Management of Data for Health Performance Measurement in the Dairy Herd

Ph.D. thesis
Mogens Agerbo Krogh
2012

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

Ph.D. thesis

Mogens Agerbo Krogh
University of Copenhagen,
Faculty of Health and Medical Sciences,
Department of Large Animal Sciences,
Grønnegaardsvej 2,
DK-1870 Frederiksberg, Denmark

Future address:
Kvægdyrlægerne Midt,
Klochsvej 2b,
DK-7441 Bording.
mok@raskedyr.dk

Principal Supervisor:
Professor Carsten Enevoldsen
University of Copenhagen,
Faculty of Health and Medical Sciences,
Department of Large Animal Sciences,
Grønnegaardsvej 2,
DK-1870 Frederiksberg, Denmark

Assessment Committee:
Chairman
Professor Anders Ringgaard Kristensen
University of Copenhagen,
Faculty of Health and Medical Sciences,
Department of Large Animal Sciences,
Grønnegaardsvej 2,
DK-1870 Frederiksberg, Denmark

Senior Scientist Søren Østergaard
University of Aarhus,
Faculty of Agricultural Sciences,
Research Centre Foulum,
Department of Animal Science,
Blichers Allé 20, Postboks 50,
DK-8830 Tjele, Denmark

Professor Henri Seegers
Centre de recherches INRA,
Rue de la Géraudière,
BP 71627,
F 44316 Nantes cedex 3,
France

Preface

The work presented in this Ph.D. thesis was conducted at the Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen in the period from October 2005 to April 2012. The project (2.5 yr) was funded by the Faculty.

Quite obviously, it has been a long journey. Also too long. Nevertheless, I believe that the work presented in the thesis and the insights I have gained from the project have matured and developed into knowledge that can be very valuable for the dairy industry. To me, dissemination and implementation of the thoughts is the ultimate judge of this work.

I express my sincere thanks to my supervisor, Professor Carsten Enevoldsen for always being supportive and never ever giving up on me.

I also thank all my fellow Ph.D. students and senior colleagues at the faculty for the opportunity to be a part of a research environment for almost 2 years. Also, I acknowledge all the more (or less) anarchistic practicing veterinarians out there who have contributed with data and always have shown great interest in my work.

Finally, I would like to thank my wife, Hanne and my sons Emil and Alfred for almost never complaining that I, at times, was far more interested in my computer than in them.

Silkeborg, April 2012

Mogens A. Krogh

Summary

This thesis deals with health-related data from dairy herds; how they are created, collected, managed, and evaluated for the purpose of health performance measurement. The overall objective is to suggest a coherent concept for management of data for health performance measurement that is suitable and sufficient for the diverse contexts of industrialized Danish dairy herds and associated veterinary practices.

Chapter 1 is a general introduction to the research area with specification of the objectives. Chapter 2 is a brief description of the context of the Ph.D. project, the framework for health-related data and their collection, and the applied analytical methods. Chapter 3 initially presents summaries of five manuscripts also included in chapter 3. These five manuscripts are the core of the work performed in the Ph.D. project. Chapter 4 is a brief discussion of general problems related to implementation in veterinary practice of the concepts and tools identified in chapter 3. Chapter 5 is an overall conclusion. Chapter 6 outlines the future possibilities for implementation of the concepts and tools and future research perspectives. Essential terminology is explained in an appendix.

The purpose of "Identification of principles and tools for management of health performance data from the industrialized dairy herd" was to identify principles and tools for analysis of herd health data in industrialized dairy herds. A further goal was to include the additional complexity that arises when human behavior and associated changes, including legislative changes, must be taken into account. The results were summarized into a concrete 7-step plan of action, which I consider to be adequate.

The purpose of the manuscript, "A tool to detect rater-introduced bias in clinical ratings", was to develop a screening tool that can identify errors arising from clinical ratings. The method can be applied in larger veterinary practices to perform quality assurance of ratings of clinical conditions. The tool was applied to body condition scores, which is a widely used scoring system with well-defined categories. Despite the standardization efforts, misclassifications were revealed.

The purpose of the manuscript, "Latent class evaluation of a milk test, a urine test, and the fat-to-protein percentage ratio in milk to diagnose ketosis in dairy cows", was to demonstrate a Latent Class Model in which one can evaluate a diagnostic test without having a perfect test (a gold standard). The model and the principles behind it have considerable potential not only to evaluate a diagnostic test's performance but also to rank the performance of diagnostic tests in relation to each other. The results show that KetoStix (Bayer Diagnostics Europe Ltd., Dublin, Ireland) has the best sensitivity and good specificity. The fat-to-protein ratio in milk has the same sensitivity as KetoLac (Sanwa Kagaku Kenkyusho Co. Ltd, Nagoya, Japan) but lower specificity.

The purpose of "A framework for integration of benchmarking and within-herd analysis in dairy herd management – analysis of lactation curves as a case" was to demonstrate an initial analysis of

a complex health and production indicator like the lactation curve. The lactation curve analysis is complex because the relevant information cannot be expressed by one figure, and information about several components of the lactation curve can be used for several purposes at several organizational levels. Results showed that both the within- and the across-herd models gave similar and valid estimates of cow and herd levels. In addition, estimates of the lactation curve were prepared for use in benchmarking at the herd level. The options for inclusion of additional explanatory variables in the model were demonstrated with inclusion of age at first calving. Herd-specific estimates regarding the relationship between explanatory variables and milk production were robust and showed variability in herd-level coefficients, which illustrates the need for herd-specific multivariable and multilevel analyses in herd management.

The purpose of the "Evaluation of effects of disease control in a complex dairy herd health management program" was to evaluate the effect of a systematic examination for metritis as part of a larger health management program that involves simultaneously having to take into account that basic recordings changed at the beginning of the health management program. The results showed that the negative effect of metritis on milk yield was reduced by 17% for first-calf cows. For second-parity and older cows, the health management program contributed 129 kg and 80 kg Energy Corrected Milk, respectively. There were indications that the effect of the health management program was mediated through the metritis examinations.

Sammendrag (Danish summary)

Denne afhandling omhandler håndteringen af data til præstationsmåling i sundhedsstyringen i malkekvægsbesætninger. Præstationsmåling skal forstå som hele den proces, der foregår fra datafødsel, dataopsamling og databehandling til evaluering af resultater. Afhandlingen præsenterer et logisk sammenhængende koncept til håndtering af sundhedsrelaterede data, der er praktisk anvendeligt og tilstrækkeligt til at beskrive de forskellige forhold, der gør sig gældende i danske malkekvægsbesætninger og tilhørende dyrlægepraksis.

Kapitel 1 er en generel introduktion til fagområdet med specifikation af målene for ph.d.-projektet. Kapitel 2 er en kort beskrivelse af den sammenhæng, hvori ph.d.- projektet skal ses, rammerne for sundhedsdata og opsamling heraf samt de anvendte analytiske metoder. Kapitel 3 præsenterer først af alle de delkonklusioner, der er resultatet af de 5 manuskripter, der følger i 5 delafsnit. De 5 manuskripter er kernen i det udførte arbejde. Kapitel 4 er en diskussion af de problemstillinger, der er tilvejebragt i kapitel 3. Kapitel 5 er en samlet hovedkonklusion, hvori jeg foreslår et sammenhængende koncept. I kapitel 6 bliver fremtidige muligheder for implementering af værktøjerne og fremtidige forskningsperspektiver beskrevet. Til sidst er der et kort appendiks, hvor central terminologi er forklaret.

Formålet med “Identification of principles and tools for management of health performance data from the industrialized dairy herd” var at identificere principper og værktøjer til analyse af sundhedsdata i industrialiserede malkekvægsbesætninger. Derudover at forholde sig til den yderligere kompleksitet, der opstår, når man skal tage hensyn til menneskelig adfærd og ændringer i disse, herunder lovgivningsmæssige ændringer. Resultaterne er sammenfattet i en 7-trins handlingsplan, som anses for at være dækkende.

