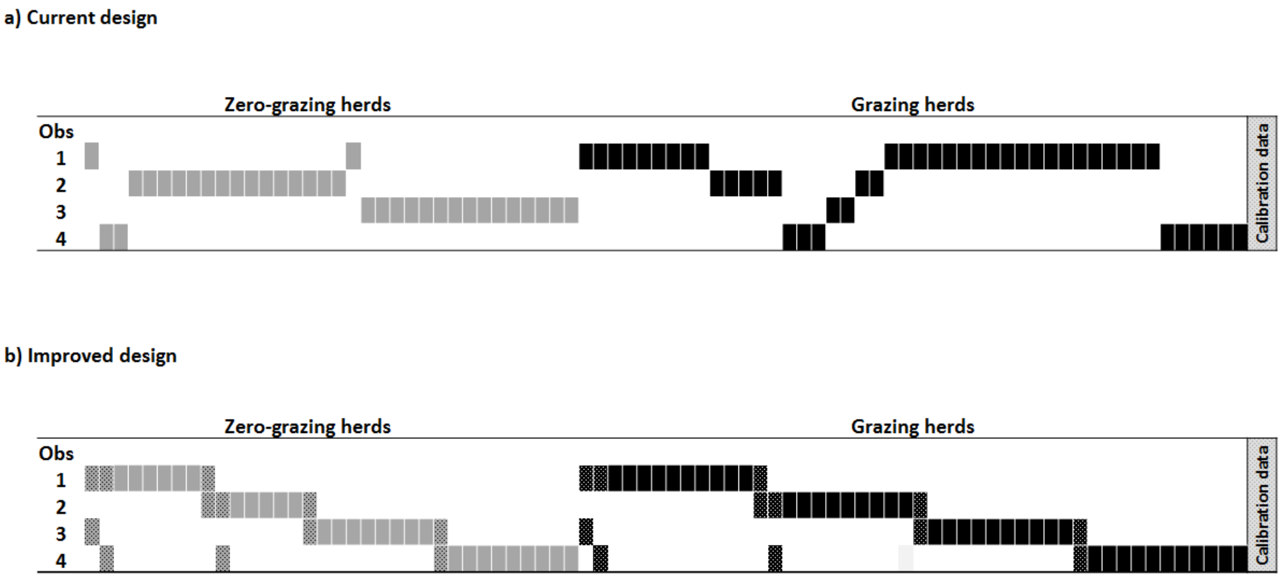


Figure 2. Illustration of study designs. Fig. 2a) depicts the current design with four observers (Obs) only having observations within the calibration material in common, since they visited herds individually. Fig. 2b) shows the improved design where observers have both the observations from the calibration material and an equal number of herds from each population (G=grazing herds and ZG=zero-grazing herds) in common with other observers.

Study population and sampling

The data used in the study originated from two independent studies; both using random samples of Danish dairy herds with more than 100 cows and loose-housing with cubicles. Study one was designed for investigating the effects of grazing on animal welfare. Study two investigated models to predict animal welfare in dairy herds. The first study only sampled amongst grazing herds, while the second study sampled herds regardless of their grazing status. Within each herd, a minimum of 50 animals were examined. For further information on the study population and within-herd sample size calculations see Burow et al. (2013). The amalgamated study population for this study consisted of 5,073 cows from 80 herds, 46 grazing and 34 non-grazing herds (Table 1).

Table 1. Summary descriptive characteristics of the 80 herds visited. The distribution of herds visited stratified by observer with the percentage herds visited within the given risk factor population (grazing and zero-grazing) in brackets.

	N herds	Grazing	
		Yes (%)	No (%)
Total	80	46 (57.5)	34 (42.5)
Observer			
1	30	28 (61)	2 (6)
2	22	7 (15)	15 (44)
3	17	2 (4)	15 (44)
4	11	9 (20)	2 (6)

Clinical protocol

Each herd was visited once by one of four trained observers in order to make an on-farm animal welfare assessment. The welfare assessment protocol consisted of seven clinical measures including lameness. The examined cows were lameness scored (for grazing herds during their housing period) walking in a straight line on a hard and level surface for a minimum of 10 meters. Groups of cows on deep litter (e.g. dry cows) were not assessed.

Lameness was scored on an ordinal scale: 0 = normal gait, 1 = moderately lame = impaired stride and/or rhythm with reduced weight bearing on one limb and 2 = severely lame = no weight bearing or more than one limb affected by lameness. For our study, a cow was regarded as either lame with a score of 1 or 2 or non-lame (score 0).

Observer calibration and inter-observer agreement

The four observers were all experienced in lameness scoring of cattle. They were trained in the clinical protocol on three on-farm occasions as well as by evaluating video sequences prior to data collection. Upon completion of the data collection, a reliability check was performed based on scoring of 39 video sequences in order to assess the inter-observer agreement. Inter-observer agreement for the lameness scoring protocol was assessed using the prevalence-adjusted bias-adjusted kappa (PABAK) (Byrt *et al*., 1993) taking the categories of the given measure into account: PABAK= $[((k*p)-1) / (k-1)]$, where k is the number of categories and p the proportions of matches. In the present study lameness was assessed at two levels: lame or non-lame. To check for systematic differences between observers tetrachoric correlation coefficients (r*) were assessed using the R-package polycor() (Fox, 2010).

Results

The average posterior apparent prevalence was higher in grazing herds, but only in 10% of the samples (Table 2). However, the average posterior true prevalence was higher in zero-grazing herds in 95.2% of the samples (Table 2).

Table 2. The effects of the risk factor grazing on the mean posterior apparent (AP) and true (TP) within-herd lameness prevalence in 80 Danish dairy herds. (POPR = posterior probability of the estimates being different, i.e. equivalent to a frequentist p-value).

Prevalence	All herds	Grazing vs.	Zero-grazing	POPR
TP (%)	39.2			
		35.9	42.3	0.048
AP (%)	19.6			
		20.2	19.0	0.100

Inter-observer agreement assessed by PABAK showed a “good” mean agreement of 0.69 according to the acceptability levels defined by Landis & Koch (1977) and similar to the initial agreement levels of 0.59 after the first training session for the lame/non lame scale evaluated within the WQ (Brenninkmeyer *et al.,* 2007). Pairwise comparisons showed a systematic difference between the two observers 1 and 4 versus observers 2 and 3, since observers 1 and 4 generally assigned higher score to cows (i.e. lameness score 2) than observers 2 and 3. The r* showed highest correlation between observers 2 and 4, second highest between observers 2 and 3 and third highest between 1 and 4, which corresponds well with the PABAK values. Furthermore, observer 1 consequently assigned higher scores compared to the other observers followed by observer 2 as shown in Table 3. The correlations could be modelled, but are not relevant for the prevalence estimation in this case, since they do not deal with the problem of misclassification.

Table 3. Prevalence-adjusted bias-adjusted kappa values, tetrachoric correlation coefficients (r*), asymptotic standard error (ASE) for r* and number of observations where a given rater assigned a higher score than the other in a reliability study with 39 observations.

Rater *x*	Rater *y*	PABAK	r*	ASE	N mismatches in scores
					$Rater_x > Rater_y$ / $Rater_y > Rater_x$
1	2	0.692	0.907	0.078	9/2
	3	0.590	0.863	0.104	12/1
	4	0.795	0.940	0.060	9/3
2	3	0.795	0.956	0.044	6/2
	4	0.692	0.999	0.000	5/8
3	4	0.487	0.797	0.139	3/11

The estimated Se and Sp for observers from the full model (calibration and herd visit data) showed a general higher Sp (0.85-0.95) than Se (0.24-0.81). Observers 1 and 4, and observers 3 and 4 showed best concordance with the smallest differences in regards to Se. In terms of Sp large differences were found between the same observers. Plotting AP against TP in Fig. 3 shows the effects that observer specific Se and Sp exert on the TP revealing a general underestimation of lameness prevalence. Using Jeffreys priors for the observers changed the posterior median estimates of Se and Sp slightly, but the overall pattern remained. The conclusions regarding apparent and true prevalence did not change.

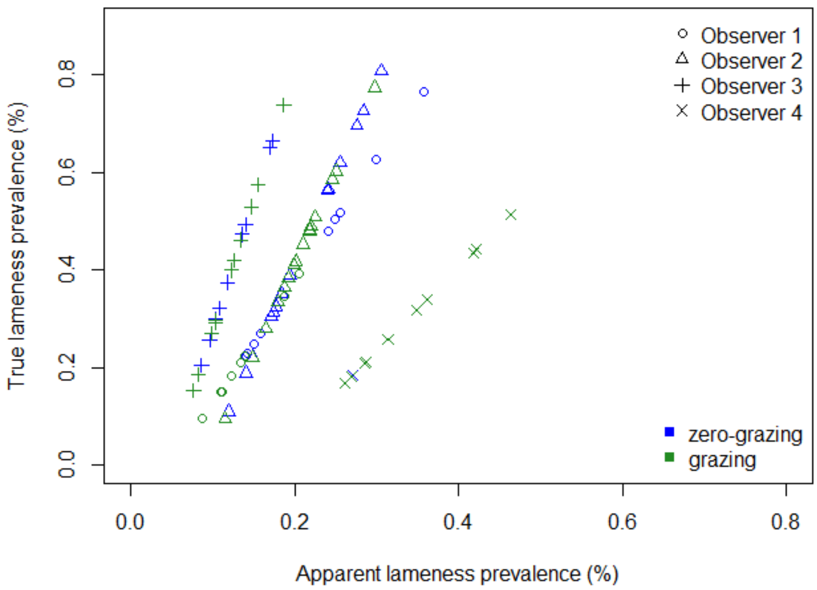


Figure 3. Relationship between the apparent lameness prevalence (AP) and the true lameness prevalence (TP) for all four observers based on 80 herds.

Discussion

This study proposed a modelling approach to alleviate the potential bias in multi-observer studies by making cross-population comparisons through the use of true, rather than apparent prevalences. This was exemplified in our study by grazing being a protective factor when the true prevalences were used, while the apparent prevalences suggested a potential negative effect of grazing. The posterior prevalence estimates yielded a posterior average true lameness prevalence of 39.2% compared to the AP of 19.6%. Due to a generally low Se and high Sp among observers, the apparent prevalence was an underestimation of the true lameness prevalence. This was even more pronounced in zero-grazing herds, as the two most frequent observers in this group showed lowest Se (observer 3 with 0.24 and observer 2 with 0.36). This efficiently masked the effect of grazing as a protective factor for lameness.

The general lack of sensitivity of clinical scorings has been addressed previously by Baadsgaard and Jørgensen (2003) who also stressed the importance of evaluating the uncertainty of clinical observations. Furthermore, as most clinical states are complex in their aetiology and visual manifestations, the "true" state must be regarded as a latent variable, which can be dealt with in a latent class setting. By applying a latent class approach, we have provided a solution to deal with the misclassification of the clinical scoring as well as differences between observers. The Bayesian regression model that we have adopted can account for the imperfect Se and Sp of the given tests when estimating prevalence or investigating risk factors. Bayesian regression with misclassification has previously been applied in studies involving intra mammary infections (Dufour *et al.,* 2012; Koop *et al.,* 2012), and gastro-intestinal infections in cattle (Cardona *et al*., 2011; Weber *et al.,* 2009; Nielsen & Toft, 2007).

The posterior TP estimates in the present study were essentially driven by the 39 observations from the calibration test. Still, by combining the calibration and risk factor data in one model, posterior inference about individual observer Se and Sp was improved, so that estimates both reflect the properties of the calibration data as well as obey the rules regarding true prevalence being above 0 and below 1 in each herd. Despite the added information from the combined data, inference based on such a small sample is associated with a high degree of uncertainty. To get better posterior estimates more information is needed from the data. By using a cross-over design as in Fig. 2b the collected data could also have been used for calibration purposes. Here observers examine (blinded) the same cows in at least one herd for each pairs of observers; ideally a herd from each stratum in the primary (hypothesized) risk factor. This implies that all pairs of observers are evaluated with sufficient data to allow the estimation of Se and Sp of all observers. The cost is that for X observers, an additional X*(X-1) visits are needed, in our case 4*3=12 (Fig. 2b. pattern filled boxes). However, designs with less overhead are possible.

The assumption of conditional independence between observers given lameness status of the cow can be challenged. A possible improvement of the current model would have been the use of a conditional dependence structure for the observers. Dependency structures between more than three observers have been modelled using different approaches (Espeland & Handelman, 1988, 1988; Yang and Becker, 1997; Qu *et al*., 1998) and with varying conclusions. Albert and Dodd (2004) found it impossible to distinguish between dependence structures when more than four tests/observers were applied. For the current study, a pairwise dependence between some of the observers could perhaps have been justified due to differences in training and background. However, it was not deemed possible to get inference about this considering the small sample size of the reliability study.

It is possible to formulate more complicated/flexible models for the true prevalence, i.e. models that also allows for more covariates. Since the study was intended as proof of concept, we choose a simplistic model regarding evaluation of the differences between the two populations – grazing and zero-grazing. Observer agreement for the given protocol was good, but still the apparent lameness prevalence showed a general underestimation compared to the true prevalence estimated in our model. Additionally, the biologically plausible associations between the risk factor grazing and lameness were masked by the observer effect. This emphasizes the dilemma of clinical observations, where subjectivity still persists even though observers have undergone extensive training and calibration. The present study shows the lack of reliability of the lameness protocol, much in contrast to the results from the Welfare Quality® Group. Despite a good inter-observer agreement, misclassification due to observer subjectivity in clinical scoring still had a considerable effect on the outcome and conclusions.

Comparisons between studies are difficult. Not only because of differences in lameness scoring systems and housing systems, but the present study also illustrates the pit-falls of relying on IOA values of the given lameness scoring system. Furthermore, with true lameness prevalences as high as 40%, the findings of the present study could suggest that many of the previously reported prevalences and effects of risk factors could be underestimated.

Conclusion

The use of latent class methods for adjustment of observer effects proved to be operational to avoid misinterpretation of prevalence estimates and associated risk factors. In summary, the findings of this study emphasized the problem of the reliability of lameness scoring protocols and the magnitude of the lameness problems in modern Danish dairy production, as the lack of sensitivity caused an underestimation of the lameness prevalence. Over one third of all cows in an average herd were truly lame. Grazing was associated with a positive effect on the true lameness prevalence but not on the apparent prevalence.

Acknowledgements

The authors gratefully acknowledge Elke Burow for her cooperation by collecting, editing and providing parts of the data, Tine Rousing for discussing the assessment protocol.

References

Albert P.S. and Dodd, L.E. 2004. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. Biometrics 60(2): 427-435.

Baadsgaard, N.P. and Jørgensen, E.. 2003. A Bayesian approach to the accuracy of clinical observations. Preventive Veterinary Medicine 59: 189-206.

Brenninkmeyer, C., Dippel, S., March, S., Brinkmann, J., Winckler, C. and Knierim, U. 2007. Reliability of a subjective lameness scoring system for dairy cows. Animal Welfare.16: 127-129.

Burow, E., Thomsen, P.T., Rousing, T. and Sørensen, J.T. 2013. Daily grazing time as a risk factor for alterations at the hock joint integument in dairy cows. Animal 7(1): 160-166.

Byrt, T., Bishop, J. and Carlin, J.B. 1993. Bias, prevalence and kappa. J Clin Epidemiol. 46(5): 423-429.

Cardona, G.A., Carabin, H., Goni, P., Arriola, L., Robinson, G., Fernandez-Crespo, J.C., Clavel, A., Chalmers, R.M. and Carmena, D. 2011. Identification and molecular characterization of Cryptosporidium and Giardia in children and cattle populations from the province of Alava, North of Spain. Science of The Total Environment 412: 101-108.

Dufour, S., Dohoo, I.R. and Barkema, H.W. 2012. Epidemiology of coagulase-negative staphylococci intramammary infection in dairy cattle and the effect of bacteriological culture misclassification. Journal of Dairy Science 95(6):3110-3124.

Fox, J. 2010. polycor: Polychoric and Polyserial Correlations. R package version 0.7-8. http://CRAN.R-project.org/package=polycor

Espeland M.A. and Handelman, S.L.. 1989. Using latent class models to characterize and assess relative-error in discrete measurements. Biometrics 45: 587–599.

Hui S.L. and Walter, S.D. 1980. Estimating the error rates of diagnostic tests. Biometrics. 36: 167-171.

Jørgensen, H.B.H., Pedersen, L.D., Sørensen, M.K, Thomsen, P.T. and Norberg, E.. 2010. Born to be a loser cow? Journal of Dairy Science 93: 4386-4390.

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

Koop, G., Toft, N., Gardner, I.A., van Werven, T. and Nielen, M. 2012. Bayesian logistic regression model of risk factors for intramammary infection in dairy goats. Proceedings of The Society for Veterinary Epidemiology and Preventive Medicine, Glasgow, 28-30 March 2012 (2012), pp: 13-25.

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

Nielsen, S.S. and Toft, N.. 2007. Assessment of management-related risk factors for paratuberculosis in Danish dairy herds using Bayesian mixture models. Preventive Veterinary Medicine 81(4): 306-317.

Qu YS, A. Hadgu. 1998. A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. Journal of the American Statistical Association 93(443): 920-928.

Thomas, A., O'Hara, B., Ligges, U., Sturtz, S. 2006. Making BUGS Open. R News, 12-17.

Thomsen, P.T., Kjeldsen, A.M, Sørensen, J.T. and Houe, H. 2004. Mortality (including euthanasia) among Danish dairy cows (1990-2001). Preventive Veterinary Medicine 62(1): 19-33.

Toft N., Innocent, G.T., Gettinby, G. and Reid, S.W.J. 2007. Assessing the convergence of Markov Chain Monte Carlo methods: An example from evaluation of diagnostic tests in absence of a gold standard. Preventive Veterinary Medicine 79: 244-256.

Toft, N., Jørgensen, E. and Højsgaard, S.. 2005. Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. Preventive Veterinary Medicine 68: 19-33.

Weber, M.F., Verhoeff, J., van Schaik, G. and van Maanen, C. 2009. Evaluation of Ziehl-Neelsen stained faecal smear and ELISA as tools for surveillance of clinical paratuberculosis in cattle in the Netherlands. Preventive Veterinary Medicine 92(3): 256-266.

Yang I. and Becker, M.P. 1997. Latent variable modeling of diagnostic accuracy. Biometrics 53: 948–58.