Formålet med ”A tool to detect rater-introduced bias in clinical ratings” var at udvikle en analytisk metode, hvormed man kan identificere fejl, der opstår i forbindelse med kliniske scoresystemer. Metoden har anvendelsesmuligheder i større dyrlægepraksis, hvor man ønsker at lave kvalitetssikring af kliniske registreringer. Resultatet var, at der på trods af standardiserings-bestræbelser kunne være systematiske fejlklassificeringer i dyrlægers anvendelse af huldscore ved andenkalvs og ældre køer.

Formålet med ”Latent class evaluation of a milk test, a urine test, and the fat-to-protein percentage ratio in milk to diagnose ketosis in dairy cows” var at demonstrere en latentklassemodel, hvormed man kan evaluere en diagnostisk test uden at have en perfekt test at sammenligne med. Modellen og principperne bag har betydeligt potentiale for ikke blot at vurdere en diagnostisk tests præstation, men også at kunne rangere forskellige diagnostiske test i forhold til hinanden. Resultaterne viser at KetoStix (Bayer Diagnostics Europe Ltd., Dublin, Ireland) har den bedste sensitivitet og bedste specificitet. Fedt-Protein forholdet i mælk har samme sensitivitet som KetoLac (Sanwa Kagaku Kenkyusho co. Ltd, Nagoya, Japan) men lavere specificitet.

Formålet med "A framework for integration of benchmarking and within-herd analysis in dairy herd management – analysis of lactation curves as a case" var at demonstrere en indledende analyse af et komplekst sundheds- og produktionsmål som laktationskurven. Kompleksiteten består i, at laktationskurven ikke kan udtrykkes ved hjælp af ét tal, og at information om adskillige komponenter af kurveforløbet kan bruges til flere formål. Resultaterne viste, at både modellen inden for og på tværs af besætninger gav ensartede og gyldige estimater på ko- og besætningsniveau. Derudover at estimaterne af laktationskurven er brugbare til benchmarking af besætninger. Inklusion af yderligere variabler i modellen til forklaring af mælkeproduktion gav robuste estimater. Disse varierede fra besætning til besætning, hvilket illustrerer behovet for besætningsspecifikke multivariable og hierarkiske statistiske modeller af laktationskurvens form.

Formålet med "Evaluation of effects of disease control in a complex dairy herd health management program" var at evaluere virkningen af systematisk undersøgelse for børbetændelse som en del af et større sundhedsstyringsprogram, når man samtidigt skal tage højde for, at grundlæggende registreringer ændrer sig i forbindelse med opstart af sundhedsstyringsprogrammet. Resultaterne viste, at den negative virkning af behandlede tilfælde af børbetændelse på mælkeydelsen blev reduceret med 17% for førstekalvskøer. For andenkalvs og ældre køer bidrog sundhedsstyringsprogrammet med henholdsvis 129 kg og 80 kg EKM. Der er tegn på, at effekten af sundhedsstyringsprogrammet var medieret igennem undersøgelserne for børbetændelse.

Table of contents

1 General introduction

In medicine, an essential task has always been to evaluate the disease and health history of individual patients (anamnesis) as a component of the diagnosis. Subsequently, it is essential to evaluate whether the progression of various disease symptoms is satisfactory after some therapeutic intervention. The same principle is applicable to populations. Some awareness of development of disease occurrence in populations probably arose centuries ago, as indicated by the use of quarantine principles to protect the population from contagious disease as far back as Roman times (ref. Schwabe et al. 1977, p. 35). Numerical methods obviously are needed to deal effectively with occurrence and spread of disease or other events in populations. As an example, John Snow's studies of cholera epidemics in London in the mid-1800s demonstrated a systematic approach to studying development of disease occurrence in a population (ref. Schwabe et al. 1977, pp. 7–8). Without knowing the etiology of cholera, he identified how cholera was transferred (sewage outflow to drinking water) by means of comparing death rates between city districts and water supply companies during epidemics. These epidemiological principles have now become sophisticated and applied to many types of diseases, not only contagious disease.

As animal production became more and more intensified and herd size grew, it was obvious to apply the same principles to identify, prevent, or eradicate disease within animal herds. In Denmark, efforts to eradicate diseases like tuberculosis and brucellosis in dairy herds were greatly facilitated by the organizational structure around cooperative milk-processing plants and access to milk for diagnostic testing. Artificial insemination and systematic recording of milk yield in individual cows were introduced at a large scale in the mid-1900s. These technologies were mainly introduced for breeding purposes, but their usefulness for management support was soon realized. Enevoldsen (1993) reviewed the technologies and management tools developed for use in dairy herd health management during the 1900s. Over the last two decades, computer technology, automatic milking systems (AMS), and other automated data collection tools have increased the amount of available data further.

The production process in a herd and the need to evaluate performance are basically the same as in any other processes in manufacturing or service-providing organizations. Therefore, the manager of a dairy herd can apply similar techniques as the managers in other types of enterprises, businesses, or organizations. The (business or herd) manager will be responsible for establishing systems to evaluate the health-related performance of the production process continuously, exactly like using continuous collection of health data to evaluate the health state of individual patients in medicine. With the growing herd size in the dairy industry, the number of hired personnel also has increased. Consequently, management of human resources is becoming important. In addition, the detailed legal regulation of dairy herds in Denmark requires provision of documentation to the public veterinary authorities. This requirement raises performance measurement issues that are similar to performance measurement and management in the public sector (so-called New Public Management).

The legal regulations regarding health-related issues in dairy herds give the herd veterinarian a number of duties with respect to health performance measurement (documentation) based on recordings in the herd and recordings of the veterinarian's activities. Because veterinarians are getting organized in larger networks or companies, they also experience the need to become deeply involved in establishing efficient systems for performance measurement in the clients' herds and in their own practice organizations.

In veterinary medicine and herd management science, terms like 'surveillance', 'monitoring', 'benchmarking', and 'control' are widely used. In public management, essential terms are 'monitoring', 'performance management', and 'evaluation'. In this work, I use the term *performance measurement* because the combination of the words 'measurement' and 'performance' directly signals an evaluation of the current state of the system. That is, it signals an evaluation of how the process of interest is functioning at any time, which does not make sense without some criteria for distinguishing between acceptable or unacceptable. This approach is in contrast to systems that merely present collections of recordings without any attempt to evaluate. Because the term 'health' cannot be measured explicitly (it is a latent or unobservable trait), measurement of health performance inevitably requires a number of indicators (variables).

In addition to the general organizational issues described above, the current context for Danish veterinarians working with larger industrialized dairy herds is characterized by the following:

- The number of cows in the herds will typically be from around 50 up to around 1000.
- The owner of the herd will often be the manager (smaller herd), but in larger herds, a hired manager may be the key decision-maker.
- The degree of automatic data collection (AMS, activity measurement, etc.) is quite variable.
- Complex movement of animals occurs between herds with the same owner because of environmental regulations about harmony between number of animals and farm land. Replacement heifers brought up on separate herds and bull calves sold to special meat production herds are common.
- There is a high degree of legal regulation, not only related to herd size but also to "negative" performance measurement like mortality and use of antibiotics that may lead to "penalties" in terms of intensified public supervision. Avoidance of public control is a strong motivating factor for some farmers.
- Attitudes towards veterinary services are highly variable among dairy producers (Kristensen and Enevoldsen, 2008).
- There is a need to use scores (ratings) to describe signs of disease like lameness or poor body condition. The quality of these recordings can be very problematic (Lastein et al. 2009).
- Farmer attitudes and values can be important determinants of health and health promotion initiatives, but these attitudes may change (Andersen & Enevoldsen, 2004).

Consequently, the needs for health performance measurements may differ from herd to herd and from veterinarian to veterinarian because of variation in contexts. Context probably will be particularly important for problems involving human activity (e.g., measurement and evaluation of 'heat detection activities' or 'prudent usage of antibiotics'). In contrast, a purely technical measurement of electrical conductivity in quarter milk in a given AMS will probably be virtually context free. The consequence of a problem being context dependent is that it is unlikely that we can find a measurement system that is valid across herds and practices.

The data structure of health-related data from dairy herds is quite complicated. Numerous indicators are needed for sufficient health performance measurement in a dairy herd because we need to be able to diagnose numerous diseases or signs of ill-health that are measured at time intervals from milliseconds to years and at udder quarter level to the herd or veterinary practice level. Consequently, each herd and each practice must have measurement concepts that can adapt to each specific context and organizational level. Overall, one fixed type is very unlikely to fit all; at the least, a claim that one fixed type does fit all would require extraordinary supportive evidence.

The overall objective of this thesis is to

- suggest a coherent concept for management of data for health performance measurement that is suitable and sufficient for the diverse contexts of industrialized Danish dairy herds and associated veterinary practices.

The specific objectives are to:

- Identify principles and tools for management of data for health performance measurement in industrialized Danish dairy herds. This part includes demonstration of key tools for time series analysis applicable to the Danish contexts, and it addresses the complexity that arises when attempts are made to measure human activities; in particular, when legal regulations are imposed. The outcome is a coherent approach for management of health performance measurement.
- Demonstrate a tool to detect the bias that may be associated with personal (clinical) judgments of the health state of dairy cows (ratings). This tool will be particularly useful for a larger network of veterinarians who wish to explore the quality of the clinical work and evaluate whether a given type of rating is useful for benchmarking or across-herd statistical analyses. The outcome demonstrates the need to be close to the research object.
- Demonstrate an application of latent class analysis to evaluating diagnostic tests while accounting for the very frequently occurring situation when diagnostic tests are imperfect. This tool may be quite useful for evaluation of the quality of the diagnostic work in the herds. The outcome also demonstrates a useful approach to clinical diseases that cannot be defined by one diagnostic test alone.
- Demonstrate an approach to preparing an example of a complex health-production measurement like the lactation curve for use in benchmarking. Each component and various aggregations of the lactation curve can be used for both benchmarking across herds and for time series analysis within the herd. The uncertainties of each component and correlations

between components at the cow and herd levels are estimated. Inclusion of an explanatory variable is described to demonstrate the potentials for causal analyses with the multilevel lactation curve analysis.

- Evaluate the possible effects of the introduction of intensified clinical examinations of individual cows in the dairy herd when the examination routine is part of a general herd health management program (HHMP). The study demonstrates the changes in recording routines that inevitably take place when a HHMP is introduced causing complexity that makes evaluation of causal effects complicated.

The thesis is organized as follows:
Chapter 2 is a brief description of the context of the Ph.D. project, the data collection framework, and the applied analytical methods.

Chapter 3 gives a summary of the results obtained in five studies that address each of the specific objectives. Subsequently, the five specific objectives are addressed in five separate manuscripts with the following titles:

- Identification of principles and tools for management of health performance data from the industrialized dairy herd
- A tool to detect rater-introduced bias in clinical ratings
- Latent class evaluation of a milk test, a urine test, and the fat-to-protein percentage ratio in milk to diagnose ketosis in dairy cows
- A framework for integration of benchmarking and within-herd analysis in dairy herd management – analysis of lactation curves as a case
- Evaluation of effects of disease control in a complex dairy herd health management program

In chapter 4, I discuss the general problems identified in chapter 3. I also suggest options for a coherent combination of the tools presented in chapter 3.

Chapter 5 provides a summary of the conclusions derived in chapters 3 and 4, and I suggest options for implementation of the principles and tools that are developed in the thesis for practical application in dairy herds and in veterinary practice.

In chapter 6, I address the most important needs for future development of dairy herd health management tools and concepts.

2 Study context, data collection framework, and analytic methods

2.1 Study context

During the last half century, the Danish dairy sector has undergone a tremendous structural development. Since the early 1960s, the number of herds with cattle has been approximately halved for every decade. Around the year 2000, there were about 10,000 milk-producing herds with a herd average of 60 cows. The Danish Cattle Federation (2012) stated that by the end of 2011, there were around 3,900 dairy herds with an average herd size of 148 dairy cows. One consequence of these larger herds is that the number of people working in the herd (apart from the family) also increases. This increase implies that it may be more difficult to identify the decision-maker for specific areas (like newborn calves) or that there will be a separation between the decision-maker and the persons who are implementing the decisions. Around the millennium, there were two trends in Danish dairy production. One was working with a low-cost, uneducated workforce (typically non-Danish) and the other was in a high-tech direction with sophisticated technologies like AMS and automatic feeding systems with a demand for a (smaller) educated workforce. Ten years after the turn of this century, AMS has been installed in approximately 25% of the herds (VFL, 2012); thus, the high-tech trend appears to have gained popularity.

The following brief review of the development in veterinary practice is primarily based on my knowledge about activities and debates in the veterinary organizations and my own employment and involvement in organizational work. During the last 3–4 decades (and because of the structural development in the dairy herds?), the veterinary practices developed from units with typically one to three veterinarians covering all species to practices with up to 20 veterinarians working only with cattle production. Other veterinarians working intensively with cattle organized themselves into corporate structures where they share brand, business development, and continuous education activities.

The role or work area for the veterinarians working in dairy cattle practice has also changed. In particular, after the first law concerning herd health management in 1995, we have seen the following major changes:

- Diagnosis and treatment of individuals and focus on individual animals has decreased.
- Consumer concerns about food safety, welfare, and microbial resistance play a much more important role not only for the producer but also for the work of the veterinarians. This prominence was especially emphasized in the Danish 2009 herd health legislation, in which the role of the veterinarian changed because public duties were placed on the veterinarians concerning auditing of animal welfare legislation (Danish Veterinary and Food Administration 2011, pp: 4–7). This legislation to some degree changed the relationship

between veterinarian and herd manager from the veterinarians being a consultant of herd health (collaborator) to having some degree of public supervision (authority).

In Denmark, control (or eradication) of contagious diseases in the cattle population has a long tradition. Denmark was declared free of Infectious Bovine Rhinotracheitis in 1992; in 1996, the eradication program for Bovine Virus Diarrhea was turned into a surveillance program. Additional programs against paratuberculosis and *Salmonella dublin* were initiated in 2006 and 2007, respectively. What is common among these programs is that they were initiated by the farmers' organization as voluntary. After some period of time, they were incorporated into national legislation. The requirements in the legislation have then been continuously strengthened until not complying with the program can make it difficult and too expensive to be a dairy producer.

Compared to virtually all other countries, Danish dairy production has been highly regulated concerning the use of antimicrobials. Any use of antibiotics requires a prescription in Denmark. Until 1995, all antibiotic treatments of cattle were to be done by a veterinarian. In 1995, the first national legislation concerning a herd health program was implemented (voluntary). In essence, the herd was to have monthly visits, and the farmer was given access to follow-up antibiotic treatment of adult cattle. Diagnosis and the first treatment were still to be made by a veterinarian. For calves, the legislation was more liberal. Around 2002, the Danish public veterinary authorities wanted to explore the consequences of lifting the very strict regulation of dairy farmers' access to antibiotics that until then required veterinary presence for a prescription. As a pilot study initiated in 2004, farmers who entered a specific herd health management program (see below) were permitted to initiate treatment with antibiotics without the presence of a veterinarian for a limited number of diagnoses. The link to the herd health management program was to ensure sufficient health surveillance and prudent use of antibiotics. Within the pilot project period, requirements were very strict for delivery of registrations from both farmers and veterinarians to a central database. Hill (2005) evaluated the effects of this pilot project, which was subsequently incorporated into national legislation. In 2006, the link between more liberal access to antibiotics and the herd health management program was formalized in a new law concerning herd health. In 2009, the legislation was further liberalized so that farmers could treat almost any disease without veterinary assistance. In addition, the very strict requirements for registration of disease treatments and results from veterinary examinations were abandoned. However, this liberalization was not without constraints. The veterinary authorities imposed thresholds concerning the amount of antibiotics used in the production and rates of dead cows and calves. These thresholds are used for classifying herds and affected the mandatory number of the herd veterinarians' visits and the risk of control visits by the veterinary authorities.