Appendix – BUGS code

Data for two populations pop0 (zero-grazing) and pop1 (grazing)

n.cows in calibration test = 39

test scores lame=1 and sound=0 for t1=observer1, t2= observer2, t3=observer3,

t4=observer4

r0/1= order of observers in the given population

n0/1= total number of animals scored in each herd

l= total number of lame cows for each herd in the given population

DATA

list(n.cows = 39, t1 = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0), t2 = c(1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1), t3 = c(0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0), t4 = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1,1, 1, 0, 1, 0, 1, 1, 1, 0), pop = c(1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1),n.pop0=34,

r0= c(2, 2, 4, 2, 3, 3, 1, 1, 2, 1, 1, 1, 1, 1, 3, 2, 1, 1, 1, 3, 1, 3, 2, 1, 1, 3, 3, 2, 3, 2, 2, 2, 3, 1, 1), n0 = c(56, 55, 70, 53, 74, 63, 62, 55, 59, 63, 57, 58, 63, 55,57, 60, 62, 61, 64, 62, 67, 60, 62, 77, 69, 68, 58, 65, 60, 66, 68, 72, 59, 67), l0 = c(5, 3, 18, 7, 18, 8 , 7, 10, 5, 8, 9, 9, 15, 16, 17, 8, 16, 4, 11, 9, 20, 14, 8, 26, 9, 17, 6, 9, 11, 5, 12, 27, 14, 11),n.pop1=46, r1= c(2, 2, 4, 4, 2, 2, 1, 2, 2, 2, 3, 1, 1, 3, 2, 1, 1, 2, 3, 3, 2, 2, 4, 4, 4, 2, 3, 4, 3, 4, 4, 3, 1, 1, 1, 3, 1, 1, 4, 4, 1, 1, 1, 1, 1, 1), n1 = c(63, 60, 77, 56, 70, 58, 61, 69, 61, 56, 56, 77, 57, 62, 55, 64, 58, 64, 56, 73, 53, 59, 62, 58, 60, 56, 57, 59, 63, 59, 66, 74, 88, 60, 70, 55, 53, 58, 64, 69, 78, 59, 76, 57, 53, 57), l1=c(9, 5, 19, 14, 10, 9, 9, 6, 13, 7, 5, 13, 10, 11, 6, 8, 3, 5, 11, 7, 6, 3, 17, 27, 25, 2, 6, 21, 8, 18, 18, 9, 18, 15, 15, 3, 13, 19, 22, 29, 15, 13, 14, 12, 10, 12

)

)

INITS

list(se=c(0.9,0.9,0.9,0.9),sp=c(0.9,0.9,0.9,0.9),prev=c(0.4))

model{

for (i in 1:n.cows) {

Probablity the i'th cow is lame, depends on the prev of the population

p[i] ~ dbern(prev[pop[i]])

test results

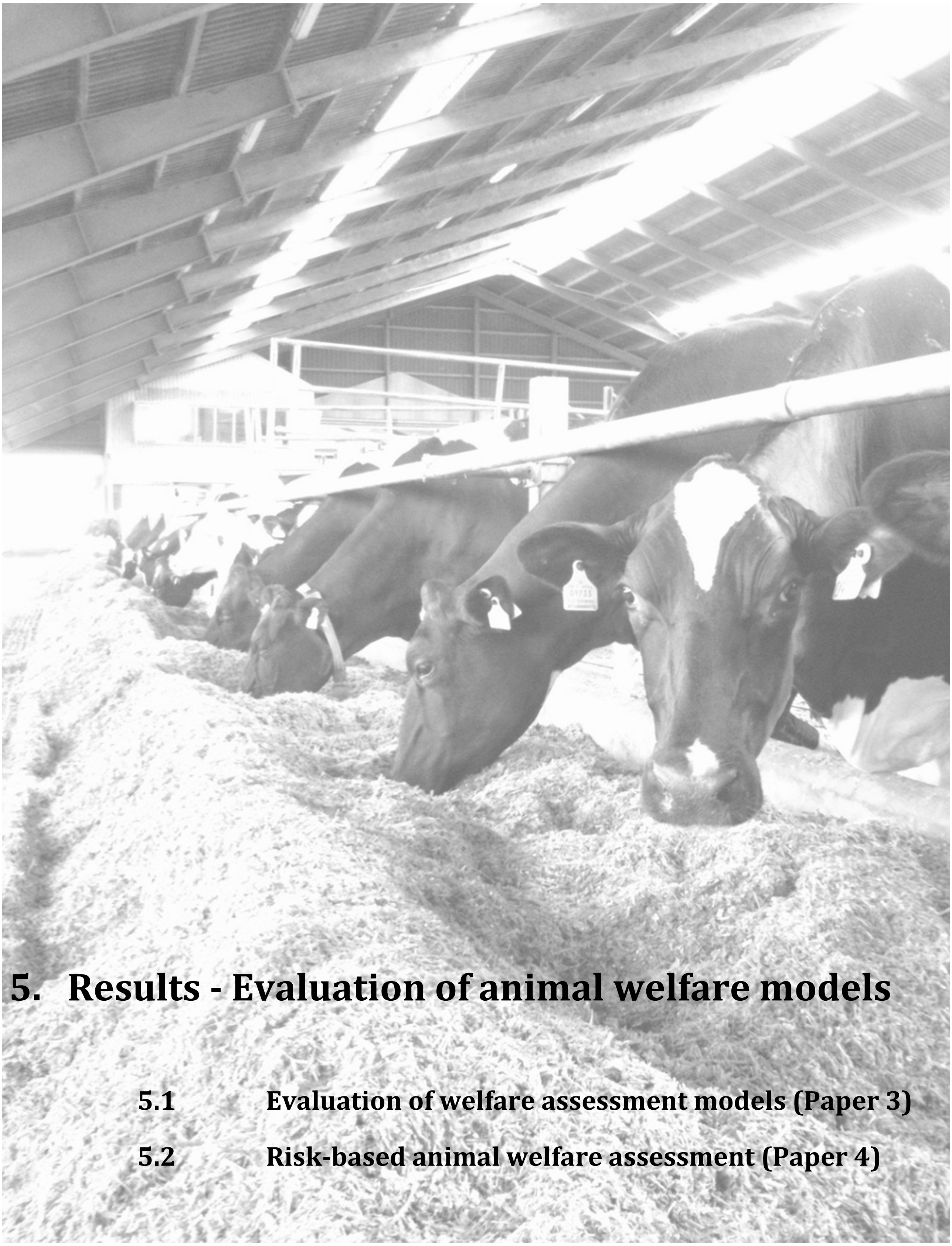
```
ap.t1[i] <- se[1]*p[i] + (1-sp[1])*(1-p[i])
ap.t2[i] <- se[2]*p[i] + (1-sp[2])*(1-p[i])
ap.t3[i] <- se[3]*p[i] + (1-sp[3])*(1-p[i])
ap.t4[i] <- se[4]*p[i] + (1-sp[4])*(1-p[i])

t1[i] ~ dbern(ap.t1[i])
t2[i] ~ dbern(ap.t2[i])
t3[i] ~ dbern(ap.t3[i])
t4[i] ~ dbern(ap.t4[i])

}
# Priors
for (t in 1:4) {
se[t] ~ dbeta(1,1)
sp[t] ~ dbeta(1,1)
}
# prev as beta
for (s in 1:1) {
prev[s] ~ dbeta(1,1)
}
#estimating TP based on AP
for (i in 1:n.pop0){
l0[i] ~ dbin(ap0[i],n0[i])
ap0[i] <- se[r0[i]]*tp0[i] + (1-sp[r0[i]])*(1-tp0[i])
#beta distribution non-informative prior
tp0[i] ~dbeta(1,1)
}
mu.tp0 <- mean(tp0[])
mu.ap0 <- mean(ap0[])
```

```
for (i in 1:n.pop1){
l1[i] ~ dbin(ap1[i],n1[i])
ap1[i] <- se[r1[i]]*tp1[i] + (1-sp[r1[i]])*(1-tp1[i])
#beta distribution non-informative prior
tp1[i] ~dbeta(1,1)
}
#testing for differneces in AP and TP
mu.tp1 <- mean(tp1[])
tp.effect <- step(mu.tp1-mu.tp0)
mu.ap1 <- mean(ap1[])
ap.effect <- step(mu.ap1-mu.ap0)


}
```



5. Results - Evaluation of animal welfare models

Paper 3 – Welfare assessment models

Quantification of animal welfare in dairy herds using different sources of data

Nina Dam Otten[1], Tine Rousing [2], Peter T. Thomsen[2], Hans Houe[1], Jan Tind Sørensen[2]

[1] Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Groennegaardsvej 8, DK-1870 Frederiksberg C, Denmark

[2] Department of Animal Science, Aarhus University, P.O. Box 50, DK-8830 Tjele, Denmark

The manuscript has been submitted in the present form to Animal Welfare on September 16th 2013 and will be revised according to reviewer comments before resubmission.

Quantification of animal welfare in dairy herds using different sources of data

Nina Dam Otten[1], Tine Rousing[2], Peter T. Thomsen[2], Hans Houe[1], Jan Tind Sørensen[2]

[1] Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Groennegaardsvej 8, DK-1870 Frederiksberg C, Denmark

[2] Department of Animal Science, Aarhus University, P. O. Box 50, DK-8830 Tjele, Denmark

Submitted to *Animal Welfare*

Abstract

The present study investigated the potential of using feasible register data to assess animal welfare compared to the more costly on-farm welfare assessments. An animal Welfare Index (AWI) was created for the three different information sources: register-based measures (AWI 1), resource-based measures (AWI 2) and animal based measures (AWI 3) using expert opinion to weight measures. AWI 1 measures were drawn from the Danish Cattle Database for three different time-periods, 365 days (period a), 180 days (period b) and 90 days (period c) prior to the on-farm assessments. Additionally, a combined index between the resource- and animal-based measures was created (AWI2 + 1) for each time period. Spearman Rank correlation was assessed between all indices. Significant, but weak negative correlations were found between AWI 3 and AWI1 for period b and positive correlations with AWI 2+1b; significant positives correlations were found between AWI2 and AWI 1 for period a. These contradictory findings highlight the challenges in defining the right time period for register data and hence, further investigations are needed before these feasible register data can be used to predict animal welfare at herd level.

Introduction

The assessment and quantification of animal welfare has been of major concern over the past decades and from this interest several assessment protocols have emerged, such as the *Tiergerechtheitsindex* (TGI) in Austria (Bartussek, 1999) and the RSPCA Freedom Food farm assurance scheme in UK (Main *et al.,* 2001). Since the aims of these assessments often differ between either serving for certification or as an advisory tool (Johnsen *et al.,* 2001), the methods, the underlying welfare definition and the included measures will also vary among protocols depending on these overall aims. Resource based measurements (i.e. the assessment of the nearby environment of the animals) are widely used as welfare measures in quality assurance programmes (Webster *et al.,* 2004). Resource based measures should be related to animal based measures because the former can be seen as causing animal welfare problems measured by direct outcome measures called animal based measures. However, studies indicate that animal welfare measured by animal based measures may vary within the same or similar housing systems and overall management regimes that are alike (Whay *et al.,* 2003; Rousing *et al.,* 2007).

Therefore, the focus has been on developing animal welfare assessment protocols based on direct observations or tests of the animals, e.g. cleanliness or blood sample results - (hereafter termed primary animal measures), (e.g. Winckler *et al.,* 2003; Webster *et al.,* 2004; Keeling 2009). An extensive international cooperation stated the consensus on modern animal welfare assessment by the development of the Welfare Quality® (WQ) assessment protocol in 2009 (Welfare Quality® 2009). A major drawback for implementation of the WQ protocol is costs due to the major time-consumption used for data collection. Furthermore, in order to validate the objective nature of the included measures, welfare assessors need to perform ongoing calibration, all adding to the total cost of these animal based assessment protocols. The estimated time-consumption for a full WQ assessment is 7-8 hours for a 200 head dairy herd (Welfare Quality® 2009). Milk production, reproduction and health recordings have been widely investigated for their association with a variety of animal based measures. In a recent review, de Vries et al. (2011) advocate the possible use of these fairly cheap and readily accessible indicators as an alternative welfare assessment. This approach using data collected for other purposes than welfare assessment for assessing animal welfare has been tried (Sandgren *et al.,* 2009; Dewey *et al.,* 2009; Nyman *et al.,* 2011, Kelly *et al.,* 2011).

However, the knowledge on how well an animal welfare assessment based on indirect animal based measures (hereafter termed secondary animal based measures) – that is routine registrations in central databases – correlate to an animal welfare assessment based on primary animal based measures is scarce. The objective of the present study was to evaluate the correlation between three overall welfare indices. The three welfare indices were defined from the following three different sources of information, the two low-cost sources including register-based indicators from routine registrations in central databases as secondary animal based measures and the resource based measures and the high-cost primary animal based measures. A secondary objective was to evaluate register-based indicators for three different time periods in order to identify the most appropriate time period in terms of correlation with the assumed true welfare state given by the animal based measures. Finally, key indicators among register-based indicators were identified by their correlation with welfare index for animal based measures.

Materials and methods

Study herds

The study herds were drawn from a list of 401 responders to a survey on grazing strategies carried out by the Danish Cattle Federation and Aarhus University in 2009. The target population was a Danish dairy herd with more than 100 cows while the following additional inclusion criteria were set for the study herds in the present study: more than 100 cows and loose-housing systems with cubicles. Out of the 401 respondents, two random samples were drawn for two separate studies and hereafter amalgamated for the present study. The first sample of 41 herds came from a study on the effect of grazing on the overall welfare (Burow *et al.,* 2013). From the same pool of initial respondents a new sample of 90 herds was drawn of which 45 were willing to participate. Finally, 86 herds were included in the present study.

Herd visits and observers

Each herd was visited once by one of four trained observers during a period from period April 2010 to July 2011. Due to non-compliance with the inclusion criteria upon data collection at the visit, four herds were excluded. Further nine herds were excluded due to missing or incomplete data sheets. Finally, 73 herds remained in the study for further analysis of which 40 herds used summer grazing. Within herd sampling was done as described by Burow et al. (2013), with a minimum of 50 cows per herd.

Each herd visit started approximately one hour after morning feeding. Observers were trained by two on-farm sessions as well as one video and picture session before conducting the visits. Upon completion of data collection, a second video and picture session was performed to assess inter-observer agreement. For each clinical and behavioural measure, a prevalence-adjusted bias adjusted kappa (PABAK; Byrt *et al.,* 1993) was calculated. PABAK values for inter-observer agreement ranged from 0.25-0.83.

Measures

Secondary animal based measures (register-based indicators)

A list of 28 register-based indicators was made reflecting the different aspects of a dairy cow's lifecycle (reproduction, milk production, treatments, mortality and abattoir remarks) and representing the aspects of productivity, health and management related to the given categories of welfare (Table 1). Data were extracted from the Danish Cattle Database for three different time periods prior to the farm visit date of the given herd: a) 365 days prior and b) 180 days prior and c) 90 days prior to the date of the visit in the given herd. All indicators were extracted as within herd level prevalences or means for the given period.

In order to be used in an animal welfare index (AWI), all indicator values were transformed into categorical values based on the distribution of indicator percentiles among the sample. A score 0 was given for indicator values among the 25% best herds within the given indicators, score 1 for indicator values> 25th and < 75th - percentiles, score 2 for indicator values ≥ 75th and < 90th percentile and score 3 for indicator values among the worst 10% in the sample. Since the spread of abattoir remarks (total of nine) between farms was very low, only remarks on lung disorders, liver abscesses, peritonitis, cirrhosis and liver flukes, old fractures and chronic inflammation were kept in the model, leaving 24 indicators for further analysis.

Resource-based measures

At each herd visit, a total of 127 variables were recorded. These included qualitative scoring of e.g. cleanliness of water points, the slipperiness of the floor, light and sufficiency of bedding material and quantitative measures of e.g. cubicle dimensions, passage width, number of feeding slots etc. All variables were then combined into 16 main indicators (Table 1) reflecting the overall aspects of: feed and water provision, resting area, movement and space, sick pens and barn equipment. In this step, indicators were evaluated based on the current Danish recommendations concerning animal welfare and when applicable also on Danish animal welfare legislation (Danish act on keeping dairy cattle and their offspring). Herds were given a value ranging from zero to one, based on the compliance with the given recommendations for the given indicator. Compliance was given a score 0 while non-compliance resulted in a value of 1, whereas partial compliance was regarded as a fraction of the non-complying measures: *n* measures yielding non-compliance/ total possible. In other words, a herd not fulfilling the recommendations for one passage width out of a total of four passages would receive a value 1/4 = 0.25. Missing values were assigned a non-informative score of 0.5. The maximum AWI score for level 2 (AWI 2) was 52.

Primary animal based measures

A clinical assessment protocol with ten clinical and two behavioural measures was used to score a random sample of cows consisting of both lactating and dry cows in each herd. The clinical protocol was modified from the WQ protocol to fit Danish settings, excluding a number of measures due to very low prevalences. The remaining measures included: hygiene leg, hygiene hind quarter, hygiene udder, integument alteration on carpus, tarsus and body, claw conformation, body condition score (BCS), lameness and the avoidance distance (AD). Furthermore, the measures hair coat and rising behaviour were added to the protocol based on their implementation in previous evaluation methods (Thomsen *et al., 2007;* Rousing *et al.,* 2007). Eight of the clinical measures were assessed at graded levels 0-2 (absent, moderate and severe impairment), while claw conformation was assessed on a binary scale (normal or overgrown), BCS on an ordinal scale with quarterly intervals (1-5) and rising behaviour on an ordinal scale ranging from 1-5. The AD was assessed at the beginning of each herd visit and measured in centimetres (cm) in 10 cm intervals. Finally, the three latter measures were transformed into the graded scale of normal, moderate or severe impairment as follows: 1.75 >BCS ≤ 2.5 as 'lean' (moderate impairment) and ≤ 1.75 as 'thin' (severe impairment), rising behaviour was graded as 1-2: 'normal', 3: 'interrupted' (moderate) and 4-5: 'abnormal' (severe), 0.5 ≤ AD < 1 meter as 'sceptical' (moderate) and AD ≥ 1 meter as 'shy' (severe). All measures were used as mean within herd prevalences and weighed by their severity in the AWI.

Animal Welfare Index (AWI)

Establishing measures and weights

The on farm welfare assessment was based on Welfare Quality® (WQ) Assessment Protocols (www.welfarequality.net/everyone). The WQ incorporates measures referring to four main principles Good Feeding, Good Housing, Good Health and Appropriate Behaviour described by a total of 12 criteria. Measures are firstly aggregated to criteria and then to principle scores, from which the individual farm may be categorised as either 'not classified', 'acceptable', 'enhanced' or 'excellent'. In this present study, we chose a different aggregation approach skipping the criteria and principle scores and where measures directly were aggregated to individual farm scores expressed as herd Animal Welfare Indexes (AWI). The individual farm welfare index directly reflected the 'proportion' of animals in the herd with remarks referring to the measures included (specified as prevalences as mentioned in the previous section) (Burow *et al.,* 2013). Like in the WQ, all measures were assigned to one of the four categories *feeding, housing, health and management,* with the latter replacing appropriate behaviour as a more descriptive term (Table 1). Furthermore, an expert panel was used to decide 1) measure weights describing the relative 'impact' of measures and 2) the relative impact of individual measure levels (answering the question: "what is the relative weight of severe vs. moderate level of graded measures as e.g. lameness?") - both as regards to measure contribution to the welfare index. The herd animal welfare index (AWI) was calculated based on the graded and non-graded welfare measures.

Expert panel opinions

Initially, 20 out of 32 appointed potential panel participants within the field of expertise of production advice (2), practicing veterinarians (5), researchers (3), industry affiliated experts (6), official welfare control officers (3) and animal-rights organisation representative (1) responded to an online questionnaire during December 2012 and January 2013. The experts were for each measure to give scores on a five point scale ranging from 1 (non-important) to 5 (very important). Relative weighing was ensured in as much as panel participants further had to make sure that the mean was 3 (= equally importance). The relative weight of severe vs. moderate level of graded measures was formulated as open questions – for what reason expert were not limited as no range was specified beforehand. The median weights of experts The graded (moderate/severe) and non-graded measures were first multiplied by the median weight assigned to respective measure and graded level by the experts and summed up to calculate the AWI.

Aggregation model for calculation of an Animal Welfare Index (AWI)

A simple additive and weighed index was created for the register- (AWI 1) and resource-based (AWI 2) measures:

$$AWI = \sum_{j=1}^{k} N_j W_j$$

where N denotes the *jth* measure and W the *jth* median weight assigned by the experts.

For the animal-based measures (AWI 3), an extended model was used, taking the graded measures into account as well using the below formula:

$$AWI = \sum_{i=1}^{k} (M_i MW_i + S_i) W_i + \sum_{j=1}^{k} N_j W_j$$

(Modified from Burow *et al.,* 2013)

where M, S and N were the herd's adjusted prevalence of moderate measure levels, severe measure levels and non-level graded measure, respectively. MW and W were the expert panel medians of relative measure level weights and measure weights; i was the proportion of the individual level graded measure, j the proportion of the individual non-graded measure. For the graded measures k was 9 and for non-graded measures k=1. Hence, the theoretical maximum score for the AWI 1 based on secondary animal based measures 1 (AWI 1 a,b,c) was 196.5, given that all indicators were amongst the 10% worst herd values (score 3). The theoretical maximum scores for AWI 2 based on resource based measures (all measures with a score 1) was 52 and finally the maximum score was 3900 for AWI 3 based on primary animal based measures, given that all measures were at a 100% prevalence for the severe levels. Finally, the two low-cost indices (AWI 1 and AWI 2) were added to an overall index (AWI 2+1a/b/c) and assessed for correlation with the high-cost index AWI 3.

Data analysis

All data editing, AWI calculations and statistical analyses were made in SAS© 9.3 (SAS Institute Inc., SAS Campus Drive, Cary, North Carolina). Correlations between all five AWI´s (AWI 1a, WI 1b, AWI 1c, AWI 2, and AWI 3) and the three combined AWI´s (AWI 2+1a, AWI 2+1b, AWI 2+1c) and the AWI 3 were assessed by the non-parametric Spearman Rank Correlation Coefficient using the **proc corr spearman** procedure. Additionally, correlations between the AWI´s for the three different time periods were assessed. In order to identify key indicators among the register-based indicators, significant correlations between individual register-based indicators and the AWI 3 for animal-based measures were also investigated.

Results

The median weights of all measures are given in Table 1 and the level weights for the graded measures (moderate vs. severe) are given in Table 2. The biggest in-between expert variation was found for the measures calf mortality and proportion of abattoir remarks for the secondary animal measures; the scraping system (exposed vs. covered) for resource-based measures; and lameness and rising behaviour for the primary animal based measures.

Table 1. Median weights identified by expert opinion for welfare indicators and measures from three different information sources (register data from routine registrations, resource-based measures and animal-based measures) used for the calculation of an Animal Welfare Index (AWI) for dairy cattle.

Information level	Indicator measure	Category	Weight
Level 1 Register-based	Lean cows at slaughter[a]	Feeding	3
	Abattoir remarks liver cirrhosis		2
	Bulk tank somatic cell count	Health	3
	Veterinary treatments per 100 cow years[b]		3
	Proportion of locomotor disorders/100 cow years		3
	Proportion of abattoir remarks		3
	Abattoir remarks:		
	Lung disorders		3
	Liver abscesses		3
	Peritonitis		2.5
	Liver flukes		2
	Chronic inflammation		2
	Cow mortality	Management	5
	Heifer mortality		4
	Calf mortality		4
	Annual average milk yield per cow year[b]		4
	Milk yield per lactation group (1st, 2nd or ≥ 3rd)		4
	Standard deviation of milk yield per lactation group)		4
	Age at first calving		3
	Standard deviation of age at first calving		3
	Abattoir remarks old fractures		2

Table 1 continued

Information level	Indicator measure	Category	Weight
Level 2 Resource-based	Water supply	Feeding	4
	Water cleanliness		3
	Number of feeding slots		4
	Occupancy rate bed stalls	Housing	4
	Bed stall length		3
	Bed stall width		3
	Passage ways:		
	Width		3
	Skid resistance		3
	Flooring		3
	Dead ends		3
	Calving pen size	Health	4
	Separation of animals		4
	Sick animals not in sick bay		3
	Harmful/damaged equipment		4
	Brushes		3
	Scraping system		3
Level 3 Animal-based	Body condition score	Feeding	4
	Hygiene:	Housing	2
	Leg		
	Hind		2
	Udder		2
	Rising behaviour		3
	Integument alterations	Health	4
	Carpus		4
	Tarsus		4
	Body		2.75
	Hair coat		4.5
	Lameness		
	Claw conformation	Management	3.75
	Avoidance distance		3

[a] Cows with a fata score 1 according to the EU Beef Carcase Classification
[b] Sum of feeding days of all cows per herd/365 days

Table 2. Weights assigned by expert opinion to the two levels of moderate or severe impairment of 11 graded welfare measures for an Animal Welfare Index for dairy cattle based on clinical and behavioural measures (modified after Burow *et al* 2013).

Measure	Weight ratio (moderate vs. severe)
Lameness	0.33 vs. 1
Integument alterations (carpus, tarsus, body)	0.33 vs. 1
Body condition	0.33 vs. 1
Rising behaviour	0.33 vs. 1
Hair coat	0.50 vs. 1
Hygiene (hind, udder)	0.50 vs. 1
Avoidance distance	0.50 vs. 1
Hygiene leg	0.67 vs. 1

Descriptive results show relative little spread of AWI scores among herds across all information sources (Table 3).

Table 3. Descriptive summary statistics for Animal Welfare Index (AWI) scores for dairy cattle based on different information sources (AWI 1= register-based measures, AWI 2= resource-based measures, AWI 3 = animal-based measures) and for different time periods prior to data collection on-farm for resource- and animal-based measures (AWI 1a = 365 days, AWI 1b = 180 days and AWI 1c = 90 days), N=73 herds.

Variable	Mean	Median	Standard Deviation	Minimum	Q1	Maximum	Q3	Possible range
AW1 a	70.03	69.00	16.69	41.00	57.00	125.00	78.00	0-196.5
AWI 1b	69.93	69.00	14.34	41.00	61.00	119.00	76.00	0-196.5
AWI 1c	73.51	71.00	12.73	49.00	65.00	108.00	81.00	0-196.5
AWI 2	33.23	33.50	5.16	18.00	30.50	43.50	36.24	0-52
AWI 3	944.65	937.85	262.17	449.99	741.19	1757	1096.02	0-3900

Correlation of the ranking of herd AWI's based on three different sources of information

Significant correlations were found between the AWI 3 and the AWI 1 for data from the period of 180 days prior to visit (period b) and between the AWI 2 and the AWI 1 for data from the period of 365 days prior to visit (period a). All Spearman Rank Correlation Coefficients are shown in Table 4. Similar, but lower correlations were found for the combined AWI from both the register- and resource-based index for time period b and the AWI 3 (Table 5). All significant correlations are depicted in Figure 1.

Table 4. The Spearman Rank Correlation Coefficients (*ρ*) between the Animal Welfare Index (AWI) scores for three different information sources (AWI 1 register-based, AWI 2 resource-based and AWI 3 animal-based) for dairy cattle. Significant correlations are highlighted by an asterix (P*< 0.05, ****P*<0.0001).**

	Register-based			**Resource-based**	**Animal-based**
	Period a	**Period b**	**Period c**		
	AWI 1a	AWI 1b	AWI 1c	AWI 2	AWI 3
AWI 1a	-	0.648***	0.639***	0.231*	-0.089
AWI 1b		-	0.630***	0.004	-0.291*
AWI 1c			-	0.166	0.069
AWI 2				-	0.093
AWI 3					-

Table 5. The Spearman Rank Correlation Coefficients (*ρ*) between the Animal Welfare Index (AWI) scores for three combined indices for the three different time periods of the register- and overall resource-based indices against the animal-based index scores for dairy cattle. Significant correlations are highlighted by an asterix (P*< 0.05, ****P*<0.0001).**

	Combined index			**Animal-based**
	Period a	**Period b**	**Period c**	
	AWI 2+1a	AWI 2+1b	AWI 2+1c	AWI 3
AWI 2+1a	-	0.215***	0.217***	-0.005
AWI 2+1b		-	0.197***	0.015*
AWI 2+1c			-	0.005
AWI 3				-

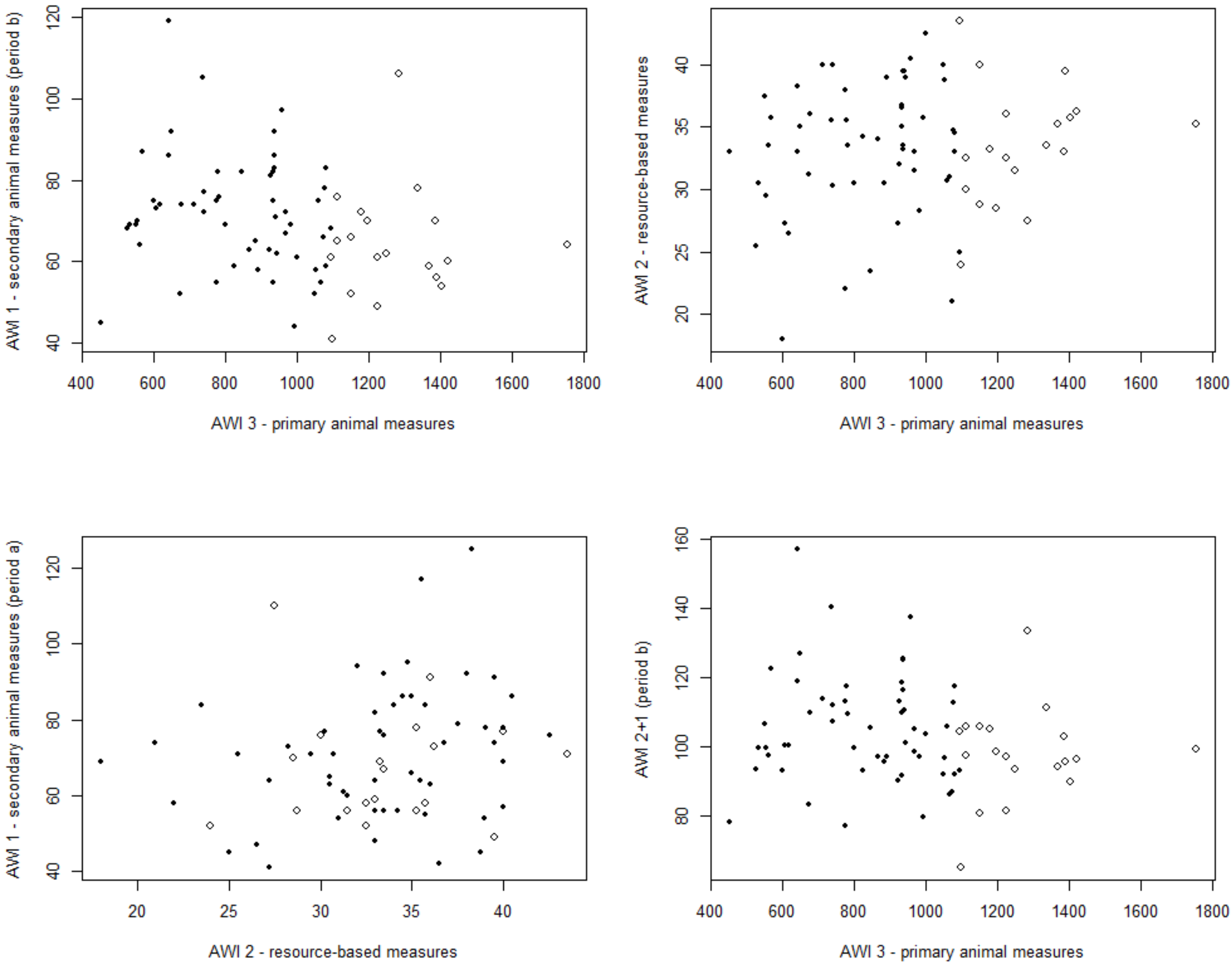


Fig.1. Plots depicting the significant correlations between four animal welfare indices (AWI) based on different sources of information (AWI 1 register data for period a (360 days prior to herd visit), period b (180 days prior to herd visit), and period c (90 days prior to herd visit), AWI 2 resource-based measures, and AWI 3 animal-based measures and the combined index for AWI 2 and AWI 1. Herds with an AWI 3 score above the third quartile are considered as potential problem herds and are marked ◇ and acceptable welfare herds are marked •.

Correlation between key indicators and AWI based on primary animal based measures

Only a certain group of register-based indicators were significantly correlated ($P< 0.05$) with AWI 3 following a consistent pattern for the three different time periods. All seven abattoir indicators (proportion of abattoir remarks, remarks on lung lesions, liver abscesses, peritonitis, liver flukes and cirrhosis, old fractures and chronic inflammation) for period a showed significant correlation with the AWI 3 based on primary animal based measures. Abattoir remarks were accompanied by the mean bulk tank somatic cell count for period b, but with a smaller number of abattoir indicators, namely the proportion of abattoir remarks, abattoir remarks for liver abscesses and chronic inflammation. Significant correlations between indicators in period c were found for the proportion of abattoir remarks and the mean bulk tank somatic cell count, furthermore a tendency towards significant correlation was seen for the calf mortality (P= 0.06).

Discussion

The aim of the study was to investigate the correlation between three welfare indices based on different information sources. This was done by applying the same model framework for the three Animal Welfare Indices (AWI´s) using an additive weighted model, where weights were derived by expert opinion. For the register-based AWI three different time periods for indicator creation were assessed. Significant negative correlation was found between the register-based AWI (AWI 1) for the period of 180 days prior to the on-farm evaluation and the animal-based AWI (AWI 3). Despite the relative small variations in AWI 1 scores between periods, no significant correlations were found for the other two time periods for AWI 1 and the AWI 3 or between the resource-based AWI (AWI 2) and the AWI 3. However, the AWI 2 showed a significant positive correlation with the AWI 1 for the time period of 365 days prior to on-farm evaluation. Combining the information from AWI 1 and 2 did not improve correlation to the AWI 3. In general, all correlations to the AWI 3 were weak (ρ< 0.39).

The findings of the current study highlight the challenges in finding alternative and cheaper welfare assessment methods, than the laborious but yet more direct animal based assessments. In contrast to the WQ protocol, which presents an integrated approach covering different data sources like routine registrations, resource measurements as well as direct animal observations, the present study investigated each data source individually. Additionally, although measures were inspired by the WQ protocol measures, there was a certain need to accommodate the assessment protocol to the study purpose and Danish production settings. Due to operational issues some animal based measures were excluded or replaced based on literature reviews. All register-based measures included in the WQ were excluded, since they were conflicting with the research question whether an index based on these measures would correlate with an index based solely on animal based measures. In the same manner the resource-based measures were also excluded and treated individually in the present study. The resource-based measures were extended to cover 16 indicators compared to the seven indicators in the WQ. Danish legislation requires the use of analgesics when dehorning and tail-docking and furthermore, tail-docking is only allowed upon medical indication, not as a preventive management procedure. Therefore, the measures reflecting the criteria *absence of pain by management induced procedures* were not applicable under Danish production settings. Hence, not all 12 criteria were covered by the measures included in the present AWI 1, 2 and 3. The register-based measures did only cover a part of all criteria, as only criteria *absence of prolonged hunger, absence of injuries, and absence of disease* were represented. In contrast, the animal based measures covered the broadest range by the criteria: *absence of pro-*

longed thirst, absence of prolonged hunger, comfort around resting, absence of injuries, absence of disease and good human-animal relationship. This discrepancy definitely contributes to the lack of correlation. A better coverage of all criteria was found in combining the register-based measures with the resource-based measures. However, it would still be difficult to directly reflect all of the WQ criteria associated with *Appropriate behaviour*, which makes direct comparisons with the integrated WQ score impossible. Results also showed no improvement in correlation, when these two indices were combined. Other studies have investigated the potential of register-based data to predict on-farm animal welfare (Sandgren *et al.,* 2009; Nyman *et al.,* 2011; de Vries, 2013). However, none of these studies used an overall welfare index as they chose a different approach in defining the animal welfare levels by using percentiles to categorise welfare measures into acceptable or unacceptable. Hence, so far investigated associations between register data and welfare indicators has been limited to single indicators – not to an overall welfare index score like the present study.

The lack of correlation between the resource-based and the primary animal based measures indicate that management has a major impact on dairy cattle welfare. This is illustrated by the discrepancies between scores, as cows in a “bad” system can have acceptable welfare scores (based on animal based measures) and vice versa. This also becomes visible in the negative correlation between the secondary animal measure AWI 1a & b and the primary animal measure AWI 3. A distinct feature in the current study was the moderate spread of both AWI scores and the single measures included in the AWI´s across herds. This could be caused by the somewhat biased sampling of herds among positive respondents in the initial survey. The inclusion criteria for the herds were set to be as representative of the Danish dairy cow population as possible. But it is very likely, that this has influenced the moderate spread, as the sample did not contain any extreme differences in resource-based measures, as identical production systems were evaluated. However, the AWI 2 did cover the widest range by covering 49% of the possible AWI spectrum compared to the 34% coverage by the AWI 3. Looking for differences in similar herds is a very difficult task and the moderate spread between AWI´s in herds is most likely to be held responsible for the lack of correlation as also documented by Andreasen et al .(2013).

Major challenges lie in defining, selecting and aggregating welfare indicators based on the existing data sources that cover all aspects of animal welfare of biological functioning, affective state and natural living (Fraser *et al.,* 1997) and which at the same time are well correlated with the actual welfare status of the given herd. This resulted in the Welfare Quality® development of a comprehensive and multi-dimensional welfare index. A comprehensive index based on a state of the art assessment protocol. However the lack of transparency became clear in regards to the aggregation of measures into an overall score. Much effort was put into the modelling part, much of this based on expert opinion. As reported by Bonde et al. (2009), it was found that experts never followed linear curves when asked to score virtual datasets and expert answers were therefore modelled by non-linear utility functions, based on which experts’ measure weightings were indirectly defined. But, as seen retrospect WQ researchers concluded that although underlying calculations had been explained in detail these appeared very complex (for example in Veissier *et al.,* (2011), and Welfare Quality® (2009 a,b,c)) and therefore were only accessible for a very narrow target audience. Hence, in order to obtain transparency, the current study chose a simple linear model also based on expert opinion using direct and indirect estimation techniques for scaling as proposed by Scott et al. (2001). However, the interpretation of AWI scores based on sums can hold potential hazards, as it does not regard compensation and trade-off between measures or categories (Botreau *et al.,* 2007) as welfare advantages are not included as negative or deductive variables. However, Botreau et al (2007) argued that the use of such sums could be beneficial for assessments of welfare subsets, an approach also used by the Swedish Dairy Association (Sandgren *et al.,* 2009) using seven different focus areas throughout the dairy cow life cycle to assess the on-farm welfare.

Finally, the validity of the chosen measures and the AWI model should be mentioned. The primary animal based measures were all selected based on previous validation within the WQ work and in other welfare assessment schemes. The model framework was validated in a study by Burow et al. (2013) comparing the AWI scores in herds in the winter housing versus summer grazing period.

Conclusion and animal welfare implications

In conclusion, the secondary animal based measures did show weak correlation with both the primary animal based and the resource-based measures. The current AWI´s were calculated for the three different measure types in contrast to the integrated WQ score. Further investigations are needed not only in order to find the most predictive combination of measures and refining the current protocols but also to determine which time period the secondary animal-based measures should be based on. In this study three different time periods were used for extracting register data and further research determining the most predictive time period is needed. Dairy herds are very volatile entities with different impact of actions and their consequences; hence a rolling average and aberrations from this might have better perspectives in terms of predicting animal welfare than using means from fixed time periods. However, the use of cheaper and feasible welfare measures, i.e. secondary animal based and resource-based, should only be used for screening purposes, while the final welfare assessment should consist of primary animal based measures complemented by additional measures (register and resource based) to ensure coverage of all welfare aspects.

References

Andreasen, S., Wemelsfelder, F., Sandøe, P. and Forkman, B. 2013. The correlation of Qualitative Behavior Assessments with Welfare Quality® protocol outcomes in on-farm welfare assessment of dairy cattle. Animal Welfare 143(1): 9-17.

Bartussek, H. 1999. A review of the animal needs index (ANI) for the assessment of animals' well-being in the housing systems for Austrian proprietary products and legislation. Livestock Production Science 61: 179-192.

Bonde, M., Boutreau, R., Bracke, M.B.M., Butterworth, A., Capdeville. J., De Rosa, G., Engel, B., Keeling, L.J., Perny, P., Rousing, T., Sørensen, J.T., Van Reenen, K. and Veissier, I. 2009. Final model of multicriterion evaluation of animal welfare. Deliverable 2.28, subtask 2.3.1.3, EU Food-CT-2004-506508.

Botreau, R., Veissier, .I, Butterworth, A., Bracke, M.B.M. and Keeling, L.J. 2007. Definition of criteria for overall assessment of animal welfare. Animal Welfare 16: 225–228.

Burow, E., Rousing, T., Thomsen, P.T., Otten, N.D. and Sørensen, J.T. 2013. Effect of grazing on the cow welfare of dairy herds evaluated by a multidimensional welfare index. Animal 7(5): 834-842.