In the late 1990s, an Israeli-inspired herd health management system (Enevoldsen, 1997b; Nir - Markusfeld, 2003) came into some use in Danish dairy production. This Danish version of the Israeli herd health system was based on regular clinical examinations of groups of cows with a particularly high risk of contracting disease. The results of these examinations were collected in a database, and a performance report was developed to use these recordings to evaluate health in the

herds (Jensen et al., 1997). In essence, this herd health program formed the foundation for the pilot project concerning herd health and the 2006 changes in national legislation.

The context of this thesis is within the area of herd health veterinarians who work with an industrialized dairy herd given the setting and constraints described above. In this setting, 'industrialized' does not only mean large in number of cows but also that the personnel should be considered as part of an organization in which the decision-maker for specific procedures in the herd can be difficult to identify. As stated previously, parts of the Danish dairy industry use a high degree of technology, which requires a well-educated work force. The education affects the ways personnel can best be managed. As such, there seem to be two major directions for workforce management. One is working with standard operating procedures requiring limited individual decision making. The other is working with continuous education and empowerment so that the personnel to a large extent are permitted to make management decisions themselves in some areas.

2.2 Data collection

In Denmark, almost all recordings on individual dairy cows and young stock are collected in the National Danish Cattle Database (NDCB), a database owned by the dairy industry. A subset of these recordings is transferred to the public authorities' public ('CHR register') and non-public databases ('Vet-stat'). Virtually all dairy farmers seem to be willing to contribute to the NDCB, probably because of the strong cooperative tradition in dairy industry and the success of the NDCB for breeding purposes. For this reason, stand-alone in-herd computerized management systems have had little success on the Danish market. They have proved to be not very useful or extremely difficult to use unless they could exchange (two-way) data with the NDCB. This kind of exchange with such systems rarely happened before 2005. Up until the late 1990s, farmers' recordings to the NDCB were based on hand-written registrations of events occurring in the herd (calving, culling, birth or sex of the calf), professional personnel doing inseminations, the veterinarians' disease recording, and recording from the milk yield recording program. Information from the NDCB was limited to standardized reports on paper to the farmer and consultants. As of 1995, it was almost impossible for herd veterinarians to gain access to NDCB's underlying recordings of events for a dairy (raw data). An explanation could be that there was no real need for extraction of raw data because the advisory services had been very limited with little use of herd-specific analyses.

In 1994, the analytic concept developed by Thysen and Enevoldsen (1994) was offered to veterinarians as a freeware PC program (HerdView). Using this program to answer herd-specific questions about herd health and production required herd-specific input files. On request from individual veterinarians, these files could be ordered from the NDCB. Around the same time, the development of a stand-alone PC program for analysis of milk production (lactation curves) led to the creation of raw text files that contained the necessary information to do additional data analysis. In 1997, results from standardized within-herd multivariable regression (Israeli-inspired) analyses were offered to the veterinarians based on these files (Enevoldsen, 1997ab; Jensen, 1997).

In 2002, a database client for the NDCB was developed. With this user client, farmers could make the mandatory registration from the farm online and retrieve production reports. In addition, it was possible to retrieve raw data from the NDCB as a special kind of text file.

To further support the development of the Israeli-inspired herd health program and establish a platform for herd health–oriented teaching and research, a non-commercial SAS®IntrNet (SAS Institute Inc., 2008) portal was created in 2003. The portal is called VPR, which is the Danish acronym for 'Veterinary Production Consultancy'. As the name indicates, the VPR portal is only for practicing veterinarians. This portal has features for uploading the clinical recordings in the herds, verification of the correctness of the registrations (e.g., unique cow numbers), and merging these registrations with data about calvings and milk production. Information about calvings, milk production, inseminations, and similar routine recordings are all downloaded from the NCDB by the veterinarian and subsequently uploaded to the VPR portal. The VPR portal is a non-commercial service with the overall purposes of 1) supporting the veterinarian for their use of data from the herd; 2) providing a place where new tools for herd health management can easily be put into work in the field; and 3) serving as a data-collection portal. During the last 10 years, VPR has become a central part of herd health management for many veterinarians, and it is under constant development.

The author of this thesis has, to a large extent, carried out the development work with the VPR portal. The work has yielded major discoveries and insights concerning the quality of data collected from the numerous dairy herds and veterinarians. Especially useful was the work with a short-term prognostic plan for calving, drying-off, and reproduction, which uncompromisingly revealed data errors because the plan was used for very concrete decision support. This process demonstrated that assurance of high-quality data concerning disease recording and clinical scores requires active use of the recordings for herd health management. Various peculiarities were detected. For example, the number of possible sizes (categories) of the calf were dependent on whether the registration was done electronically or on paper, and the number of milk control dates per year should be 11 and approximately equally spaced over the year; however, checking the data showed that some herds had only 9 or 10, and that even with 11 control dates, they were not equally spaced. Further, registration of dry-cow treatment can automatically give a recording of dry-off, which is highly redundant.

The data used in this thesis all came from the VPR portal. This feature carries the implication that the applied data all are of some interest to veterinarians – data that have been used for herd health management to a varying degree. Alongside the work with this thesis, the author has been working as a practicing veterinarian in typical Danish dairy herds (approximately 3 years of full-time work). This employment has provided additional insights into the entire flow of the recordings from the establishment of a clinical diagnosis to the analytic outcome of the VPR portal.

2.1 Analytic methods

The analytical methods applied in this thesis are primary based on various types of multi-level random regression (section 3.3, 3.5 and 3.6). In section 3.4, the parameters of a Hui-Walter model were estimated with a Bayesian approach using a Markov Chain Monte Carlo sampling algorithm.

3 Results

3.1 A summary of the major results

This section gives a summary of the major results of this Ph.D. project. First, I summarize and demonstrate concepts and tools of particular relevance for a systems approach to dairy herd health management within the context of this Ph.D. project. Then, I summarize four independent studies of specific tools, which are of particular relevance for health performance measurement.

3.1.1 Identification of principles and tools for management of health performance data from the industrialized dairy herd

Problem statement: One part of dairy herd management is to handle disease occurrence by means of health promotion, disease prevention, timely medical treatments, or eradication of disease. It is an essential task for the cattle veterinarian to support this part of herd management. The study objective was to identify principles and tools for analysis of herd health data in industrialized dairy herds. The analysis takes into account the additional complexity caused by changes in behavior among herd managers and herd personnel due to, for instance, legislative changes to promote animal welfare or food safety **Approach:** Methods from herd management science were combined with context-specific information about social mechanisms. **Results:** The results were summarized into a concrete 7-step plan of action. 1) The foundation is the continuing use of process behavior charts primarily based on animal level data. 2) Assure strict definition of the measurements considering purpose, collector, and meaning in terms of biology and management. 3) Interpret the patterns in the process behavior charts, and search for and remove causes of exceptional variation in a dialogue with the herd manager. 4) Search for options to reduce routine variation. Multivariable or multivariate statistical models can give additional information because of their ability to reveal hidden sources of variation. 5) Set targets at tactical and strategic level while accounting for costs and benefits with appropriate methods suggested in the paper. Issues related to non-financial effects are addressed. 6) Adjust measurement and intervention theory. The previous 5 steps should initiate an iterative process, where the intervention is evaluated and updated based on the results achieved this far. 7) Develop a framework in the veterinary practice unit to support the health performance measurement process. The activities in step seven will almost certainly require expert statistical assistance.

3.1.2 A tool to detect rater-introduced bias in clinical ratings

Background: I suggest a 'screening test' to examine large data files with clinical ratings for the occurrence of rater-introduced bias prior to using the data for quantitative analyses. The test is based on a statistical model in which a well-standardized interval-scale outcome (for example, milk yield) is related to clinical ratings (for example, body condition scores) obtained from multiple contexts (for example, dairy herds).