Byrt, T., Bishop, J. and Carlin, J.B. 1993. Bias, prevalence and kappa. Journal of Clinical Epidemiology 46: 423–429.

Dewey, C., Haley, C., Widowski, T., Friendship, R., Sunstrum, J. and Richardson, K. 2009. Using data collected for production or economic purposes to research production animal welfare: An Epidemiological Approach. Journal of Applied Animal Welfare Science 12(2): 105-113.

DeVries M, Bokkers EA, Dijkstra T, van Schaik G, and de Boer IJ 2011 Invited review: associations between variables of routine herd data and dairy cattle welfare indicators. Journal of Dairy Science 94: 3213-3228.

deVries, M. 2013. Assuring dairy cattle welfare: towards efficient assessment and improvement. PhD thesis. Wageningen Univeristy dissertation no. 5486.

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

Johnsen, P.F., Johannesson, T. and Sandøe, P. 2001. Assessment of farm animal welfare at herd level: many goals, many methods. Acta Agriculturae Scandinavica Section A- Animal Science S30: 26-33.

Keeling, L. (Ed.). 2009. An Overview of the Development of the Welfare Quality Project© Assessment Systems. Welfare Quality Reports no.11. Cardiff University, United Kingdom.

Kelly, P.C., Moore, S.J., Blake, M. and Hanlon, A.J. 2011. Identification of key performance indicators for on-farm animal welfare incidents: possible tools for early warning and prevention. Irish Veterinary Journal 64: 13.

Main, D.C.J., Webster, A.J.F. and Green, L.E. 2001. Animal Welfare Assessment in Farm Assurance Schemes, Acta Agriculturae Scandinavica, Section A – Animal Science, 51 (S30): 108-113.

Nyman, A.K., Lindberg, A. and Sandgren, C.H. 2011. Can pre-collected register data be used to identify dairy herds with good cattle welfare? Acta Agriculturae Scandinavica, Section A – Animal Science, 53 (S1): S8.

Rousing, T., Jakobsen, I.A., Hindhede, J., Klaas, I.C., Bonde, M. and Sørensen, J.T. 2007. Evaluation of a welfare indicator protocol for assessing animal welfare in AMS herds: Researcher, production advisor and veterinary practitioner opinion. Animal Welfare 16: 213-216.

Sandgren, C.H., Lindberg, A. and Keeling, L.J. 2009. Using a national dairy database to identify herds with poor welfare. Animal Welfare 18: 523-532.

Scott, E.M., Nolan, A.M. and Fitzpatrick, J.L. 2001. Conceptual and methodological issues related to welfare assessment: A framework for measurement. Acta Agriculturae Scandinavica Section A- Animal Science. Suppl 30:5-10.

Thomsen, P.T., Østergård, S., Sørensen, J.T. and Houe, H. 2007. Loser cows in Danish dairy herds: Definition, prevalence and consequences. Preventive Veterinary Medicine 79, 116-135.

Webster, A.J.F., Main, D.C.J. and Whay, H.R. 2004. Welfare assessment: indices from clinical observation. Animal Welfare S93-98.

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

Welfare Quality® .2009 b. Welfare quality assessment protocol for pigs (sows and piglets, growing and finishing pigs). Welfare Quality® Consortium, Lelystad, Netherlands.

Welfare Quality®. 2009 c. Welfare quality assessment protocol for poultry. Welfare Quality® Consortium, Lelystad, Netherlands.

Whay, H.R., Main, D.C.J., Green, L.E. and Webster, A.J.F. 2003. Assessment of the welfare of dairy cattle using animal-based measurements: Direct observations and investigation of farm records. The Veterinary Record 153: 197–202.

Winckler, C., Capdeville, .J, Gebresenbet, G., Hörning, B., Roiha, U., Tosi, M. and Waiblinger, S. 2003. Selection of parameters for on-farm welfare assessment protocols in cattle and buffalo. Animal Welfare 12: 619-624.

Veissier, I., Jensen, K.K., Botreau, R. and Sandøe, P. 2011. Highlighting ethical decisions underlying the scoring of animal welfare in the Welfare Quality® scheme. Animal Welfare 20(1, S1): 89-101.

Paper 4 – Risk-based animal welfare assessment

Register-based predictors of violations of animal welfare legislation in Danish dairy herds

Nina Dam Otten[1], Liza Rosenbaum Nielsen[1], Peter T. Thomsen[2], Hans Houe[1]

[1] Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Groennegaardsvej 8, DK-1870 Frederiksberg C, Denmark

[2] Department of Animal Science, Aarhus University, P.O. Box 50, DK-8830 Tjele, Denmark

The manuscript has been submitted in the present form to Animal on December 2nd 2013.

Register-based predictors of violations of animal welfare legislation in Danish dairy herds

Nina Dam Otten[1], Liza Rosenbaum Nielsen[1], Peter T. Thomsen[2], Hans Houe[1]

[1] *Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Groennegaardsvej 8, DK-1870 Frederiksberg C, Denmark*

[2] *Department of Animal Science, Aarhus University, P.O. Box 50, DK-8830 Tjele, Denmark*

Submitted to *Animal*

Abstract

The assessment of animal welfare can include resource-based or animal-based measures. Official animal welfare inspections in Denmark primarily control compliance with animal welfare legislation based on resource measures (e.g. housing system) and usually do not regard animal response parameters (e.g. clinical and behavioural observations). Herds selected for welfare inspections are sampled by a risk-based strategy based on existing register data. The aim of the present study was to evaluate register data variables as predictors of dairy herds with violations of the animal welfare legislation (VoAWL) defined as occurrence of at least one of the two most frequently violated measures found at recent inspections in Denmark, namely a) presence of injured animals not separated from the rest of the group and/or b) animals in a condition warranting euthanasia still being present in the herd. A total of 25 variables were extracted from the Danish Cattle Database and assessed as predictors using a multivariable logistic analysis of a dataset including 73 Danish dairy herds, which all had more than 100 cows and cubicle loose-housing systems. Univariable screening was used to identify variables associated with VoAWL at a p-value < 0.2 for the inclusion in a multivariable logistic regression analysis. Backwards selection procedures identified the following variables for the final model predictive of VoAWL: increasing standard deviation of milk yield for second lactation cows, high bulk tank somatic cell count (≥ 250,000 cells/mL) and suspiciously low number of recorded veterinary treatments (≤ 25 treatments per 100 cow-years) The identified predictors may be explained by underlying management factors leading to impaired animal welfare

in the herd, such as poor hygiene, feeding and management of dry or calving cows and sick animals. However, further investigations are required for causal inferences to be established.

Introduction

Traditionally, animal welfare assessment protocols have sought to ensure the basic animal needs as described in the "Five Freedoms" by the Farm Animal Welfare Council (1992) being the freedom from hunger and thirst; freedom from discomfort and pain, injury and disease; freedom to express normal behaviour; and the freedom from fear and distress. In the legislative context these needs have been addressed by establishing minimum standards regarding housing equipment and management (e.g. daily care and intervention), aspects typically assessed by resource-based measures. Recently published guidelines on the risk assessment of animal welfare by the European Food Safety Authority (EFSA 2012), are based on the traditional approach of quantifying risk exposure scenarios and their welfare consequences on single adverse effects of animal welfare, e.g. lameness. However, the risk assessment regarding animal welfare as a complex entity is still in its early phase.

The Danish Veterinary and Food Administration (DVFA) has conducted risk-based sampling of livestock herds for the official animal welfare inspections since 2008. This sampling is targeting 5% of the herds with more than ten animals covering cattle, swine, broilers and mink. Welfare inspection in these herds implies checking of compliance with current animal welfare legislation, which is primarily based on resource-based measures for the given species (i.e. for dairy cattle *Act on the keeping of Dairy Cattle and their Offspring* 2010) and the Animal Welfare Act (2013). The risk parameters used in the targeted sampling vary for each year and each species, and for cattle, they have included herd size, antimicrobial consumption, mortality, production type (dairy or beef calves) and abattoir remarks. The novelty in the Danish risk-based welfare inspection system lies in using risk parameters derived from incidence data in national databases for a more targeted sampling. Previous studies have evaluated prediction models based on register data for assessment of animal welfare (Sandgren *et al.,* 2009; Nyman *et al.,* 2011; deVries, 2013) by investigating both uni- and multivariable associations between register data variables and single animal-based measures. However, these previous studies have focused on animal welfare assessed by comprehensive, yet costly, animal-based measures.

Hence, the present study set out to estimate the risk of herds having violations of animal welfare legislation (VoAWL) by investigating data from register data as predictors.

Materials and methods

Study design

This study was designed as a cross-sectional prevalence study of compliance with animal welfare legislation combined with a retrospective follow-up with register data extracted for a one year period ahead of the actual herd visits where compliance with the legislation was controlled among other resource-based welfare measures for a different study purpose by one of four trained observers during March 2010 to July 2011.

Target and study population

The target population for the present study was Danish dairy herds with more than 100 cows and loose-housing systems with cubicles. In 2011, the average herd size for Danish dairy herds was 132 cows, and 64% of the Danish dairy herds had more than 100 cows (Agricultural Statistics 2011). The study population consisted of an amalgamated sample from two other studies. Thus, 88 herds were sampled from a pool of 401 respondents from a questionnaire regarding grazing strategies (Kristensen, 2010). Amongst the current sample, 42 herds used summer grazing and 46 herds did not use summer grazing. The herds were distributed geographically dispersed all over Denmark. A total of 15 herds were excluded from the present study. Eight herds were excluded due to missing or incomplete registrations, four herds due to having fewer than 100 cows on the day of visit, two herds due to deep bedding, and one herd due to missing milk production recordings. Finally, 73 herds remained in the study for further analysis.

Official animal welfare control

Since 2008, an annual risk-based sample of five percent of all livestock herds with more than ten animals kept for farming purposes have been visited by inspectors from the Danish AgriFish Agency (Ministry of Food, Agriculture and Fisheries of Denmark). On average, 669 cattle herds have been inspected per year. The welfare inspection evaluates farms based on the two overarching acts being the Danish Animal Welfare Act (2013) and the Act on the keeping of Dairy Cattle and their Offspring (2010, subsequently referred to as *2010 Act*) as well as acts and executive orders concerning farmed animals, euthanasia, protection of calves, tail-budding and castration, disbudding/dehorning, the use of electrical aggregates, ear tagging and livestock owners' use of pharmaceuticals. The *2010 Act* particularly defines the minimum housing and management standards for dairy cattle based on the most recent recommendations.

Outcome measure: violation of animal welfare legislation (VoAWL)

All barns built after July 2010 are obliged to be designed according to the minimum standards given in the *2010 Act*, while barns built before 2010 are obliged to upgrade the facilities according to given paragraphs related to five of the distinct transition terms (i.e. July 1st 2014, 2016, 2022, 2024, and 2029) until full implementation July 1st 2034. Hence, the resource-based measures were selected in accordance to the Danish animal welfare legislation; however, since none of the barns included in the present study were built after 2010, all minimum standards within the *2010 Act* did not apply to our study herds and resource-based measures were reduced to the following: a) presence of injured animals not separated from the rest of the herd and/or b) animals in a condition requiring euthanasia still being present in the herd. These outcome measures were chosen because they were the most frequently violated aspects in recent welfare inspections (Anonymous, 2011). In case of non-compliance requiring further guidance, inspectors issue warnings; enforcement notices with follow-up are issued in cases where the non-compliance is of more severe character and further guidance is not considered sufficient; finally, the farmer can be reported to the police in case animals are treated recklessly or when previous enforcements have not been met satisfactorily. A farm can receive both a warning, enforcement notice and be reported to the police for different non-compliance issues at the same inspection visit. In the present study, any herd having non-compliance with a minimum of one out of the two measures was regarded as a non-compliant herd with impaired welfare irrespective of the action from the authorities (warning, enforcement notice or police report).

Explanatory variables extracted from register data

The Danish Cattle Database (DCD) is a comprehensive database compiling all mandatory and voluntary routine registrations from the official (Central Husbandry Register, VetStat) and privately owned databases (e.g. milk recording scheme, breeding organisations, abattoirs, laboratory results, and veterinary treatments). A literature review resulted in an initial list of 36 register-based variables with assumed associations with dairy cattle welfare. However, due to a large number of missing observations 11 variables were omitted, leaving 25 variables for the present analyses. These variables covered health recordings, reproduction results, milk recordings, abattoir data and culling data (Table 1). Data were extracted from the DCD for each herd for a one year period ahead of the actual herd visit by the study project. Due to very low within-herd prevalence, the abattoir remarks on lean cows, new fractures, arthritis, joint luxation, locomotor disorders, and bruises were excluded from further analyses. Additionally, based on their distribution in the groups of herds with and without VoAWL the following variables were dichotomized: high bulk tank somatic cell count (≥ 250,000) and normal range (< 250,000), suspiciously low number of veterinary treatment records (≤ 25 treatment records per 100 cow-years) and normal level (> 25 treatment records per 100 cow-years), and unrealistically low proportions of abattoir remarks (no remarks) and normal proportions (> 0 remarks).

Data analysis

All 25 register-data measures were used as explanatory variables accompanied by information on grazing and milking system (automated or conventional) and by information from the Central Husbandry Register on region, production type (organic or conventional) and the *Salmonella* Dublin status in a national surveillance programme (test-negative or test-positive).

Explanatory variables were initially screened in a univariable logistic regression in the statistical software *R* (R Development Core Team, 2012) using the *glm* function. Odds ratios (OR) and 95% confidence intervals (95% CI) were calculated for change in intervals (increase/decrease) depending on the explanatory variable (Table 1). Finally, a multivariable logistic regression model was used to identify predictors of herds with VoAWL, and to estimate the probability of VoAWL as a function of the predictor values.

Results

Presence of sick animals not kept in separation or sick pen were found in 22 herds, while only one herd had the presence of animals requiring euthanasia, yielding a total of 23 herds with VoAWL.

Univariable analysis

The initial univariable screening of associations between explanatory variables and the classification as herds with VoAWL resulted in five variables with a *p*-value < 0.2 (Table 1). Among these five variables, significant associations at a 5% level were found between being a problem herd and an increase in standard deviation of the average milk yield for both first and second lactation cows as well as an association with a decrease in the proportion of abattoir remarks. Figure 1 illustrates the probability of VoAWL for the single continuous variables bulk tank somatic cell count and recorded veterinary treatments, showing an increased probability of VoAWL for an increase in bulk tank somatic cell count and for a decrease in recorded veterinary treatments.

Table 1. Descriptive characteristics of continuous register data variables investigated as explanatory variables for dairy cattle welfare in 73 Danish dairy herds with loose-housing and cubicles and the p-values from the analysis of variance of differences in means in the two groups with and without violations of animal welfare legislation. Odds ratios (OR) and 95 % confidence intervals are given for variables passing the initial screening (*p* <0.2) at the given intervals.

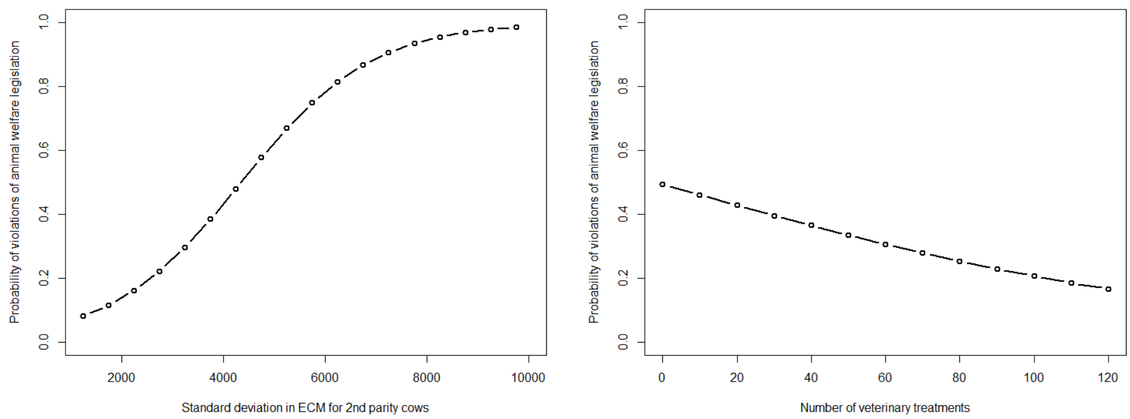


Figure 1. The predicted probabilities of herds having violations of animal welfare legislation plotted against the continuous risk factors standard deviation in milk yield (kg ECM) for second parity cows and the number of registered veterinary treatments.

Multivariable model

Variables identified in the univariable screening with a p-value < 0.2 were considered for the multivariable model. However, the milk yield variables were strongly correlated and only the SD in ECM for 2nd lactation cows was chosen for further analysis because this led to the best fitting model. The proportion of abattoir remarks was low in both outcome groups and hence omitted from the final model, because conditions that are rarely present are not suitable as predictors (only three of herds with VoAWL also had no abattoir remarks). Hence, the resulting model estimating the probability of VoAWL included SD in ECM for 2nd lactation cows, dichotomized bulk tank somatic cell count and dichotomized number of recorded veterinary treatment per 100 cow-years (Table 2).

Table 2. Significant explanatory variables from a univariable screening included in a final prediction model for violations of animal welfare legislation (VoAWL) in 73 Danish dairy herds with specified risk estimates (Odds ratio) for each of the included variables.

Variables	Levels	N herds with VoAWL	Odds ratio (OR)	95% Confidence limits	*p*-value
Herds with VoAWL (N=23)					
SD in ECM 2nd lactation[a]		23	2.2	[1.08;4.83]	*0.03*
Qualitative variables					
Bulk tank somatic cell count	High (≥ 250,000)	12	1.8	[0.9;3.42]	*0.17*
	Low (< 250,000)	11			
Veterinary treatment records	Low (≤ 25)	7	2.2	[1.16;4.13]	*0.07*
	High (> 25)	16			
Abattoir remarks	No remarks	3	2.5	[0.84: 7.55]	*0.11*
	Remarks	20			

[a] *Standard deviation of mean milk yield for the given lactation group*

The predicted probabilities of VoAWL are shown in Figure 2 for the single variables and for the final model in Figure 3. The figure shows the probability of VoAWL for the different combinations of variables included in the final model, e.g. a herd with a high bulk tank somatic cell count (≥ 250,000 cells/mL) and less than 25 recorded annual veterinary treatments per 100 cow-years would have an estimated probability of VoAWL at welfare inspections of approximately 70% if the standard deviation in milk yield among second parity cows was 3000 kg. Model parameters are summarized in Table 3.

Table 3. Parameter estimates, standard error (s.e.), 95% confidence intervals and *p*-values for explanatory variables in the final model for the prediction of violations in animal welfare legislation in 73 Danish dairy herds.

Predictor	Estimate (s.e.)	Lower and upper 95 % CI limits	*p*-value
Intercept	-2.4 (1.4)	[-5.3;0.3]	0.1
Standard deviation in ECM of second lactation cows per 1000 kg	1.2 (0.4)	[0.4;2.1]	0.01
Annual veterinary treatments			
> 25 recorded treatments	-2.1 (0.8)	[-3.8;-0.6]	0.01
≤ 25 recorded treatments	0	0	-
Bulk tank somatic cell count			
< 250,000 cells/mL	-1.1 (0.6)	[-2.3;0.03]	0.06
≥ 250,000 cells/mL	0	0	-

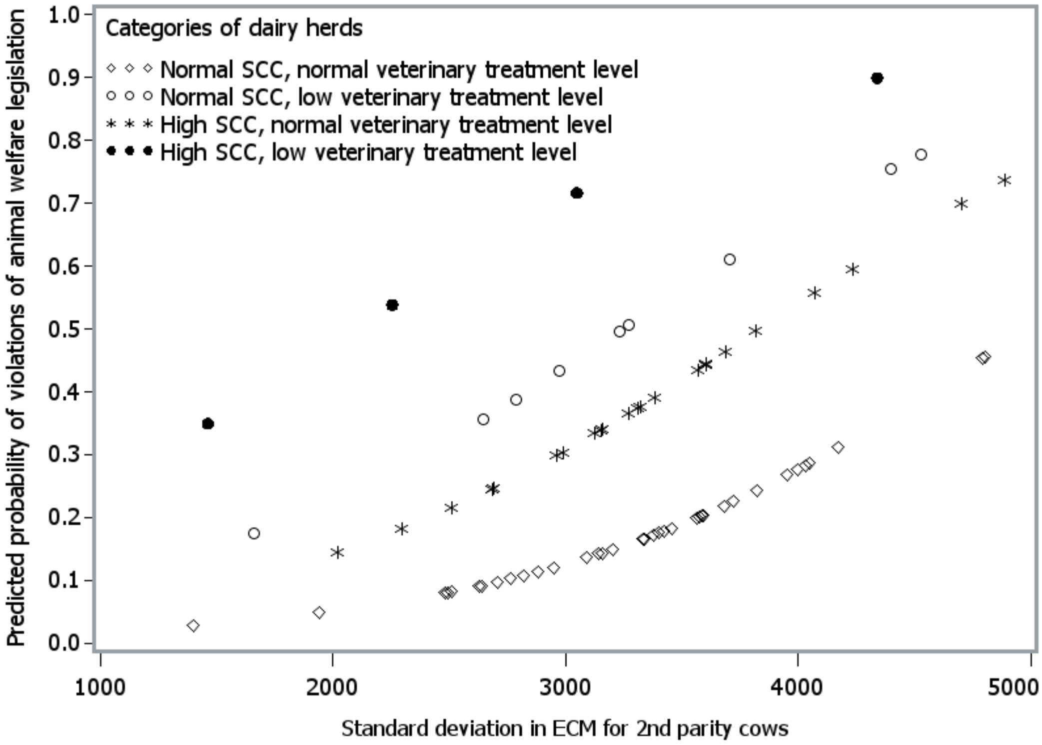


Figure 2. The predicted probabilities of herds having violations in animal welfare legislation based on their levels of three explanatory variables from the final model including the standard deviation in milk yield (kg ECM) for second parity cows based on the qualitative variables high levels of bulk tank somatic cell count (High SCC, ≥ 250,000 cells/mL) or normal range (Normal SCC, < 250,000 cells/mL), low levels of veterinary treatments (≤ 25 annual recordings per 100 cow-years) and normal level of veterinary treatments (> 25 annual recordings per 100 cow-years).

Discussion

The study aimed at identifying predictor variables easily extractable from routine registrations within the Danish Cattle Database to be used in the identification of herds with an increased risk of VoAWL at welfare inspections. The present study identified a limited number of variables associated with violations of animal welfare legislation. Because this study had a cross-sectional study design, the detected associations do not necessarily imply causal relationships between the significant variables and the outcome. Since the associations are more likely to reflect management aspects than animal welfare, caution should be exerted when including these as future predictors for animal welfare. Future field validation of the predictive power of the model is therefore warranted.

Four variables were considered in the multivariable model. Abattoir remarks were excluded, as only three problem herds also had unrealistically low abattoir remarks (no remarks). The final model from the present study differed from previous prediction models. de Vries (2013) reported variables related to demographics (e.g. herd size, distribution of age groups and mortality), milk composition and yield, management and fertility being the most frequently included explanatory variables in final prediction models for single animal-based measures, not an overall composite welfare definition. Mortality of cows less than 60 days in milk was the most frequently included variable in the final models. Two Swedish studies used the same dataset to investigate the predictive potential of identifying herds with poor welfare (Sandgren *et al.,* 2009) and herds with good welfare (Nyman *et al.,* 2011). Here, Sandgren et al. (2009) identified late on going artificial insemination (>120 days), heifers not inseminated at 17 month of age, and calf mortality (age 2-8 month) for the classification of herds with poor welfare at a sensitivity of 62%. However, herds with good welfare, on the other hand, could be classified at a higher sensitivity of 96% based on the same two fertility measures, cow mortality, stillbirth rates, and incidences of mastitis and feed-related diseases.

Assessment of the variation in milk parameters as predictors of VoAWL

The standard deviation of mean milk yield for first and second lactation groups were both significantly associated with the VoAWL; an increase of 1000 kg ECM in standard deviation was associated with a 2.2 times higher odds of VoAWL. Bulk tank somatic cell count was not significantly associated with VoAWL (p= 0.19) in the univariable analysis, but was included as a confounder in the final prediction model.

The association to increased variation in milk yield for lactation groups within a herd and increased bulk tank somatic cell count with VoAWL might be explained by farmers not performing well on feeding, dry cow management, calving management and milking practices, or controlling high levels of lameness, which subsequently are associated with less uniform milk yields and higher somatic cell counts. Known management-related risk factors for high somatic cell count are short post-milking standing time (Watters *et al.,* 2013), metabolic diseases (Nyman *et al.,* 2008) and hygiene aspects (Barkema *et al.,* 1999).

Assessment of the number of veterinary treatment records as a predictor of VoAWL

A decrease in the number of recorded veterinary treatments was not significantly associated (p= 0.07) with VoAWL in the univariable analysis yielding only an OR=1.1. However, the association with the probability of VoAWL and decreasing number of treatment records was not proportional, and dichotomization of the variable yielded an OR=2.2 (p=0.07) and contributed significantly to the multivariable model. The low number of veterinary treatment records could reflect the farmers' treatment threshold, which is in line with findings by Kielland et al. (2010) showing that farmers with low empathy scores towards cattle also ranked painful conditions lower than farmers with a positive attitude and high empathy scores. Additionally, farmers with positive attitudes also had better cow welfare due to lower prevalence of skin lesions. Furthermore, a suspiciously low number of veterinary treatment records in the database may occur, if farmers do not record treatments, which again can be associated with unorganised herd health management that also affects other aspects of the herd welfare level, such as detection and removal or management of sick animals or cattle that should be euthanized.

Kelly et al. (2011) investigated key performance indicators from the Irish Department of Agriculture, Fisheries and Food´s national databases on livestock herds to characterize 18 case cattle herds with recorded farm animal welfare incidents. Four indicators were identified including late registration of calves, carcase disposal by use of on-farm burial, increased transport of animals to incineration plants over time, and records of movements of animals to herds unknown. These indicators correspond very well to accessible data from the DFVA and late registration of calves and mortality rate have been used as risk-parameters for sampling in previous Danish welfare inspections.

Assessment of the proportion of abattoir remarks as a predictor of VoAWL

A 10% decrease in proportion of abattoir remarks yielded a 4.4 times higher odds of VoAWL in the univariable analysis. However, dichotomization of the variable left only three herds with no abattoir remarks and VoAWL. Furthermore, the proportion of abattoir remarks is a variable that is not only influenced by herd management, but also by the abattoirs, so this was assessed as a less suitable variable for prediction of herds with VoAWL.

Nonetheless, the potential association between VoAWL and low number of abattoir remarks was unexpected. The findings in the present study did not show increases in abattoir remarks as a risk but a decrease in abattoir remarks as a potential risk, which was much unexpected. This might be due by the very little between-herd variation for this variable. A speculation might be, that farmers with on-going animal welfare problems do not send injured or recently ill animals to slaughter either because they risk fines for delivering animals in a poor health condition, or because these cows are not suited for transportation or actually die on-farm either assisted or unassisted.

Neither herd size nor mortality showed any significant associations with VoAWL in the present study. The lack of agreement between the present findings and previous studies is most likely to be found in the differences of the outcome definition - whether animal welfare is defined by minimum legislative standards or by direct observations of the animals (animal-based welfare assessment). Certainly the small sample size also contributed to the lack of finding any coherent risk factors for high levels of welfare related disorders. The current sample only considered herds with more than 100 cows, while the official surveillance scheme may target herds from as little as ten animals per herd. However, there are no official results indicating what herd size group might be overrepresented in the non-compliant group found in the official control. The use of privately owned routinely collected data has not been used in the official welfare inspection so far, as milk recordings and breeding data are privately owned. Nonetheless, the incorporation of these parameters might have potential to identify more true case herds and thereby improve the efficiency of the welfare inspections.

Outcome of interest as an animal welfare indicator

It should be stressed that the present study did not investigate the actual prevalence of poor animal welfare as a complex entity. The official animal welfare inspection is not assessing animal welfare per se, but is controlling compliance with current animal welfare legislation. Hence, it could be possible for a herd with VoAWL to have low prevalences of clinical or behavioural poor welfare measures. Results might have been different, if the outcome definition had been based upon animal-based measures instead. Previous risk factor studies have only investigated single components of animal welfare like animal-based measures e.g. lameness prevalence (Alban *et al.,* 1996; Green *et al.,* 2002; Haskell *et al.,* 2006; Dippel *et al.*, 2009), hock lesions (Rutherford *et al.,* 2008; Kielland *et al.,* 2009) or mortality (Alvåsen *et al.,* 2012; Thomsen *et al.,* 2004). Although, other studies have evaluated register data performance in predicting herd animal welfare, there have been major differences in the case definition of animal welfare. deVries (2013) used similar range of register data to predict the single welfare measures included in the Welfare Quality® (WQ) protocol, but no associations between the register data and the overall WQ® score were made. It would be notoriously difficult to obtain meaningful register-based predictors of a poor WQ® score, because the score in itself is a composite number based on many different observations and evaluations weighted on different scales.

However, welfare assessment and welfare inspection are fundamentally different. Thus, the aim differs substantially whether welfare is evaluated quantitatively in a large-scale cross-sectional welfare assessment system or whether it is to control minimum requirements settled within legislation. Where the welfare assessment defines the welfare level based on a very explicit welfare definition combined into e.g. an animal welfare index, the inspection is based upon proxies for welfare. Evaluating resource-based measures as proxy for animal welfare has been practiced in other settings than the Danish, i.e. the *Tiergerechtheitsindex* (TGI) being implemented in Austria in 1995 (Bartussek, 1999) or in quality assurance schemes like 'Freedom Food' defined by the Royal Society for the Prevention of Cruelty to Animals in the UK. Further refinement of e.g. animal welfare legislation should also be built upon existing knowledge and implementation of appropriate risk evaluation methods (Marahrens *et al.,* 2011). Risk assessment is already an integrated part of food safety and could provide a standardized methodological framework for assessing the effects of certain exposures on animal welfare as suggested by others (Marahrens *et al.,* 2011; Collins *et al.,* 2012; Paton *et al.,* 2013). However, the major requirement for an animal welfare risk assessment method would be the general agreement on the framework in order for it to be applied consistently (Paton *et al.,* 2013).

Conclusion

The present study identified variation in milk yield in first and second parity cows, ≤ 25 veterinary treatment records per 100 cow-years and bulk tank milk SCC ≥250.000 as significant predictors of dairy herds with welfare problems expressed as non-compliance with animal welfare legislation. However, since risk factors are highly dependent on the welfare definition there is a need for further investigation of possible risk factors for animal welfare covering more than just minimum legislative standards.

References

Act on the keeping of dairy cattle and their offspring. 2010. Lov om hold af malkekvæg og afkom af malkekvæg Lov nr. 520 Danish Ministry of Jusitice.

Agricultural Statistics. 2011. Landbrug 2011. Statistik om landbrug, gartneri og skovbrug. Danmarks Statistik. Available online: www.dst.dk/Publ/Landbrug.

Alban, L., Agger, J. F. and Lawson, L.G. 1996. Lameness in tied Danish dairy cattle: the possible influence of housing systems, management, milk yield, and prior incidents of lameness Preventive Veterinary Medicine 29(2): 135-149.

Alvåsen, K., Jannson Mörck, M., Hallén Sandgren, C., Thomsen, P.T. and Emanuelson, U. 2012. Herd-level risk factors associated with cow mortality in Swedish dairy herds. Journal of Dairy Science 95: 4532-4362.

Anonymous. 2011. Dyrevelfærdsrapport 2011. Videnscenter for Dyrevelfærd, Ministeriet for Fødevarer, Landbrug og Fiskeri, Fødevarestyrelsen, Glostrup.

Barkema, H.W., Deluyker, H., Schukken, Y. and Lam, T.G.J.M. 1999. Quarter-milk somatic cell count at calving and at the first six milkings after calving. Preventive Veterinary Medicine 38:1–9.

Bartussek, H. 1999. A review of the animal needs index (ANI) for the assessment of animals' well-being in the housing systems for Austrian proprietary products and legislation. Livestock Production Science 61: 179-192.

Burow, E., Rousing, T., Thomsen, P.T., Otten, N.D. and Sørensen, J.T. 2013. Effect of grazing on the cow welfare of dairy herds evaluated by a multidimensional welfare index. Animal 7(5): 834-842.

Cochran, W.G. 1977. Sampling Techniques. John Wiley & Sons. New York, pp 74-76.

Collins, L.M. 2012. Welfare risk assessment: the benefits and pitfalls. Animal Welfare 21(S1): 73-79.

Danish Animal Welfare Act .2013. Dyreværnsloven LBK nr 252 Danish Ministry of Food, Agriculture and Fisheries.

DeVries, M. 2013. Assuring dairy cattle welfare: towards efficient assessment and improvement. PhD thesis. Wageningen Univeristy dissertation no. 5486.

Dippel, S., Dolezal, M., Brenninkmeyer, C., Brinkmann, J., March, S., Knierim, U. and Winckler, C. 2009. Risk factors for lameness in freestall-housed dairy cows across two breeds, farming systems, and countries. Journal of Dairy Science 92: 5476-5486.

EFSA Panel on Animal Health and Welfare. 2012. Guidance on risk assessment for animal welfare. EFSA Journal 10: 2513.10.2903/j.efsa.2012.2513

Farm Animal Welfare Council. 1992. FAWC updates the Five Freedoms. Veterinary Record 17: 357.

Green, L.E., Hedges, V. J., Schukken, Y.H., Blowey, R.W. and Packington, A.J. 2002. The impact of clinical lameness on the milk yield of dairy cows. Journal of Dairy Science 85(9): 2250-2256.

Haskell, M.J., Rennie, L. J., Bowell, V.A., Bell, M.J. and Lawrence, A.B. 2006. Housing system, milk production, and zero-grazing effects on lameness and leg injury in dairy cows. Journal of Dairy Science89: 4259-4266.

Keeling, L. (Ed.), 2009. An Overview of the Development of the Welfare Quality Project© Assessment Systems. Welfare Quality Reports no.11. Cardiff University, United Kingdom.

Kielland, C., Ruud, L.E., Zanella, A.J. and Østerås, O. 2009. Prevalence and risk factors for skin lesions on legs of dairy cattle housed in freestalls in Norway. Journal of Dairy Science 92: 5487-5496.

Kristensen, T. 2010. Produktionssystemer i de danske malkekvægbedrifter. Landbrugsinfor 2117. Available online: https://www.landbrugsinfo.dk/Kvaeg/Foder/Grovfoder/Afgraesning/Sider/Produktionssystemeridanskemalkekvaegbedrifter.aspx. Accessed 10-02-2013.

Marahrens, M., Kleinschmidt, N., Di Nardo, I., Velarde, A., Fuentes, C., Truar, A., Otero, J.L., Di Fede, E. and Dalla Villa, P. 2011. Risk assessment in animal welfare – Especially referring to animal transport. Preventive Veterinary Medicine 102:157-163.

Nyman, A.K., Emanuelson, U., Holtenius, K., Ingvartsen, K.L., Larsen, T. and Persson Waller, K. 2008. Metabolites and Immune Variables Associated with Somatic Cell Counts. Journal of Dairy Science 91: 2996-3009.

Nyman, A.K., Lindberg, A. and Sandgren, C.H. 2011. Can pre-collected register data be used to identify dairy herds with good cattle welfare? Acta Agriculturae Scandinavica, Section A – Animal Science, 53 (S1): S8

Paton, M.W., Martin, P.A.J. and Fisher, A.D. 2013. Risk assessment principles in evaluation of animal welfare. Animal Welfare 22: 277-285.

Rutherford, K.M.D., Langford, F.M., Jack, M.C., Sherwood, L., Lawrence, A.B., Haskell, M.J. and Carre, X.W.o.S.S.l.c.d. 2008. Hock injury prevalence and associated risk factors on organic and nonorganic dairy farms in the United Kingdom. Journal of Dairy Science 91(6): 2265-2274.

Sandgren, C.H., Lindberg, A. and Keeling, L.J. 2009. Using a national dairy database to identify herds with poor welfare. Animal Welfare 18: 523-532.

Scott, M., Fitzpatrick, J.L., Reid, J. and Wiseman, M.L. 2003. Evaluation of welfare state based on interpretation of multiple indices. Animal Welfare 12(4): 457-468.

Thomsen, P.T., Kjeldsen, A.M., Sørensen, J.T. and Houe, H. 2004. Mortality (including euthanasia) among Danish dairy cows (1990-2001). Preventive Veterinary Medicine 62(1): 19-33.

Watters, A.M.E, Meier, K.M.A, Barkema, H.W., Leslie, K.E., von Keyserlingk, M.A.G. and DeVries, T.J. 2013. Associations of herd- and cow-level factors, cow lying behavior, and risk of elevated somatic cell count in free-stall housed lactating dairy cows. Preventive Veterinary Medicine 111(3-4): 245-255.

6. General Discussion

The thesis investigated the possibilities of identifying dairy herds with potential animal welfare problems based on different levels of information. The investigations involved a stepwise bottom-up process, starting with evaluating welfare indicators and measures before combining these into an index, and finally constructing a predictive model for the presence of animal welfare problems. This chapter will summarize the findings and discuss the methodology, validity and outcome of the research performed in the pursuit of the three objectives. Despite having an overlap in data quality concepts each objective presents its own specific challenges concerning data quality. Hence, validity will be discussed in relation to each of the three objectives Findings and challenges will be discussed in relation to the overall aim of the thesis in order to reach to a conclusion of the thesis and to present perspectives for further research.

6.1 Animal welfare indicators

6.1.1 Register data as diagnostic indicators

In paper 1 the diagnostic potential of register data to predict clinical lameness was evaluated. A predictive model of annual mortality rate, bulk tank somatic cell count, lean cows at slaughter and the standard deviation in age at first calving showed a predictive ability to detect herds with high lameness prevalences of AUC=0.79, implying that the model would correctly classify herds in 79% of the cases. However, the model did not significantly improve its AUC by adding on more variables than mortality. Looking at sensitivity, mortality also prevailed with a combination of sensitivity (Se) of 100% and specificity (Sp) of 53%, yielding the highest differential positive rate (DPR). The associations found in paper 1 were not unexpected, as lameness has been positively associated with mortality in several studies (Esslemont & Kossaibati, 1997; Rajala-Schultz & Gröhn, 1999; Boot *et al.*, 2004; Bicalho *et al.*, 2007), and low age at first calving (Rutherford *et al.*, 2009), and low BCS (Espejo *et al.,* 2006; Dippel *et al.*, 2009).

An interesting finding, however, was the issue of optimized versus pre-defined cut-offs. The pre-defined cut-offs were evaluated as they form the basis of the current risk-based identification scheme, using register data as risk parameters in their targeted sampling. To the authors' knowledge, at present there are no studies reporting risk factors for an overall estimate of animal welfare *per se*, hence, the use of these risk parameters have been based on the assumption that register variables can be used as a proxy for the underlying or latent condition of welfare. Instead of quantifying the risk parameter a cut-off was set for the discrimination between a problematic level of the given risk parameter and the non-problematic levels. For mortality the cut-off has been based on the 50th-percentile (Lund Nielsen, 2013 *personal communication*), where herds among the 50% highest mortality levels were regarded at risk of animal welfare problems. This assumption is debatable. Thus as argued by Thomsen et al.