Findings: 84,968 calvings from 279 herds, with subsequent body condition scores performed by 117 veterinarians within the first 21 days postpartum were analyzed with a multilevel random coefficient regression model. The model included an independent variable, where body condition

score was centered within veterinarian. This is a socalled comparison effect to describe possible rater-introduced bias in the body condition scores. A highly significant comparison effect was found for second and older parities, indicating occurrence of possible rater-introduced bias in this large multi-herd data file.

Conclusions: A within-group centering technique (the comparison effect) appeared to be useful for discriminating between biased and unbiased clinical scores. In some cases, this test for bias should prohibit further analysis of the data and divert the focus of study to the calibration of raters or alternative study designs.

3.1.3 Latent class evaluation of a milk test, a urine test, and the fat-to-protein percentage ratio in milk to diagnose ketosis in dairy cows

In this study, three commonly used tests to diagnose ketosis were evaluated with a latent class model to avoid the assumption of an available perfect test. The three tests were the KetoLac BHB (Sanwa Kagaku Kenkyusho Co. Ltd., Nagoya, Japan) test strip that tests milk for β-hydroxybutyrate, the KetoStix (Bayer Diagnostics Europe Ltd., Dublin, Ireland) test strip that tests urine for acetoacetate, and the fat-to-protein percentage ratio (FPR) in milk. A total of 8,902 cows were included in the analysis. The cows were considered to be a random sample from the population of Danish dairy cattle under intensive management, thus representing a natural spectrum of ketosis as a disease. All cows had a recorded FPR between 7 and 21 d postpartum. The KetoLac BHB recordings were available from 2,257 cows and 6,645 cows had a KetoStix recording. The recordings were analyzed with a modified Hui-Walter model, in a Bayesian framework. The specificity of the KetoLac BHB test and the KetoStix test were both high [0.99 (0.97–0.99)], whereas the specificity of FPR was somewhat lower [0.79 (0.77–0.81)]. The best sensitivity was for the KetoStix test [0.78 (0.55–0.98)], followed by the FPR [0.63 (0.58–0.71)] and KetoLac BHB test [0.58 (0.35–0.93)].

3.1.4 A framework for integration of benchmarking and within-herd analysis in dairy herd management – analysis of lactation curves as a case

Comparison of livestock herds' key performance indicators is a widely used management tool for farmers and herd management consultants (benchmarking). The results of this comparison should lead to thorough within-herd and between-herd analyses to identify causes of exceptional variation (unsatisfactory performance). Milk production is obviously the key output from a dairy herd. However, the distribution of milk production between calvings (the 'shape of the lactation curve') can also be an indicator of production efficiency, including health traits. The objectives of this study were to describe the variability of the shapes of lactation curves within and between dairy herds and suggest frameworks for integration of benchmarking and within-herd analysis in herd management. A total of 51,311 lactations and 345,595 test-days of Energy Corrected Milk (ECM) yield from 170 Danish Holstein herds were used in the analysis. A random coefficient test-day model which allows a 'break' around 60 days in milk was applied to parity-groups 1, 2, and older. A single-herd model was applied for each parity group (3) within each herd (170) and a multi-herd model was applied once for each parity group. Internal validity of the models was acceptable. The multi-herd model lactation curve estimates were closer to the overall mean but the numeric differences in lactation

curve estimates between the models were marginal. To demonstrate a principle for within-herd identification of causes of unsatisfactory performance, age at first calving was included in the multi-herd model as a random variable on herd-level and a fixed variable on cow-level. The result was that the impact of age at first calving on herd lactation curves was highly herd-specific. In addition, age at first calving modified the persistency of the lactation curves through an interaction. For a 1 month increase in age at first calving from 26.6 months, herd-level estimates ranged from 24 Kg ECM to 126 Kg ECM per 295 d per cow. The conclusions of the study were 1) both the single-herd and multi-herd model provided similar and valid estimates of the major lactation curve parameters at herd and cow level, 2) the estimates of the lactation curves derived from the models are unbiased and useful for benchmarking of herds, 3) inclusion of other determinants of milk production is possible and provides herd-specific estimates about the relation between the determinant and the milk production, and 4) important information about how the determinant influences the lactation curve is achieved.

3.1.5 Evaluation of effects of disease control in a complex dairy herd health management program

Evaluating the effects of all interventions in a dairy herd, including the effects of various herd health management programs (HHMP), is highly relevant. A traditional randomized controlled trial is the gold standard but is likely practically impossible or prohibitively expensive to use for a general evaluation of a HHMP. Generalizability may also be poor because of the dynamics of the production contexts. In this study, we demonstrate an approach for evaluating the effects of a HHMP in the field, specifying an intervention theory for an ongoing HHMP in the context of the Danish dairy industry. As an example, I suggest one statistical model for studying the possible effects on milk production of systematic post-partum examinations of vaginal discharge, which is supposed to improve detection and treatment of metritis or endometritis. This routine is one component of the HHMP. The data consisted of 121 herds and 76,953 lactations over a 15-year period. For parity group 1, the negative effects of metritis (with treatment) on 305-d milk production after a normal calving were reduced by 17% after enrollment in the HHMP. For parity group 2 and parity group >2, enrollment in the HHMP resulted in a 129 kg and an 80 kg energy-corrected milk yield increase in milk production, respectively. There was some indication that the effect of the HHMP was mediated through improved metritis detection. This study demonstrates the importance of a clear-cut intervention theory although even with a theory, the research question can be too context (herd) specific. In such a case, a within-herd randomized controlled trial study design seems to be the only way to achieve a valid result for a given herd, and acquiring valid results from an observational multi-herd study will be very difficult.

3.2 Identification of principles and tools for management of data for health performance measurement in the industrialized dairy herd

Mogens A. Krogh & Carsten Enevoldsen

Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Grønnegårdsvej 2, DK-1870 Frederiksberg C, Denmark

Manuscript

Identification of Principles and Tools for Management of Data for Health Performance Measurement in the Industrialized Dairy Herd

M. A. Krogh[a*] & C. Enevoldsen[a]

[a]*Department of Large Animal Sciences*
Faculty of Health and Medical Sciences, University of Copenhagen,
Groennegaardsvej 2, DK-1870 Frederiksberg C, Denmark

[*] Corresponding author: email: mok@life.ku.dk (Krogh, M.A.); tel: +45 23659010; fax: +45 96609000

Abstract

Problem statement: One part of dairy herd management is to handle disease occurrence by means of health promotion, disease prevention, timely medical treatments, or eradication of disease. Supporting this part of herd management is an essential task for the cattle veterinarian. The study objective was to identify principles and tools for analysis of herd health data in industrialized dairy herds. The analysis takes into account the additional complexity caused by changes in behavior among herd managers and herd personnel due to, for instance, legislative changes to promote animal welfare or food safety **Approach:** Methods from herd management science were combined with context-specific information about social mechanisms. **Results:** The results were synthesized into a concrete 7-step plan of action, as follows: 1) As the foundation, use continuously process behavior charts primarily based on animal-level data. 2) Assure strict definition of the measurements considering purpose, collector, and meaning in terms of biology and management. 3) Interpret the patterns in the process behavior charts and search for and remove causes of exceptional variation in a dialogue with the herd manager. 4) Search for options to reduce routine variation. Multivariable or multivariate statistical models can give additional information because of their ability to reveal hidden sources of variation. 5) Set targets at the tactical and strategic levels while accounting for costs and benefits with appropriate methods suggested in the paper. Issues related to non-financial effects are addressed. 6) Adjust measurement and intervention theory. The previous five steps should initiate an iterative process in which the intervention is evaluated and updated based on the results achieved thus far. 7) Develop a framework in the veterinary practice unit to support the health performance measurement process. The activities in step 7 will almost certainly require expert statistical assistance.