(2004) a high mortality rate could reflect the herd managers' choice to euthanize animals to prevent pain and suffering actually improving welfare in the given herd. Nonetheless, an increased mortality rate as a reflection of a higher morbidity is indicative of impaired animal welfare. In regarding a parameter exceeding a certain threshold to classify itself as a problem this threshold is, however, not equivalent to the threshold needed for that given parameter to explain another parameter, here being the clinical finding of high lameness prevalence. In this situation, the optimal threshold level has to be established separately for the specific purpose. This is clearly illustrated by the results in paper 1, where the thresholds or cut-offs maximizing the DPR were consequently lower than the national means, which constituted the predefined cut-offs.

In paper 1 the focus was on maximizing DPR, i.e. to get the best trade-off situation between Se and Sp. However, another approach could have been to maximize Se (or Sp) instead. The choice is highly dependent on the purpose of the application of the diagnostic tool and the population it is to be applied within as Se and Sp are population specific parameters under specified conditions (Greiner & Gardner, 2000). Since register data here are regarded as screening tests for a given cross-sectional state, the intuitive measure of interest would be sensitivity in order to catch as many true positive cases as possible. In regards to the sampling scheme for welfare inspections, this is also the case, as the initial sampling could be regarded as a screening test with high Se followed up by the actual welfare inspection with high Sp, aiming at increasing the positive predictive values (PPV), i.e. the probability that a herd with a positive risk parameter actually also has animal welfare problems, by reducing the number of false positives (Dohoo *et al.*, 2009). On the other hand, from the farmers' perspective an initial screening should be of high Sp in order to avoid false incrimination if the welfare is actually acceptable. However, the validity of the included measures is equally important as it contributes to the overall accuracy of the risk parameters.

Data quality

Secondary databases with animal-disease information offer a great potential for epidemiologic research (Lawrenson *et al.*, 1999; Mörk *et al.,* 2009). Advantages of such secondary databases are evidently the availability of the vast entity of data at low cost. Although the potential is great, a number of pitfalls should be kept in mind when building research upon these databases. This issue is discussed by Lawrenson et al. (1999) for the use of databases in human disease recordings and by Mörk et al. (2010; 2009) for veterinary databases. The disadvantages of these data are simply the researchers' lack of influence on data collection and data quality unless the secondary database is validated (Mörk *et al.*, 2009).

Validity in terms of register data refers to the *completeness* (i.e. sensitivity of recording) and the *correctness* (i.e. positive predictive value) (Jordan *et al.*, 2004). Combining these test characteristics yield the *accuracy*, this is also termed as correctness by the EFSA report (2012a) on animal-based measures. The register data within the DCD are gathered on cow level with varying levels of sensitivity depending on the outcome variable. Highest Se estimates are to be expected for the objective and automated measures of milk quality and yield data, as these are gathered on a regular basis for all lactating cows. Reproductive data are troublesome, as they are indeed prone to the managers choices, e.g. the of age at first service for heifers – whether this regimen follows the weight or height of heifers; or the length of the calving-to-service interval following a fixed target or depending on the cows lactation curve. The veterinary treatments show varying Se estimates from 0.37 (Lind *et al.*, 2012) for locomotor disorders to

0.84-0.88 for metabolic diseases (Espetvedt *et al.*, 2012) and 0.94 for mastitis cases (Wolff *et al.*, 2012). Abattoir data on specific findings have varying Se estimates as well. Hence, Bonde et al. (2010) found large variation Se for meat inspection in swine for four different pathological groups with Se as low as 0.16 to as high as 0.92, while Sp in general was high (0.98-1). No comparable estimates were found for meat inspection in cattle.

The relative small sample size in paper 1 does challenge the external validity of the given prediction model, as the model fits for the study population (i.e. internal validity), but needs to be tested in other herds among the target population as well (Dohoo *et al.*, 2009) to ensure the representativeness. Nonetheless, results indicate that register data collected for other purposes do have a potential to predict cross-sectional findings if uncertainty concerning variables are taken into account. These uncertainties, however, are not only restricted to the register data variables – the clinical scoring by different observers are very much subject to uncertainty as the next section regarding paper 2 will discuss.

6.1.2 Animal-based measures

Register data are prone to information bias due to the farmers' threshold for e.g. euthanasia, treating diseased animals or selecting animals for slaughter. The bias associated with clinical scoring is centred on the misclassification of the given observation. This issue became evident in paper 2, as a risk factor study for lameness was used to exemplify the pitfalls of relying on traditional inter-observer agreement methods (e.g. kappa values). The study showed an unexpected association between grazing and lameness prevalence, as grazing turned out to be a risk factor for the apparent prevalence. However, the distribution of the four observers among grazing and non-grazing herds was extremely unbalanced. Observers had undergone calibration testing on video material with good agreement on the three lameness levels normal, moderately lame, and severely lame. However, thorough investigations revealed two observers scoring higher than the two others. Additionally, these two were responsible for 80% of the scorings in grazing herds. This systematic effect called for a solution in order to eliminate the bias introduced by the observer. Since no gold standard existed for the lameness evaluation the true lameness prevalence was estimated using a two-step Bayesian estimation model. In the first step observer specific sensitivity and specificity was estimated based on the information from the calibration test. In the next step, these estimates were used to calculate the true within-herd prevalence and effects of grazing were evaluated in a latent class model. This model showed the expected higher true prevalences in grazing herds. Two distinct points were illustrated in paper 2 – the importance of the study design (as discussed in paper 2) and the validity of raw estimates of clinical measures – i.e. the ability of these measures to correctly assess the welfare outcome (EFSA, 2012a).

The direct observations of animals' responses to their environment are regarded as the most essential welfare measures (Scott *et al.,* 2003) with higher **validity** than resource or management measures (Knierim & Winckler, 2009). However, this validity is perceived as the ability of the given measure to actually measure welfare outcomes, or so called **measurement validity**. The results from paper 2 indicate, that validity indeed is challenged in multi-observer studies, since clinical measures are far from objective. Physical factors like light, herd size, hygiene and floor type have shown to influence the quality of the clinical observations (Baadsgaard & Enevoldsen, 1997). Hence, personal thresholds for observers do get mixed with the rational scoring compromising the validity of the given measure – a detrimental problem if not taken into account.

The adjustment for these observer effects are, however, not that straightforward after all in the case of welfare assessment. In order to obtain a nuanced picture of the current welfare status of a herd, more measure levels are needed. Traditionally, the **reliability** of scaled measures is assessed by evaluating inter- and intra-observer agreement. According to the EF-SA report (2012a) reliability and repeatability are used to cover the **robustness** of measures. Within the evaluation of clinical measures in the Welfare Quality® protocols this has been performed using the prevalence-adjusted bias-adjusted kappa (Brenninkmeyer *et al.,* 2007; March *et al.,* 2007). Both studies highlighted the challenges in achieving robustness in agreement between observers. However, little concern has been given to the misclassification bias within the Welfare Quality® scope.

Improving agreement requires clear definitions of and optimal cut-offs between categories. The choice of cut-off is dependent on the purpose of the given study, as mentioned in Section 6.1. Thomsen & Baadsgaard (2006) stating when the purpose is to detect all cases regardless of the number of false positives a low cut-off yielding a higher Se will be chosen. However, in their study the choice of cut-off with higher specificity increased the inter-observer agreement. In addition, Brenninkmeyer et al. (2007) showed increased agreements when scoring categories were reduced from five to two (lame or non-lame). These arguments lead to the discussion of how many measure categories or levels are needed within welfare assessment? Issues addressed in the following section concerning paper 3. However, using latent class modelling is limited to binary outcome levels and thus its application within reliability assurance in welfare assessments is limited.

6.2 Aggregation of measures – welfare index

In paper 3 information from three different types of data (register data, system and clinical and behavioural observations) was aggregated into overall animal welfare indices (AWI) and agreement on the ranking of farms by these AWI´s was evaluated by the Spearman Rank Correlation Coefficient. Register data were assessed for three different time periods (365, 180, and 90 days) prior to the on-farm data collection in order to evaluate the most optimal sampling frame for routine registrations. Results showed weak and contradicting correlations between the three AWI´s, and hence the hypotheses III stating that herds with animal welfare problems can be identified by existing data without visiting the farm can be rejected. The study confirmed the challenges in assessing the multi-dimensionality of animal welfare. Previous studies have focused on evaluating associations between register data variables and single welfare measures and not with a direct overall welfare score or index score. Hence, no real foundation for comparison with previous studies was available.

Issues in animal welfare assessment – measures and indicators of animal welfare

"When you cannot measure it, when you cannot express it in numbers – you have scarcely in your thoughts, advanced to a stage of science, whatever the matter may be", Lord Kelvin (in Chapman, 1976).

This quote is very striking in relation to animal welfare science, as animal welfare scientist, not only from different welfare schools but also different ethical stand points, have argued intensely on how to conduct animal welfare assessments and measure welfare.

Historically, the early assessments of animal welfare focused on the feasible resource-based measures for quality assurance or certification purposes. The approach, however, is rather operating with **welfare indicators** giving an **indirect indication of animal welfare**, as these factors are regarded as risk factors for impaired animal welfare although animals might not be affected after all. Nonetheless, the mere assumption of causality drives the use of these welfare indicators. The probability of a potential risk can be affected by several aspects. For one, the intensity and duration of which animals are exposed to the risk play a critical part in developing adverse welfare effects. Secondly, the management of these hazards or risks can be counteracted by good management practices.

More recently scientists argued more in favour of the **direct indications of animal welfare** as expressed by the animal-based **welfare measures** (Whay *et al.,* 2003; Scott *et al.*, 2003; Blokhuis *et al.*, 2003; Webster *et al.*, 2004) – a truth that can be modified according to Sandøe & Simonsen (1992), arguing that animal welfare cannot be measured directly. They build their argumentation on the fact, that welfare can only be assessed by thorough ethical reflections and hence they argue that researchers should restrict their efforts to assess objective facts (i.e. behavioural and physical measures) and engage into inter-disciplinary discussions with philosophers and ethicists.

Many welfare assessment protocols have seen the light of day since with the most comprehensive work carried out by the EU Framework 6 funded Welfare Quality® project (2004-2009). Engaging 44 institutes and universities from Europe and Latin America, the developed welfare assessment protocols can be regarded as the state-of-the-art of animal welfare assessments. Hence, the three welfare assessment approaches used in paper 3 were based on the same framework (described in Section 2.3.3). However, the protocols were modified due to practical and operational issues. The Welfare Quality® protocol for dairy cattle is primarily based on the direct animal-based measures and supplemented by resource measures in case of insufficient coverage by the animal-based measures (e.g. water provision) (Veissier *et al.,* 2011). Since, the results of the study described in paper 3 were contradicting and not really convincing, a couple of issues need to be discussed further.

Index validity

The idea of expressing animal welfare is very appealing in regards to overall benchmarking between herds or in a certification context. An index score is a ratio that is used to describe a development over a fixed period by comparing it to a baseline score. Hence, the index provides a good overview of the development of the outcome defined by the composite measures included in the index. As discussed earlier, the evaluation of measures and measurement tools is performed in regards to validity and reliability (Scott *et al.,* 2001). This main question is: is the tool actually measuring the attribute of interest?

To ensure the **content validity** of the given assessment protocols both the included measures and the final animal welfare index (WAI) should be evaluated. The chosen register variables were based on a literature review (see Table 3.1) and therefore their measurement validity can be regarded as fair for the purpose. However, especially disease recordings are prone to be biased at many steps on the way from the observation to the actual recording in the system. Starting with the observation, measure severity is assessed by the observer. His or hers' threshold determines the scaling of the given measure and the following interpretation determines whether it will result in a record made in the database either due to his own intervention or by the veterinarians interpretation. Nonetheless, as mentioned in the previous section, estimates for validity of disease recordings do exist and can be dealt with. The individual resource- and animal-based measures are assumedly highly valid, as they have been included from previous acknowledged protocols. Hence, measures included in the applied assessment protocols are valid measures. However, the next question is, whether it can be concluded, that the content of the applied welfare assessments or AWI's are valid?

Coverage

The coverage of all of the twelve Welfare Quality® criteria (Table 2.3) was not fully accomplished by any of the three animal welfare indices, since measures were aggregated based on their characteristics (i.e. register-, resource- and animal-based) instead of an integrated approach. Hence, the three different AWI's reflect on different aspects of animal welfare and the direct comparison of the ranking of farms becomes muddled. Aggregating measures into an index might contribute to the loss of information on potentially well associated welfare indicators e.g. from register data. Nonetheless, the contents of the measures included in the three AWI's were representative.

Expert opinion on measure weights

Aggregating measures into an overall index presents fundamental challenges of a more ethical character. The optimal model should regard the multi-dimensionality of welfare and reflect the relative importance of the measures included (Botreau *et al.*, 2007a). The relative importance of measures can be derived by expert opinion as done in the present index aggregation. This approach was chosen to accommodate the **face validity** of the index content. As described by Scott et al. (2001), the face validity covers a subjective judgement by experts in regards to the validity of the tool. The purpose of the expert panel query in the present study was to obtain weights used for the scaling of measures within each index.

Weights were derived using direct estimation techniques for scaling of measures although indirect techniques are proposed by Scott et al. (2001) for assigning weights used in a score reflecting animal welfare. However, the choice between scaling methods has been debated within clinical settings favouring both one and the other (Wright & Feinstein, 1987; McDowell & Newell, 1987). The direct technique involves a subjective judgement; hence, Scott et al. (2001) question the validity of this approach. In this light another obstacle is presented by the experts' reluctance to directly assign a relative weight (Rodenburg *et al.,* 2003). However, Spoolder et al. (2003) stated that deciding on welfare measures and their importance for animal welfare assessment would require some degree of subjectivity. Rodenburg et al. (2008) argued that welfare scientists including ethologists and veterinarians are presumably better qualified than lay people to make judgement of the overall animal welfare state based on complex datasets on various welfare indicators. The present study addressed a relative wide range of potential expert panel members – varying from animal welfare scientists to practising veterinarians, production advisors and animal welfare organisation delegates in order to account for the likely differences in stakeholder opinion. The use of balanced panels with heterogeneous experts has been widely used (Main *et al.,* 2003; Webster *et al.*, 2004; Bracke *et al,* 2008; Phythian *et al.,* 2011). In contrast, Jensen et al. (2012) relied on a homogenous and highly specialized expert panel consisting of animal behaviour scientists in one part and practising veterinarians in another part of their study quantifying pain and production consequences of lameness in finishing pigs. The number of experts needed is directly linked to the choice of panel – if highly specialized experts are used within-expert variation is expected to be minimized and hence a smaller number of experts is needed. Previous studies have used five (Bonde *et al.,* 2009), eight (Jensen *et al*., 2012) and 13 experts (Rodenburg *et al.,* 2008) for specialised and 22 to 56 for the more heterogeneous panels (Main *et al.,* 2003; Bracke *et al.,* 2008; Phytian *et al.,* 2011). Hence, the number of 20 and 15 experts, respectively, in the two surveys performed in the research for this thesis could be considered sufficient substantiated by the limited variation found in measure weights.

The weights derived by the present expert survey were in line with previous findings. Lievaart & Noordhuizen (2011) investigated expert opinion on suitable welfare measures for dairy cattle and reported high ranking of the following measures: lameness, BCS, hock lesions, separate calving pen, SCC in milk, stocking rate (overcrowding), avoidance distance and floor type. Similar patterns were found in the present surveys where highest weights were assigned to mortality and milk yield data variables, water and feed supply, stocking density, calving pen and separation of sick animals, lameness, BCS and integument alterations. It could also be argued that experts assigned higher weights according to their belief in the intensity of pain inflicted by the given measures as presented in Table 2.1.

Aggregation model

Gathering all the different individual measures on cow level into one overall welfare descriptor on herd level comes with a number of challenges to ensure **construct validity** of the measuring tool. Firstly, measures are observed at different scales (e.g. clinical data on an ordinal scale normal/moderate/severe, behavioural data on a continuous scale in seconds or centimetres). Secondly, initial data are observed on cow level, but the unit of interest is the herd level. Hence, the aggregation method has to be appropriate for the types of scales.

For the aggregation model variables were aggregated onto herd level based on different approaches: register data variables were assessed based on percentiles among the study population, resource-based measures were assessed for compliance with Danish legislation and recommendations for housing, and the animal-based measures were assessed as herd level prevalences. Thus, it can be debated that direct comparison of the three indices is applicable, as one index is dependent on the distribution of variables among the study population given by percentiles, while the other two indices can be used in any population. The aggregation framework, however, remained the same for all three variables types, as the sum of weighted scores was the chosen approach to compare the ranking of herds within the study population. This approach holds the advantages of scores being obtainable on any farm and further sums allow for compensation between measures. However, this also implies a major disadvantage as the approach does not clearly illustrate whether a moderate score has arisen due to moderate mean scores across all measures or due to very low scores for some measures equalling out very high score for other measures (Botreau *et al.,* 2007a). Additionally, the approach does not consider the severity of animals having multiple severe impairments. This limits the possibility of pointing out very problematic areas in a given farm which could have been used in a constructive advisory situation. The assignment of measure weights could potentially counteract or at least reduce this bias. Finally, the choice to use a linear aggregation (i.e. simple summation scale) stood in contrast to the non-linear functions (i.e. progressive scale) applied in the Welfare Quality® scheme. Here measure weights were derived from experts scoring virtual datasets and ranking herds based on these. The ranking was the used to deduct measure weights. It turned out that non-linear functions representing expert preferences were to be used in order to match the expert opinion (Bonde *et al.,* 2009). This was due to experts varying judgement of increments in scores given to increases in the prevalences of anomalies in the bottom versus the top of the prevalence scale. It can be argued that a linear reasoning is biologically more plausible, as an increment basically is an objective state and only when we pass ethical or emotional judgement upon it, it becomes associated with non-linear reasoning. Furthermore, the linear approach yields a more transparent aggregation which should be a prerequisite of animal welfare assessment (Winckler *et al.,* 2003; Webster *et al.*, 2005).

Model validity – internal and external validity

The results in paper 3 only showed significant correlation between the animal-based index and the register based index for a period of 180 days prior to the actual herd visit. Different time periods were investigated because dairy herds are very dynamic populations with many exits and entries during a year. Reasons for the correlation between the 180 day period and the welfare state measured on-farm but not with the other periods are therefore hard to find. In most register based studies annual means are used. In the present context it could be argued that the missing correlations are due to welfare indices are in fact measuring different welfare states. Register data cannot describe behavioural aspects directly, as direct causality between production outcomes and behavioural is hard to encompass. Previous findings show associations between e.g. weight gain and fearfulness in chickens (Gross & Siegel, 1981; Hemsworth *et al.,* 1994; Hemsworth *et al.,* 1996). For dairy cattle findings have included associations between fearfulness and milk yield (Breuer *et al.,* 2000), and negative handling (Breuer *et al.,* 1997; Waiblinger *et al.*, 2002). However, similar associations have been found for several other clinical parameters e.g. lameness and milk yield etc., making the current possible associations held within the indices blurry.

The **internal validity** (described in Section 2.4.1) of the aggregation models has been evaluated by Burow et al. (2013) testing the hypothesis whether the animal welfare score of herds will improve during summer grazing. However, the welfare index used for evaluation contained both resource- and animal-based measures in contrast to the separated indices in the present paper 3. Thus, an index specific evaluation still needs to be performed.

The **external validity** should be granted as the study population is very representative in terms of herd characteristics among the Danish dairy herds. The representativeness is however challenged by the lack of extreme welfare scores among the sample. This could either be explained by a general good welfare standard amongst Danish dairy herds or it by the selection bias of herds included for the study. Practical and financial constraints due to limited budgets only enabled the present sample to include 86 herds of which complete datasets were present for 73 herds at the end. Compared to other on-farm welfare assessment studies in dairy herds the present sample size was fair. Under comparable production settings sample sizes have ranged from 41-43 in Danish studies (Andreasen *et al.*, 2013; Burow *et al.,* 2013) while Swedish studies investigated 55 herds (Sandgren *et al.,* 2009; Nyman *et al.*, 2011) and 196 Dutch herds (deVries *at al.,* 2013 *unpublished*). There is no doubt that an increased sample size could not only possibly have increased the number of associations, but also opened up for the use of other analytical methodologies.