Key words: Herd health, Dairy herd management, Veterinarian, Health performance measurement

Introduction

The size of dairy herds has increased dramatically in many countries, and it seems relevant to consider the dairy herd as any other industrialized manufacturing enterprise, service provider, or organization in general. Continuous evaluation of the performance of the production process is an essential part of herd (business) management. One part of herd management is to handle disease occurrence by means of health promotion, disease prevention, timely medical treatments, or eradication of disease. It is an essential task for the cattle veterinarian to support this part of herd management. During the last two decades, computer technology, automatic milking systems (AMS), and other automated data collection tools have dramatically increased the amount of data available for measuring and evaluating performance over time in dairy herds. These data may be especially useful for measuring occurrence of diseases with subtle signs (e.g., ketosis and mastitis), which have become relatively more important because major diseases like tuberculosis and brucellosis have been eradicated. The continuing entry and removal of numerous animals, the interaction between animals and management, and feedback mechanisms make the dairy herd a very complicated system or organization, which may make performance measurement and evaluation of performance in the dairy herd more complicated than they may be in most other industries.

Enevoldsen (1993) reviewed technologies and management tools developed for dairy herd health management up to the early 1990s. Principles and tools for measuring and evaluating performance over time were treated in some detail. Inspired by the tools and principles used in manufacturing enterprises and other organizations, including public management, we may find uses of numerous additional tools and principles to be useful. Terms like monitoring, surveillance, control, benchmarking, epidemiological or business intelligence, performance measurement, evaluation, statistical process control, and quality control are widely used. However, the definitions and distinctions between them seem to differ among disciplines, the objectives for application are often vague, and the interpretation can be complicated.

Krogh and Enevoldsen (2006) describe the so-called VPR platform. It was established in 2003 and gives Danish practicing cattle veterinarians access to a growing number of tools for management of health data. During development of the platform and support of the users, we have identified a number of barriers and needs for efficient support of data management for health performance measurement in the dairy herd. Especially when data were used for very specific decisions, errors in collection and management of data were revealed. Based on this interactive development work with veterinarians in the field together with various research and teaching based on the collected data, we will (objectives):

1) identify principles and tools that are of particular relevance to dairy herd health consultants' continuous evaluation of health performance in the industrialized dairy herd, and

2) suggest a coherent set of definitions and tools for management of data for health performance measurement in the industrialized dairy herd.

This work is organized into the following main sections: 1) time series analysis, 2) control and a systems approach to herd management, and 3) a summary of concepts and tools.

The work does not present or discuss simple graphical or tabular presentations of data without attempts to address random and systematic variation in the production process, or support evaluation of the performance of the process by means of some type of limits or criteria.

Time Series Analysis

In herd management, the most common questions are related to time. We want to know whether there are changes in the production process. Detection of changes requires some kind of comparison of the current (or future) production process with some previous production result. As part of herd management, we use a variety of measurement tools to make observations of some activity (variables) at successive points in time. Such data are called *time series data* or longitudinal data. The fundamental elements of an analysis of time series data are (Armitage & Berry, 1987, p. 349) as follows:

- Plot the data before doing any computations
- Look for extreme outliers and search for possible reasons
- Identify obvious long-term trends

The following section presents concepts and tools for such a *time series analysis* of major relevance to dairy herd health management.

The Process Behavior Chart

Figure 1, the upper panel, shows a typical example of a *time series graph* meant for measuring performance of a *process*. In this case, it is a process in a dairy herd, but it could be a process in a factory or a service industry. The data points are the fat percentage to protein percentage ratio (FPR) of individual cows at the first milk test day in the period 5 to 28 days after calving. The diagrams in Figure 1, upper and lower panels, will be described and explained in the following with the terms used by Wheeler (2000, pp. 151–155), who calls the diagrams in Figure 1 a *Process Behavior Chart* (PBC). Above and below the lines connecting the measurements of FPR (the time series graph) are so-called *Natural Process Limits* (NPL). The purpose of these limits is to separate the *routine variation of the process* (the *natural process*) from the *exceptional variation*. If the process exhibits only routine variation, demonstrated visually as all *points inside the limits*, the process will also be *predictable* (within limits). Predictability is an important and favourable characteristic of a process. Consequently, the exceptional points (*points outside the limits*) must indicate something unpredictable, and the cause(s) of the exceptional variation should be continuously explored and, if possible, removed to improve the process (make it predictable). Attempts possibly should be made to reduce routine variation, but doing so will require fundamental changes in the process. This type of change may be necessary if too many results are unacceptable from a product quality point of view. For example, electrical conductivity measurements from an AMS might show only routine variation. Still, an unacceptably large proportion of cows could have mastitis, which would require very time-consuming attention or medication. Therefore, fundamental changes in the AMS or the herd management may be justified.

Wheeler (2000) uses the term *method of continual improvement* to describe the PBC and its intended uses.

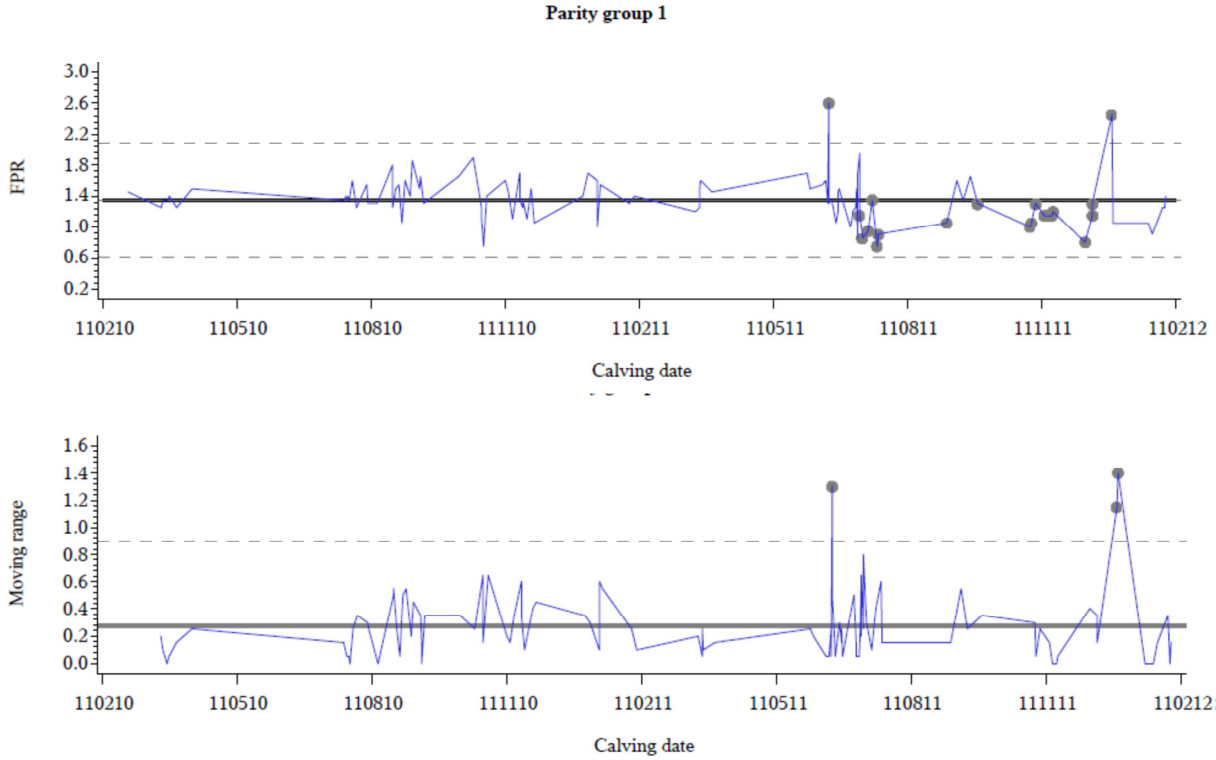

Figure 1. XmR-chart of Fat-to-Protein Ratio (FPR) in milk recorded between 5 and 28 days after calving. The solid lines are the averages and the dashed lines are the control limits. The upper panel is the X-chart. Observations crossing the control limits and observations that fall for the 'runs-rules' are highlighted. The lower panel is the 'moving Range'-chart. Observations crossing the control lines are highlighted.