Multivariate methods

The identification of welfare indicators among register data variables could be assessed by other approaches. Multivariate analyses investigating more than one outcome variable, e.g. lameness and dirty hind legs. Among the variable reduction approaches factor analysis, principal component and correspondence analysis can be mentioned (Dohoo *et al.*, 2009). Factor analysis is a technique used to determine indicator groups or unobservable features associated with the given outcome based on variable pattern loadings (Sharma, 1996). However, although experts differ in their opinions on the required minimum sample size the smallest number is either 100 observations (MacCallum *et al*., 1999) or five times the number of variables (Hatcher, 1994). Other reduction techniques are Principal component analysis (PCA) and correspondence analysis (Dohoo *et al.,* 2009).

Principal component analysis is often the method of choice within behaviour sciences. The method transforms possibly correlated variables into different non-correlated linear components yielding dimensional reduction without the loss of information. This method is also used in the evaluation of the Qualitative Behaviour Assessment (QBA) within the Welfare Quality® protocol (Wemelsfelder *et al.,* 2001). Correspondence analysis analyses the relationship between categorical variables to detect clusters of predictors. Finally, a more contradictory approach within animal welfare the Test theory approach was used by Herva et al. (2009) to validate a welfare index against weight gain in beef cattle. Test theory considers more latent constructs in contrast to the direct measures used in welfare assessment, offering a more explorative approach to large sets of variables.

6.3 Welfare indicators and risk factors

In paper 3 the correlations between single register data variables and the animal-based welfare index were investigated. Significant correlations were found for abattoir remarks and mean bulk tank somatic cell count as well a tendency to significance for calf mortality. Similar findings were made in paper 4 investigating risk factors from register data for herds having violations of animal welfare legislation that could result in a warning, enforcement notice or police report if official welfare inspections were carried out. The outcome of interest was presence of animals in a state requiring euthanasia (only present in one herd) or sick/weak animals not separated from the flock (23 herds) corresponding to 32% of the study herds. This proportion is, although slightly higher, in alignment with previous findings of 24% with animal welfare remarks as reported by Sandgren et al. (2009) and the annual rate of 17-25% of herds found with welfare remarks at the official welfare control (Anonymous 2010; 2011; 2012). Register variables were assessed and significant risk factors for violations of animal welfare legislation included increased variation (standard deviation) in milk yield for first and second lactation groups and abattoir remarks. This consistency could challenge the argumentation that the mere insurance of minimum standards as performed in the resource-based welfare inspection is not reflecting animal welfare to an optimal extent. The outcome measure of violation in paper 4 is however an animal-based measure. Looking at the classification herds based among the 25% worst scores of the animal-based AWI from paper 3 (19 herds) and herds with violations in paper 4 there is an agreement in 44% of the cases, as 10 high AWI score herds also had violations. Improvement of the agreement might have been bigger if non-compliance with more resource-based measures had been included in the case definition, but this is only speculative. The chosen outcome definition expresses some of the more severe notions within the welfare inspections and none of the expected risk factors turned up with significant effects. Neither mortality nor herd size was significantly associated with either of the outcome definitions in paper 3 and 4, and the identified risk factors all point in the direction of management issues instead. Thus, reasons for animal welfare problems in herds cannot simply be detected by using register data; their application in this context should be restricted to aid as screening tools as other factors seem to play a more evident role in the occurrence of animal welfare impairment and neglect. Anneberg (2009) investigated risk factors for farmers being reported to the police for neglect of animal care. The risk factors fell into more sociological categories as they were problems within the family, of financial or psychological character.

Applicability

In conclusion, the risk assessment of animal welfare is very complex due to the multi-factorial causality including interactions, but nonetheless it has been proposed as a transparent approach to evaluate animal welfare (Paton *et al.,* 2013) and has also been advocated by EFSA (2012b). However, a number of essential points within this scheme are still posing problems. Collins (2012) discussed the major issues regarding the risk assessment of animal welfare based on direct animal observations.

Firstly, the assumption of independence between multiple welfare hazards (i.e. risk factors) can have great implications on the risk score due to possible over- or underestimation. This is indeed a challenge within welfare evaluation, e.g. illustrated by the consequence of impaired rising-behaviour in dairy cows. Hazards could be too small/short cubicles or hock lesions – two factors also strongly associated with each other. Having these conflicts at this early stage in the process of establishing a risk assessment makes this approach quite complicated. However, the issues of changes in intensity and duration of hazards and the uncertainty of included parameters can be dealt with statistically e.g. probability distributions and stochastic modelling (Martin *et al.,* 2007) and should be explored in future research.

References

Andreasen, S.N., Wemelsfelder, F., Sandøe, P. and Forkman, B. 2013. The correlation of Qualitative Behavior Assessments with Welfare Quality® protocol outcomes in on-farm welfare assessment of dairy cattle. Applied Animal Behaviour Science143(1): 9-17.

Anneberg, I. 2009. Landmænd under pres. Østjyske landmænds oplevelser af kriser og håndtering af sygdom og sundhed. En etnografisk undersøgelse. Aarhus University, Aarhus, Denmark.

Anonymous. 2010. Kontrol af dyrevelfærd 2008-2009. ©Fødevarestyrelsen, 1. Edition, Schjølin Tryk, Copenhagen, Denmark.

Anonymous. 2011. Kontrol af dyrevelfærd 2010. ©Fødevarestyrelsen, 1. Edition, GP-Tryk, Copenhagen, Denmark.

Anonymous. 2012. Kontrol af dyrevelfærd 2011. ©Fødevarestyrelsen, 1. Edition, GP-Tryk, Copenhagen, Denmark.

Baadsgaard, N.P., Enevoldsen, C., 1997. A potential approach to support animal welfare promotion in a Danish veterinary practice context. In: Proceedings of the Society for Veterinary Epidemiology and Preventive Medicine, Chester, April 9–11, pp. 108–119.

Bicalho, R. C., Vokey, F., Erb, H.N. and Guard, C.L. 2007. Visual locomotion scoring in the first seventy days in milk: Impact on pregnancy and survival. Journal of Dairy Science 90: 4586–4591.

Blokhuis, H.J., Jones, R.B., Geers, R., Miele, M. and Veissier, I. 2003. Measuring and monitoring animal welfare: Transparency in the food product quality chain. Animal Welfare 12: 445-455

Bonde, M., Toft, N., Thomsen, P.T. and. Sørensen, J.T. 2010. Evaluation of sensitivity and specificity of routine meat inspection of Danish slaughter pigs using Latent Class Analysis. Prev. Vet. Med. 94: 165 169.

Bonde, M., Boutreau, R., Bracke, M.B.M., Butterworth, A., Capdeville, J., De Rosa, G., Engel, B., Keeling, L.J., Perny, P., Rousing, T., Sørensen, J.T., Van Reenen, K. and Veissier, I. 2009. Final model of multicriterion evaluation of animal welfare. Deliverable 2.28, subtask 2.3.1.3, EU Food-CT-2004-506508.

Booth, C. J., Warnick, L.D. Gröhn, Y.T., Maizon, D.O., Guard, C.L. and Janssen, D. 2004. Effect of lameness on culling in dairy cows. Journal of Dairy Science 87: 4115–4122.

Botreau, R., Bonde, M., Butterworth, A., Perny, P., Bracke, M.B.M., Capdeville, J. and Veissier, I. 2007. Aggregation of measures to produce an overall assessment of animal welfare. Part 1: a review of existing methods. Animal 1(8): 1179-1187.

Bracke, M.B.M., Edwards, S., Engel, B., Guist, W.G and Algers, B. 2008. Expert opinion as 'validation' of risk assessment applied to calf welfare. Acta Agriculturae Scandinavica, Section A – Animal Science 50: 29.

Brenninkmeyer, C., S. Dippel, S. March, J. Brinkmann, C. Winckler,U. Knierim. 2007. Reliability of a subjective lameness scoring system for dairy cows. Anim. Welfare.16: 127-129.

Breuer, K., Hemsworth, P.H. and Coleman, G.J. 1997. The influence of handling on the behaviour and productivity of lactating heifers. Proceedings of the 31st International Congress of the International Society for Applied Ethology, ISAE, Prague; Czech Republic.

Breuer, K., Hemsworth, P.H., Barnett, J.L., Matthews, L.R. and Coleman, G.J. 2000. Behavioural response to humans and the productivity of commercial dairy cows. Applied Animal Behaviour Science 66: 273-288.

Burow, E., Thomsen, T., Rousing, T., Otten, N.D. and Sørensen, J.T. 2013b. Effect of grazing on the cow welfare of dairy herds evaluated by a multidimensional welfare index. Animal 7(5): 834-842.

Chapman, C. R. 1976. Measurement of Pain: Problems and issues. Advances in Pain Research and Therapy 1, 345–353.

Collins L.M. 2012. Welfare risk assessment: the benefits and pitfalls. Animal Welfare 21(S1): 73-79.

deVries, M., Bokkers, E.A.M., van Shaik, G., Engel, B., Dijkstra, T: and de Boer, I.J.M. 2013. Exploring the value of routinely collected herd data for estimating dairy cattle welfare. Submitted to Journal of dairy Science. In Assuring dairy cattle welfare. Towards efficient assessment and improvement. PhD thesis, Wageningen University, The Netherlands.

Dohoo, I.R., Martin, S.W. and Stryhn, H. 2009. Veterinary Epidemiologic Research. 2 ed. VER, Inc., Charlottetown, Canada.

EFSA Panel on Animal Health and Welfare (AHAW). 2012a. Statement on the use of animal-based measures to assess the welfare of animals. EFSA Journal 10(6): 2767-2796.

EFSA Panel on Animal Health and Welfare (AHAW). 2012b. Guidance on risk assessment for animal welfare. EFSA Journal 10: 2513.10.2903/j.efsa.2012.2513.

Espejo, L. A., Endres, M.I. and Salfer, J.A. 2006. Prevalence of lameness in high-producing Holstein cows housed in freestall barns in Minnesota. Journal of Dairy Science 89: 3052–3058.

Espetvedt, M.N., Wolff, C., Rintakoski, S., Lind, A. and Østerås, O. 2012. Completeness of metabolic disease recordings in Nordic national databases for dairy cows. Preventive Veterinary Medicine 105: 25-37.

Esslemont, R. J., and Kossaibati, M.A. 1997. Culling in 50 dairy herds in England. Veterinary Record 140: 36–39.

Gross, W.B. and Siegel, P.B. 1997. Adaptation of chickens to their handlers and experimental results. Avian Diseases 23: 708-714.

Greiner, M. and Gardner, I.A. 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Preventive Veterinary Medicine 45: 3-22.

Hatcher, L. 1994. A Step-by-Step Approach to Using the SAS® System for Factor Analysis and Structural Equation Modeling. Cary, NC: SAS Institute, Inc.

Herva, T., Peltoniemi, O. and Virtala, A.M. 2009. Validation of an Animal Needs Index for cattle using Test Theory. Animal Welfare 18: 417-425.

Hemsworth, P.H., Coleman, G.J, Barnett, J.L. and Jones, R.B. 1994. Behavioural responses of humans and the productivity of commercial broiler chickens. Applied Animal Behaviour Science 41: 101-114.

Hemsworth, P.H., Coleman, G.C., Cransberg, P.H. and Barnett, J.L. 1996. Human Factors and the Productivity and Welfare of Commercial Dairy Cows. Research Report on Dairy Research and Development Council Project, Attwood, Victoria, Australia.

Jensen, T.B., Kristensen, H.H. and Toft, N. 2012. Quantifying the impact of lameness on welfare and profitability of finisher pigs using expert opinion. Livestock Science, 149, 209-214.

Jordan, K., Porcheret, M. and Croft, P. 2004. Quality of morbidity coding in general practice computerized medical records: a systematic review. Family Practice 21: 396-412.

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

Lawrenson, R., Williams, T. and Farmer, R. 1999. Clinical information for research; the use of general practice databases. Journal of Public Health Medicine 21(3): 299-304.

Lind, A., Thomsen, P.T., Ersbøll, A.K., Espetvedt, M.N., Wolff, C. and Rintakoski, S. 2012. Validation of Nordic dairy cattle disease recording databases – Completeness for locomotor disorders. Preventive Veterinary Medicine 107: 204-213.

MacCallum, R. C., Widaman, K. F., Zhang, S. and Hong S. 1999. Sample size in factor analysis. Psychological Methods 4: 84-99.

Main, D.C.J., Whay, H.R., Green, L.E. and Webster, A.J.F., 2003. Preliminary investigation into the use of expert opinion to compare the overall welfare of dairy cattle farms in different farm assurance schemes. Animal Welfare, 12, 565-569.

March, S., Brinkmann, J. and Winckler, C. 2007. Effect of training on the inter-observer reliability of lameness scoring in dairy cattle. Animal Welfare 16: 131-133.

Martin, P.A.J., Cameron, A.R. and Greiner, M. 2007. Demonstrating freedom from disease using multiple complex data sources 1: a new methodology based on scenario trees. *Preventive Veterinary Medicine* 79: 71-97.

McDowell, I. and Newell, C. 1987. Measuring Health: A Guide to Rating Scales and Questionnaires. Oxford University Press, New York.

Mörk, M.J., Wolff, C., Lindberg, A., Vägsholm, I and Egenvall, A. 2010. Validation of a national disease recording system for dairy cattle against veterinary practice records. Preventive Veterinary Medicine 93(2-3): 183-192.

Mörk, M.J., Lindberg, A., Alenius, S., Vagsholm, I and Egenvall, A. 2009b. Comparison between dairy cow disease incidence in data registered by farmers and in data from a disease-recording system based on veterinary reporting. Peventive Veterinary Medicine 88: 298-307.

Nyman, A.K., Lindberg, A. and Sandgren, C.H. 2011. Can pre-collected register data be used to identify dairy herds with good cattle welfare? Acta Agriculturae Scandinavica, Section A – Animal Science, 53 (S1): S8.

Paton, M.W., Martin, P.A.J. and Fisher A.D. 2013. Risk assessment principles in evaluation of animal welfare. Animal Welfare 22: 277-285.

Phythian, C.J., Michaelopoulou, E., Jones, P.H.,Winter, A.C., Clarkson, M.J., Stubbings, L.A., Grove-White, D., Cripps, P.J. and Duncan, J.S. 2011. Validating indicators of sheep welfare through a consensus of expert opinion. Animal, 5:6, 943-952.

Rajala-Schultz, P. J., and Gröhn, Y.T. 1999. Culling of dairy cows.Part I. Effects of diseases on culling in Finnish Ayrshire cows. Preventive Veterinary Medicine 41: 195–208.

Rodenburg, T.B., Tuyttens, F.A.M, de Reu, K., Herman, L., Zoons, J. and Sonck, B. 2008. Welfare assessment of laying hens in furnished cages and non-cage systems: assimilating expert opinion. Animal Welfare, 17, 355-361.

Rutherford, K. M. D., Langford, F.M., Jack, M.C., Sherwood, L., Lawrence, A.B. and Haskell, M.J. 2009. Lameness prevalence and risk factors in organic and non-organic dairy herds in the United Kingdom. Veterinary Journal 180: 95–105.

Sandgren, C.H., Lindberg, A. and Keeling, L.J. 2009. Using a national dairy database to identify herds with poor welfare. Animal Welfare 18: 523-532.

Sandøe, P. and Simonsen, H.B. 1992. Assessing animal Welfare: Where does Science end and Philosophy begin? Animal Welfare 1: 257-267.

Sharma, S., 1996. Applied Multivariate Techniques, John Wiley & Sons, Inc., New York, USA.

Scott, E.M., Nolan, A.M. and Fitzpatrick, J.L. 2001. Conceptual and methodological issues related to welfare assessment: A framework for measurement. Acta Agriculturae Scandinavica Section A- Animal Science. Suppl 30:5-10

Scott, E.M., Fitzpatrick, J.L., Nolan, A.M., Reid, J. and Wiseman, M.L. 2003. Evaluation of welfare state based on interpretation of multiple indices. Animal Welfare 12: 457-468.

Spoolder, H., de Rosa, G., Horning, B., Waiblinger, S. and Wemelsfelder, F. 2003. Integrating parameters to assess on-farm welfare. Animal Welfare 12: 529-534.

Thomsen, P.T. and Baadsgaard N.P. 2006. Intra- and inter-observer agreement of a protocol for clinical examination of dairy cows. Preventive Veterinary Medicine 75: 133-139.

Thomsen, P.T., Kjeldsen, A.M, Sørensen, J.T. and Houe, H. 2004. Mortality (including euthanasia) among Danish dairy cows (1990-2001). Preventive Veterinary Medicine 62(1): 19-33.

Waiblinger, S., Menke, C. and Coleman, G. 2002. The relationship between attitudes, personal characteristics and behavior of stockpeople and subsequent behavior and production of dairy cows. Applied Animal Behaviour Science 79: 195-219.

Webster, A.J.F., Main, D.C.J. and Whay, H.R. 2004.Welfare assessment: indices from clinical observations. Animal Welfare 13(S):93-98.

Wemelsfelder, F., Hunter, T.E.A. , Mendl, M.T. and Lawrence, A.B. 2001. Assessing the 'whole animal': a free choice profiling approach. Animal Behaviour 62: 209–220.

Whay, H.R., Main, D.C.J, Green, L.E. and Webster, A.J.F. 2003. Animal-based measures for the assessment of welfare state of dairy cattle, pigs and laying hens: consensus of expert opinion. Animal Welfare 12: 205-217.

Wolff, C., Epstevedt, M., Lind, A., Rintakoski, S., Egenvall, A., Lindberg, A. and Emanuelson, U. 2012. Completeness of the disease recording systems for dairy cows in Denmark, Finland, Norway and Sweden with special reference to clinical mastitis. Veterinary Research 8: 131-142.

Wright, J.G. and Feinstein, A.R. 1987. A comparative contrast of clinimetrics and psychometric methods for constructing indexes and rating scales. Journal of Clinical Epidemiology 45: 1201-1218.

Veissier, I., Jensen, K.K., Botreau, R. and Sandøe, P. 2011. Highlighting ethical decisions underlying the scoring of animal welfare in the Welfare Quality® scheme. Animal Welfare 20(1, S1): 89-101.

7. Conclusions, recommendations & perspectives

The present Ph.D. thesis reached the following main conclusions:

- Register data for mortality and bulk tank somatic cell count can potentially classify herds with lameness prevalences above a certain threshold.
- Misclassification bias/observer effect in cross-sectional studies can be alleviated by latent class modelling and be avoided to a great extent by implementing a correct study design.
- Aggregation of register data could not reveal the actual welfare state measured by resource- and animal-based measures.
- Register data can be used to evaluate risk factors for violations of animal welfare legislation.

The findings presented in the present thesis showed the possibility of using routinely collected secondary data to predict cross-sectional findings and their potential use in the identification of herds with animal welfare problems, e.g. high lameness prevalence (paper 1). Aggregating welfare measures into an overall index yielding a final welfare score used for the classification of herds holds certain validation issues. Prevalences of individual measures are highly influenced by the observer (paper 2). This problem could, however, be alleviated if the outcome of interest could be assessed on a dichotomous scale instead of the ordinal scale being favoured in welfare assessment protocols. Assessing animal welfare solely based on one type of measures (i.e. register-, resource-, or animal-based parameters) each only yields welfare estimates for certain aspects of welfare, thus making comparisons between welfare outcomes difficult and making it impossible to estimate the status of animal welfare without visiting the herd (paper 3). Nonetheless, register data could be used as a screening tool reducing the number of herds to be visited by the official welfare inspection (paper 4) if the definition of animal welfare is clearly defined. In order to obtain highest validity, the welfare definition should primarily consist of animal-based measures, as good management showed to be of great value in equalling out potential detrimental effects of housing deficits; and vice versa bad management could erode the potential benefits of a good system (paper 3).

Recommendations and Perspectives

Based on the encountered problems with the interpretation of aggregated index values register data variables should rather be assessed individually first and then be employed in a multi-variable prediction model, than implemented in an overall register data welfare index (paper 3).

In order to further promote and enhance animal welfare in dairy herds' further scientific results should aid the authorities in developing more optimal screening tools. The results from the present thesis could serve as an incitement to improve the risk-based screening tool if the highlighted challenges are further investigated. Initially this calls for a large-scale welfare assessment of a representative sample to establish the prevalence of animal welfare problems. This calls for a multi-disciplinary cooperation between all stakeholders. This cooperation should help to establish a clear case definition both regarding the opinion of people directly involved with the animals having a practical approach (e.g. farmers, production advisors, veterinarians), people with a theoretical approach (e.g. animal welfare scientists, ethicists/philosophers, policy makers) and people with an emotional approach to the animals (e.g. consumers, animal rights movements). Once a clear distinction between acceptable and unacceptable welfare is established an integrated welfare assessment like e.g. the Welfare Quality®, the DCF or the assessment protocol presented in this thesis could be performed in order to estimate the prevalence of dairy herds with unacceptable levels of welfare problems.

Modern Danish dairy production systems tend to become more uniform in regards to the barn conformation. However, big differences do exists in terms of resting areas. Bedding material varies highly between herds, i.e. straw, saw dust, sand or peat are all common materials. Additionally, summer grazing is present in both conventional and organic herds. Hence, in risk-based assessment schemes, a stratified sampling is necessary to avoid confounding effects and to enable the evaluation of the impact on animal welfare within these different subgroups of housing aspects. Further risk factor studies are needed to clarify the actual impact of these effects.

Finally, the development of a screening tool based on register data should explore the associations of these time series data for varying period lengths in order to find the most suitable time period to predict potential animal welfare problems. The present thesis only regarded data collected in the DCD, and further investigations could be made including data from more non-specific production data sources such as demographics, medical records and tax records. However, this approach would require quite some delicacy, as this type of surveillance would imply a massive intervention into the private sphere of livestock holders.

In practice, syndromic surveillance detecting aberrations from baseline levels of non-specific health data provides an interesting alternative. In syndromic surveillance historical data are used to form baseline levels and thresholds for aberrations from this result in an early warning. This concept should be investigated for the detection of animal welfare problems, as it would consider every single herd as its own control. The credibility of such a system would probably be higher ranked by farmers, since a warning would rely on a larger number of variables related specifically to the given herd rather than being a somewhat arbitrary identification based on national averages like the current risk-based inspection system.

Appendices:

Appendix A – register-based variables

Appendix B – Resource checklist

Appendix C – Clinical & behavioural scoring sheets

Table A. Variables extracted from the Danish Cattle Database used for welfare indicators

Data set	Key	Variable	DCD Label	Unit
Herd_info	Herd id	Cow years	Arsk0203/0304	Number of cow years
		Milk yield	Mlk_0203/0304	Kg ECM
		Fat	Fdt_0203/0304	Kg fat/year
		Fat %	Fdtp_0203/0304	Fat %/ year
		Protein	Prt_2003/0304	Kg protein/ year
		Protein %	Prtp0203/0304	Protein%/year
Herd_tankscc	Herd id	Herd id	K40CHRNR	CHR-number
		Control date	K40UDTDATO	Date
		Bilk tank somatic cell count	K40RESPRIMR	Cells/mL
Cow_ins	Cow id	Cow id	CKRDYRNR	CKR number
		Insemination date	Dato	Date
		Insemination number	Insnr	Count
Cow_turn	Cow id	Cow id	CKRDYRNR	CKR number
		Turn over date	Dato	Date
		Turn over	Omsatning	Slaughter/Dead/Sale
		Reason for turn over	Aarsag	Diagnosis (reproduction/locomotor disorder/mastitis etc.)
Cow_health	Cow id	Cow id	CKRDYRNR	CKR number
		Treatment date	Sygdomsdato	Date
		Disease code	Sygdom	Diagnostic code
		Treatment provider	Behandler	Veterinarian/ Farmer/Inseminator

Table A. continued

Data set	Key	Variable	DCD Label	Unit
Cow_cull	Cow id	Cow id	CKRDYRNR	CKR number
		Culling date	Dato	Date
		Reason for culling	Aarsag	Diagnosis (reproduction/locomotor disorder/mastitis etc.)
Cow_info	Cow id	Cow id	Cow id	CKR number
		Date of birth	Foedselsdato	Date
		Date of entry	Datofra	Date
		Date of exit	Datotil	Date
		Breed	Race	Breed code (1=Danish Red, 2= Danish Holstein, 3= Danish Jersey, 4= Red Holstein, 8= Mixed breed)
		Sex	Koen	Sex code (1= male, 2= female, 3=ET male, 4= ET female, 8=unknown sex)
		Calving status	fodttxt	Alive/Dead
Cow calving	Cow id	Cow id	CKRDYRNR	CKR number
		Calving number	Kaelvningsnr	Count
		Calving date	Kaelvedato	Date
		Calving process	klvforloeb	Unproblematic without help/Unproblematic with help/Difficult without veterinary assistance/Difficult with veterinary assistance/ Caesarean section

Table A. continued

Data set	Key	Variable	DCD Label	Unit
Cow_yield	Cow id	Cow id	CKRDYRNR	CKR number
		Annual Milk yield	Kgmaelk	Kg ECM
		Annual Fat content	Kgfedt	Kg Fat
		Annual Protein content	Kgprotein	Kg Protein
		Lactation number	Laktationsnr	Count
		Lactation start date	Datofra	Date
Cow_YKTR	Cow id	Cow id	CKRDYRNR	CKR number
		Control date	Kontroldato	Date
		Milk yield at control date	Kgmaelk	Kg
		Fat % at control date	Fedtpct	Fat %
		Protein % at control date	Proteinpct	Protein %
		Somatic cell count	Celletal	Cells/mL
Cow_abattoir	Cow id	Cow id	CKRDYRNR	CKR number
		Slaughter date	Slgdato	Date
		Carcase classification	Slgfedme	EUROP Beef Carcase Classification
		Slaughter code	Fundkode	Abattoir code for pathological finding

Resource check list

Farm: ID: Date: Observer:

Layout (sketch): pens/groups, feed bunk, lying areas, alleys etc., brushes, drinkers, concentrate feeders, outdoor run etc.

Parameter	**Start time**	**End time**
AD		
Laying down observations		
Clinical scoring		
Resources Protocol		

Farm: ID: Date: Observer:

General information

- **Number of cattle housed with the main herd**

'Animal category'	Number of animals
Lactating group…	
Lactating group…	
Lactating group…	
Dry cows	
Cows in calving/sickness pen	

Lactating cows

- **Horns?** ☐ all dehorned ☐ partly horned-number of animals:............... ☐ all horned
- **Are horn tips round(ed)?** ☐ yes ☐ no (=no sharp tip also in young animals)

Lying area

- ☐ freestall with cubicles ☐ deep litter system ☐ other ……………………………
- **Number of cubicles**: facing wall: .... head-to-head: .... facing passage: ....

or: length: ……….m depth: …..…..m of lying area in case of deep litter system

- **Cubicle type**: ☐ raised ☐ deep bedding ☐ raised with curb ☐ raised with deep bedding (choose main type)
- **Flooring in cubicle**: ☐ hard rubber mat ☐ soft rubber mat ☐ straw/manure mattress ☐ cow comfort mattress ☐ concrete ☐ wood ☐ sand ☐ soil
- **Bedding**: ☐ straw ☐ shredded straw ☐ sawdust ☐ wood shavings ☐ peat ☐ sand ☐ no bedding ☐ ....
- **Amount of bedding**: ☐ minimal ☐ sufficient ☐ amply

minimal: amount of bedding is very little and just to absorb moisture; ***amly:*** high bedding that gives cushioning

Farm: ID: Date: Observer:

- **Current state of lying area**: ☐ clean ☐ slightly dirty ☐ dirty
- **Hardness of lying area**: ☐ hard ☐ medium ☐ soft
- **Neck rail**: ☐ flexible ☐ solid
- **Type of cubicle partitions**: ☐ cantilever ☐ mushroom type ☐ Newton Rigg ☐ BK-Box ☐ DIY ☐ ....
- **Cubicle partitions:** ☐ flexible ☐ solid ☐ ....
- **Anti-mounting rail present:** ☐ yes ☐ no

- **<u>Measures/dimensions of free stalls</u>**

Measures in cm	**Head-to-head cubicles**			**Cubicles facing wall**		
*hardness of bed front**						
*hardness of bed rear**						
mattress intact (y/n)						
curb height						
level of lying area (raised and deep bedding cubicles)						
cubicle width						
cubicle length						
length of 'lying area'						
height of neck rail						
diagonal neck rail-curb						
height of cubicle division						
height of dividing board btw. rows (if present)						
height of front rail (if present)						

- ***(hard)** like concrete or wooden boards; you don' t drop on your knees voluntarily
- ***(medium)** like camping mat made from rubber foam; unpleasant when dropping with verve
- ***(soft)** like mattress / a 10 cm layer of sawdust; painless / soft landing

Farm: ID: Date: Observer:

Passages and flooring

- **Type of flooring (feed passage):** ☐ mastic asphalt ☐ concrete
  ☐ rubber ☐ other: …………………
  ☐ solid ☐ slatted elements ☐ other: …………………………
- **Feed passage:** length:............m width: ……… m
  gap width: ....cm slat width: ………cm
- **Slipperiness feed passage:**
  ☐ very slippery/no grip when braking, very easy spinning ☐ slippery/little grip when braking, spinning possible
  ☐ medium/rudimentary sliding and spinning ☐ good grip/sliding and spinning almost not possible
  ☐ coarse and abrasive surface/like sandpaper
- **Cleanliness of feed passage**: ☐ clean ☐ slightly dirty ☐ dirty
- **Type of flooring (main passages):** ☐ mastic asphalt ☐ concrete
  ☐ rubber ☐ other: …………………
  ☐ solid ☐ slatted elements ☐ other: …………………………
- **Main passage1**: length:............m width: ……… m
  gap width:..........cm, slat width………cm;
- **Main passage2:** length:............m width: ……… m
  gap width:..........cm, slat width………cm;
- **Main passage3:** length:............m width: ……… m
  gap width:..........cm, slat width………cm;
- **Slipperiness main passages (on average):**
  ☐ very slippery/no grip when braking, very easy spinning
  ☐ slippery/little grip when braking, spinning possible
  ☐ medium/rudimentary sliding and spinning
  ☐ good grip/sliding and spinning almost not possible
  ☐ coarse and abrasive surface/like sandpaper
- **Cleanliness of main passage**: ☐ clean ☐ slightly dirty ☐ dirty

Farm: ID: Date: Observer:

- **Connecting passage width**: 1)..........m 2)......... m 3)........... m 4)........... m
- **Number of dead ends** (<3m wide & >3m deep): ....
- **Built-in manure scraper:**
  guide of suspension: ☐ deepened ☐ hidden ☐ exposed ☐ ....
  inactive position: ☐ in the activity area ☐ out of the activity area

Water points

- **Type**: (1) trough (2) tip-over trough (3) bowl (4) bowl with reservoir (5) trough with balls (anti-frost)

type		**number**	**length**	**width**	**height (rim)**
( )		....	.... cm	.... cm	.... cm
)		....	.... cm	.... cm	.... cm
( )		....	.... cm	.... cm	.... cm
( )		....	.... cm	.... cm	.... cm

- **Mark position of water points in layout sketch!**
- **Is the water flow sufficient (20l/min → bucket test)?** ☐ yes ☐ partly ☐ no

  (Perform test just in case of bowls as drinker.)
- **Are drinkers clean?** ☐ clean ☐ slightly dirty ☐ dirty
  <u>dirty</u>: drinkers not cleaned, dirty & water dirty at moment of inspection
  <u>slightly dirty</u>: drinkers dirty (old dirt) - but water fresh and clean at moment of inspection OR only part of several drinkers clean, with clean water
  <u>clean</u>: drinkers clean & water clean at moment of inspection (some amount of fresh food acceptable)

Feeding place

- **Number of feeding places: ....** **Total length of feeding place: ....** m (in case of neck rail)
- **Type of feed rail**: ☐ standard headlockers ☐ headlocker tombstones ☐ tomb-stones ☐ solid neck rail ☐ flexible neck rail ☐...
- **Feed rail height:** ....cm

Farm: ID: Date: Observer:

- **Feeding stalls:** height..........cm depth ………cm (remember Trenthorst)
- **Feeding stall partitions** present: ☐ yes; width …………cm ☐ no
- **Level of feed trough**: ....cm (above level of feed alley)
- **Inclination of feed rail**: ☐ yes ☐ no
- **Is the feed bunk clean?** ☐ clean ☐ slightly dirty ☐ dirty
  dirty: old feed, dung, dirt, stones lying on parts of feed bunk; no regular cleaning of trough
  slightly dirty: no regular cleaning, but feed trough found to be clean; clean: trough cleaned each time before feeding;no dirt and old food rests on feeding table
- **Is feed available and seems to be sufficient until next feeding time (take into account feeding time given by the farmer)?** ☐ yes ☐ partly ☐ no
- **Is the food of same quality throughout the feeding places?** ☐ yes ☐ no
- **Are there different basic feed components fed?** ☐ yes ☐ no
- **Do you feed a total mixed ration (TMR)?** ☐ yes ☐ no
- **Do you feed different feed at different places?** ☐ yes ☐ no
- **Do you distribute different feed throughout all feeding places?** ☐ yes ☐ no
- **Does the feed contain enough roughage (hay, good silage in ration or extra available; not cut too shortly in case of TMR)?** ☐ yes ☐ partly ☐ no

Concentrate feeder

- **Number of feeders**:...............
- **Protection in feeder**: ☐ complete ☐ sides ☐ head area ☐ ....

Light conditions

- **Subjective assessment of the light conditions**: ☐ dark ☐ medium ☐ bright
- **Weather**: ☐ sunny ☐ partly cloudy ☐ cloudy ☐ rainfall ☐ snowfall

Ventilation

- **Barn climate:** ☐ cold barn ☐ warm barn ☐ ……………
- **Ventilation system**: ☐ eaves-ridge-ventilation ☐ one side open ☐ cucettes ☐ forced ventilation ☐ ……………….......................

Farm: ID: Date: Observer:

Other barn equipment

- **brushes:** ☐ yes ☐ no
- **brush type** ☐ fixed (broom): number:.......... ☐ automatic (rotating): number ....
- **salt blocks**: ☐ yes ☐ no
- **hay rack**: ☐ yes ☐ no ☐ hay at feeding rack
- **claw trimming facility (crush):** ☐ yes ☐ no **others**: ....

Malfunctioning equipment

- **Feed rail** ☐ yes .... ☐ no
- **Concentrate feeder** ☐ yes .... ☐ no
- **Water points** ☐ yes .... ☐ no
- **Brushes** ☐ yes .... ☐ no
- **Cubicle partitions/stalls** ☐ yes .... ☐ no
- **Gates** ☐ yes .... ☐ no
- **Manure scraper** ☐ yes .... ☐ no
- **Other**...

Injuries-causing equipment

- **Feed rail** ☐ yes .... ☐ no
- **Concentrate feeder** ☐ yes .... ☐ no
- **Water points** ☐ yes .... ☐ no
- **Brushes** ☐ yes .... ☐ no
- **Cubicle partitions/stalls** ☐ yes .... ☐ no
- **Gates** ☐ yes .... ☐ no
- **Manure scraper** ☐ yes .... ☐ no
- **Other**...

Farm: ID: Date: Observer:

Calving pen/sick pen

- **Calving pen available?** ☐ yes, single pen: ……… (number)
  ☐ yes; group pen: …….. (number)
  ☐ no
- **Calving pen size?** ....x…………m ....x…………m ....x…………m
- **Type of bedding?** ☐ straw ☐ other: …………………………………
- **Is the calving environment appropriate?**
  calving pen ☐ yes ☐ no
  barn/herd ☐ yes ☐ no

  yes: free movement possible; easy and save rising and lying down – soft, non-slippery floor, clean, control => no tie-stall; calving box with enough space (>9m² in single box, >6² in group box), litter deep enough; visual contact to other cows possible;
- **Do empty calving pens look cleaned?** ☐ yes ☐ no
- **Is there the possibility to separate sick animals from the main herd?**
  ☐ yes, sick pen ☐ yes, tie stall ☐ yes, calving pen ☐ no
  Sick pen available (additionally)? ☐ yes , ...x…………m ☐ no
- **Type of bedding?** ☐ straw ☐ other:
  …………………………………
- **Is the hospital pen in use?** ☐ yes ☐ no
  if yes: Is the pen clean and dry? ☐ yes ☐ medium ☐ yes

Do empty hospital pens look cleaned? ☐ yes ☐ no

Outdoor loafing area (OLA) (current situation)

- **Do cows have the possibility to get directly in touch with the weather?** ☐ yes ☐ no
- **OLA available?** ☐ yes, area: approx.. …………m² → *add to sketch* ☐ no
- **Surface of OLA** ☐ concrete ☐ mastic asphalt ☐ soil
  ☐ „green“ paddock ☐ wood chips ☐ ……………
- **Cleanliness of outdoor loafing area**: ☐ clean ☐ slightly dirty ☐ dirty

Farm: ID: Date: Observer:

- **Number of permanently open doorways** ....
- **Doorway widths**: ....m .... m .... m
- **Furnishings**: ☐ hay rack ☐ brush ☐ water point ☐ salt blocks ☐ concentrate feeder ☐ ....
- **Wind protection?** ☐ yes ☐ no ☐ solid
- **Roofing**: approx.. ....%

<u>Dry cows</u>

- **Housing system**: ☐ like lactating animals ☐ free-stalls ☐ deep litter systems ☐ tie stalls
- **Flooring**: ☐ solid ☐ slatted ☐ mixed ☐ straw

<u>Further information</u>

- **Conspicuously aggressive animals present in the herd?** ☐ yes ☐ no
- **Are obviously sick animals observed in the herd which should be kept in a hospital pen?** ☐ yes ☐ no

(animals with severe disease having difficulties to get access to food etc., e.g. high-grade lame animals, very weak animals)

- **Are obviously sick animals observed in the herd which should be euthanized (§ 20)?** ☐ yes ☐ no

Milking equipment

- Machine type: ☐ tandem ☐ auto-tandem ☐ trigon ☐ heringbone ☐ side-by-side ☐ rotary ☐ robot milking
- Number of aggregates in use during milking: ....

Recording sheets for clinical and behavioural observations (Avoidance distance, clinical/behavioural scoring, lying down)

Date:________ Observer:____________________ Page:---- ----

Farm:____________________

AD

	group /pen	collar no.	ear tag no.	test 1	test 2 (retest)	remarks
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						
30						
31						

AD = avoidance distance within herd, distance between hand and nearest part of head (nose, muzzle...)

Clinical scoring sheet

2= wound/swelling
Integument: **Location:** **N**=neck **S**= shoulder **B**= back **TC**= Tuber coxae **TI**= Tuber ischiaticum **TB**=tail basis **T**= thigh

No	Chr-Ckr	Cleanliness			Claw	Hair	Integument				Getup	Faec.	BCS	Lame
		leg	hind	udder			carpus	tarsus	body	location				
1 L		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
2 R		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
3 L		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
4 R		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
5 L		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
6 R		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
7 L		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
8 R		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
9 L		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
10 R		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
11 L		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
12 R		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
13 L		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
14 R		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
15 L		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
16 R		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
17 L		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2
18 R		1 2	1 2	1 2	1	1 2 s	1 2	1 2	1 2					1 2

CHR:________________ Dato:____________---- Observatør:____________ Side:----

Lying down/ Lægge-sig

	gruppe	Chr	CKR/konr.	varighed (sek)	kollisioner (antal)	kommentar
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						
30						
31						

Time starts when the carpus is bending and stops when the front leg(s) are stretched out/stopped to bring into position

Collidation=collidation with cubicle equipment during lying down (while the time is taken)
Sounds=if collidation is acustically heard and clearly to connect with the observed lying down situation (while the time is taken)