The data points in the lower panel of Figure 1 are the numerical differences between successive values in the upper panel. They are called *moving ranges* (mR), which directly measure the cow-to-cow variation. The *average moving range* is the average (arithmetic mean) value of the moving ranges and is shown as the lower horizontal line in the lower panel. The lower and upper NPL in the upper panel are derived from the average moving range in the lower panel by multiplication-constants that depend on the type of data (Wheeler, 2000, pp. 136–139); in this case, the constant is 2.66. Similarly, the upper range limit for the average mR is obtained by multiplication with the constant 3.27. A more conservative approach is to calculate a median mR, which may be more appropriate if some few values are very high or low. Indications of possible emerging trends are marked in the upper panel. In this case, a series of more than 7 points on one side of the average is regarded as signaling a trend. This pattern represents one of several of the so-called *runs rules*; of which some are summarized by Kristensen et al. (2009; pages 51-51).

Based on the first author's personal knowledge about the herd from which Figure 1 was derived, the interpretation of the chart can be as follows: In the upper panel, two observations cross the upper limit. These two cows are most likely associated with subclinical ketosis (Krogh et al., 2011). Based on the previously described runs rules, there is a trend towards a lower average FPR from June 2011

onwards. In this specific situation, a similar trend was not found in second and older parities (not shown). For this specific herd, this signal of change in the process was most likely related to insufficient training of fresh first-parity cows to the AMS. First-parity cows were left standing outside the AMS for up to 6 hours, thus reducing their roughage intake and leading to milk fat depression.

The issue with a chart like that in Figure 1 is that we can make two errors: 1) interpret noise (routine variation) as if it were a signal of exceptional variation or 2) fail to detect exceptional variation when it is present. The above-mentioned constants and rules to calculate the limits and define 'signals' are empirical and intended to strike a balance between these two mistakes (Wheeler, 2000, p. 32). Woodall (2000) stresses that this type of chart is "a tool of exploratory data analysis" (of historical data) and that "no assumptions of normality or independence over time need to be made. In fact, distributional assumptions cannot even be checked before the chart is initially applied…because one may not have process stability…". Woodall (2000) disputes the effectiveness of the traditionally used runs rules and suggests alternatives, as well as suggesting alternatives to using the mR chart to identify changes in variability in the process. Koutras et al. (2007) conclude that the sensitivity improvement achieved by supplementing the classical control chart by runs rules, has a trade off in the false alarm rate. In simple words, runs rules increase sensitivity but also produce more false alarms.

Wheeler (2000) vigorously stresses that no assumptions are required for the PBC. In case there are no signs of exceptional variation or trends, intervention is not warranted. In fact, intervention may distort the process (Wheeler, 2000; Woodall, 2000). Wheeler also vigorously stresses that specific knowledge about the context of the process is needed to discover causes of exceptional variation, which is the primary objective of the *method of continual improvement*.

In the increasingly automated systems, the users of the information may become detached from the management of data. To completely and fully describe the context, the user needs to know (Wheeler, 2000): Who collected the data? How, when, and where were data collected? What do values represent? If computed, how were they computed from raw data? Were there changes in formulas over time? We will add that sometimes it is crucial to know for what purpose the data are collected to understand why data can be misleading. These requirements may be a real challenge to a herd health consultant but also an important learning process.

Statistical Process Control

Classical methods

The PBC described above is one simplified version of the Shewhart Control Chart concept, which is among the body of techniques known as Statistical Process Control (SPC), widely used since the 1930s. Kristensen et al. (2009, pages 41-72) give a detailed description of what they call the classical methods for SPC and their applications to various types of herd management data. The major difference between the PBC and the SPC is that the limits in SPC usually are based on distributional assumptions of the measurements (e.g., normal or binomial) and degree of

dependencies between measurements (autocorrelation). For these reasons, these methods are separated from the PBC in this presentation. The validity and importance of these assumptions may be very questionable and hard to judge. Woodall (2000) quotes Hoerl and Palm (1992) as stating that "the underlying model (for SPC) then is only that one has a series of independent random observations from a single statistical distribution. The control chart rules are used to detect deviations from the model, including the model assumptions themselves". In statistical terminology, this concept is called model control. De Vries and Reneau (2010) discussed the effectiveness of SPC based on their comprehensive review of applications of the control charts in animal production. Their main conclusion was that an actual search for the true causes of exceptional variation is very difficult and seldom done. Papers on the practical benefits of implemented control chart schemes were not found. Run length distributions (an indicator of SPC effectiveness) were only found in papers describing simulations studies, which may be problematic because simulations usually are based on assumptions about distributions, which we rarely know in a real life setting. Wheeler (2011) claims that autocorrelation (that is, non-independencies of the series of observations) should not influence the limits for NPL. The argument is that autocorrelation will cause a trend (signal) that should be explored and the cause(s) identified and removed. If this exploration and intervention are successful, only routine variation remains, and routine variation will not contain autocorrelation.

Another major difference between SPC and the PBC is that SPC in many cases shows only data that are *filtered* or *smoothed* to better reveal patterns in the data. This process is achieved by calculating one of several types of *moving averages*. One possible choice is the average of the latest 12 months plotted for each month, which will eliminate erratic fluctuations (*smoothing*). The moving average may also be weighted so that the latest measurements of the time series are given more weight than the preceding ones. Such weighting is generally recommended to avoid reactions resulting from removal of the oldest historical data. Smoothing may also reveal harmonic variation, which often is caused by seasonal or diurnal factors in the dairy herd. Basically, smoothing serves the same purpose as the runs rules for PBC. Methods for calculating various types of moving averages are available in widely used spreadsheets. However, these simple tools do not always provide limits, probably because calculation of the standard errors becomes more complex. Wheeler claims that some methods to calculate limits applied in standard software are quite inappropriate (Wheeler, 2010).

Woodall (2000) stresses the importance of distinguishing between an initial purely explorative time series analysis like PBC (phase 1) and a subsequent SPC based on the results of the explorative time series analysis (phase 2). In phase 1, we may find justifications for assuming homogeneous processes or certain distributions (e.g., normal or binomial) that permit application of a series of parametric analytical techniques that may be used for prediction and quantification (phase 2, methods addressed below). Woodall (2000) supports the view that the PBC is very robust but also states that "there is a wide difference of opinion on how much robustness is needed in practical applications, so there may always be some disagreement on this issue". Wheeler (2011) probably represents the most extreme view by stating that "We do not need to check for normality or

transform the data to make them 'more' normal. We do not have to use sub-grouped data to receive the blessing of the central limit theorem before the chart will work. We do not need to examine our data for autocorrelation".

Performance measurement by State Space Models

Figure 2 provides an example of a concept suggested by Thysen (1993). The individual data points are the same as those in Figure 1. The solid line is the filtered prediction of the process at each data point. Outliers (another word for exceptional variation) are indicated by a circle. The solid line (the prediction) can take the following positions: Level shift or 'normal evolution'. An outlier will not affect the prediction.

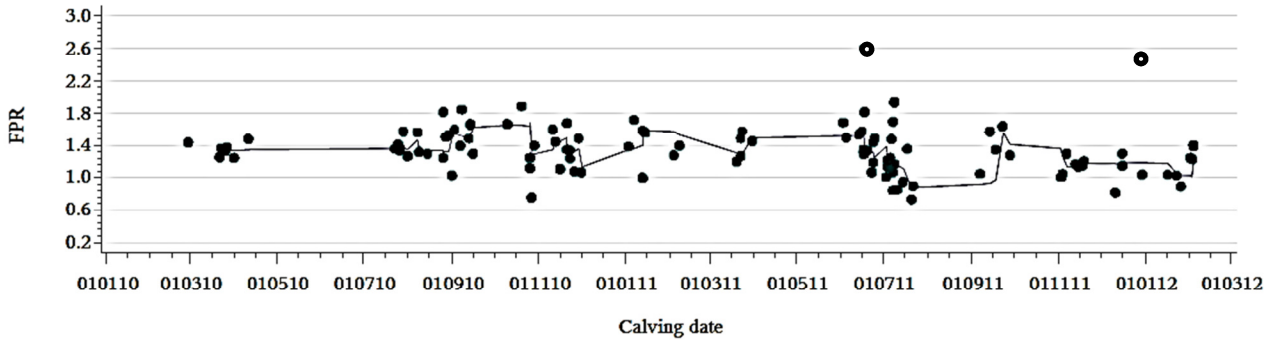

Figure 2. Fat-to-Protein Ratio (FPR) in milk recorded between 5 and 28 days after calving. Each observation is given by a dot. The solid line is the prediction of the FPR updated at each new observation. Observations with a high probability to be outliers are identified by the model (circles) and do not contribute to the prediction.

Figure 2 is one example of the so-called *state space models* (SSM). Kristensen et al. (2009) describe SSM and their potential applications for herd management in detail. The general purpose of a SSM (Kristensen et al., 2009, p. 74) is to estimate the parameters in a mathematical model (e.g., regression coefficients or variances) that combines information from the observed data (e.g., the data points in Figure 2) with some information available before data collection starts (e.g., expected effects of some intervention like changes in milking routines). A major advantage of this type of model is that it is a natural formulation of the Bayesian approach, which means that *a priori* knowledge can be combined with new information in a systematic fashion. Important assumptions can include types of distributions of error terms (e.g., normal or binomial), type of correlation between measurements, or thresholds for level-shift or outlier. A simple SSM model for dichotomous fertility data is described by Thysen and Enevoldsen (1994). The trend-line is supplemented with a graphical display of the dynamics of the raw data to support a qualitative exploration of potential causes of (exceptional) variation. This concept is implemented in freely available software for herd management support (Thysen and Enevoldsen, 2011), which is applied by a substantial number of Danish cattle veterinarians (we track the use via the download of data from the VPR-platform). The assumption of a binomial distribution behind this concept is not tested. Justification of the binomial distribution would require providing evidence that all cows in the observation period had the same chance of experiencing the events (insemination or pregnancy) (Wheeler, 2000 page 141).

In the very simple PBC concept described above, it is the manager's or the consultant's task to react to signals and start a search for causes of exceptional variation. This reaction may require some type of more or less complicated statistical analysis. In the much more complicated SSM, a statistical analysis essentially is embedded in the time series analysis. That approach may give more valid signals but at the cost that the assumptions must be justified, which may be a rather complicated task. In fact, a statistical model control is required, and outliers or lack of fit detected by means of model control tools can be considered signals of deviations from the assumed (statistical) theory. In case of signals, the managerial reaction must be directed towards a search for both causes of exceptional variation (a qualitative context-bound search) and an appropriate statistical model. We suggest it will be simpler to start with the virtually assumption-free PBC, especially in the typical dairy health management context where numerous health measurements are available and relevant. Even if a SSM is validated in one context, it is very likely that distributions and causes of exceptional variation are different in another context. Because statistical model control is a task for experts, this approach may be impractical with many herds and numerous indicator variables in each herd, as is the case for the work context of the herd veterinarian.

Multivariate Statistical Process Control

With the increasing number of herds with automatic data collection, both the number of health, fertility, and production indicators and the measurement frequency increase dramatically. Some of these indicators will be correlated. So-called *Multivariate Statistical Process Control* is an analytical concept designed to handle the correlations and the large volume of data. By 'multivariate analysis', we mean that several variables are analyzed jointly by creating a new Y-variable (response variable) that is defined by the correlations between the original variables. The new indicator may represent an unobservable (latent) condition that has an interpretation or simply a hidden data structure. The calculations are usually based on so-called *principal components*. The concept with control limits is the same as in SPC. The variance can also be exposed to time series analysis with the SPC concept. However, the interpretation of out-of-control points becomes more complicated because they cannot be directly linked to one single indicator. The concept was developed several decades ago and is implemented in standard software (e.g., MVPMONITOR procedure, SAS Institute Inc., 2011).

We are not aware of practical applications or interpretations of Multivariate Statistical Process Control for dairy herd management, and it is not addressed by Kristensen et al. (2009). Enevoldsen et al. (1996) applied second-order factor analysis (a similar statistical technique) to condense 22 herd-level indicators of health, fertility, and production into 10 and 5 first- and second-order factors, respectively (new variables), but these new variables were not used for time series analysis.

Numerous tests are available for disease diagnosis in the dairy herd (e.g., mastitis pathogens in milk or ketone bodies in urine). In fact, every comparison of performance measurement with the associated target value can be regarded as a diagnostic test. Because diagnostic tests (including performance measurements) will be used for decision support, it is necessary to evaluate the quality in terms of sensitivity and specificity. However, information about these parameters and the

associated uncertainty is often insufficient. If information about the validity and precision of a given diagnostic test is insufficient, the herd manager cannot know how an intervention based on the test results will work. Virtually all diagnostic tests are imperfect. However, knowledge about some underlying unobservable state can be obtained by combining tests similar to the multivariate technique described above. Krogh et al. (2011) used a *Latent Class Analysis* (LCA) to handle this problem for diagnosis of ketosis. The LCA might be combined with the SPC tools outlined above.

In some aspects of dairy production, we have a solid theory about the relationships between measurements that allows us to combine a number of measurements into one meaningful combination. This approach is in contrast to the purely data-driven condensation of variables by means of principal components or similar methods. An example is the so-called lactation curve. Krogh and Enevoldsen (2012a) demonstrated an analysis of milk yield recordings in which the shape of the lactation curve is defined by multiple variables in a coherent way that takes into account correlations between variables. In the case of the lactation curve, we have an example of a hierarchy of indicators and applications. We can use some components (e.g., the parameter for acceleration early postpartum) as a direct health indicator, the combination of all parameters into a lactational yield per cow, and the summation of yield from all cows into a herd-level indicator of milk delivery.

In recent years, the emergence of social media and other digital stores with vast amounts of text has created a need for automatic detection of emerging trends in, for instance, buying patterns. This search is called *text mining*. Search engines like Google are based on such tools. The increasing requirements for documentation by means of various reports in the dairy industry may create a need for development of tools for continuous text mining to support health performance measurement. Computerized text analysis has been applied by Allaki (2005) for the veterinary authorities' surveillance of health. Text mining is also implemented in standard software (SAS Text Miner, SAS institute inc., 2010).

Multilevel Statistical Process Control

In a dairy herd, data are produced at multiple organizational levels (e.g., udder-quarter, udder, lactation, cow, group of cows, and herd). The data from these levels may be correlated, and such dependencies should be accounted for. The correlations could, for instance, be taken into account by pooling the recordings from the four quarters (e.g., electrical conductivity) into one average udder-level measurement. However, important information may be lost by this aggregation. Some of the methods described above may be developed to handle this situation effectively. We are not aware of practical applications for herd management, but industrial applications are reported.

Control and a Systems Approach to Herd Management

The mainly explorative analytic approaches described in the previous sections will enable us to detect changes within the processes in the production system. However, the historical results from an actual herd will not necessarily tell us whether the resources could have been used better in that

herd. That is, was the performance acceptable, really good, or poor? Or was it *optimal* from a resource use point of view? The following presents relevant approaches to answering this fundamental question. Often this *evaluation* is called *control* in the management literature.

Benchmarking

Benchmarking is one obvious way to select targets. In its simplest form, it could merely be a herd manager asking his neighbor about the performance in his herd as a tool to judge his own results. More systematically, the principle of benchmarking is to identify several other herds with a similar combination of resources as our case herd and compare the performance measurement in our specific case herd with the range of results in these reference herds. This comparison will indicate performance level at *best practice*. For instance, what is the range of values in the best 25% of a performance indicator (e.g., milk production)? A formal comparison of targets and performance measurements may now allow us to evaluate whether we are on target or not and determine if the system is performing satisfactorily. In addition, dissemination of these targets to the farmers may motivate changes in management (Nir-Markusfeld, 2003). The selected target performance measures can also be considered a *prognosis* for the future or a *budget*.

A fundamental problem in benchmarking is to decide when a potential reference herd really is a comparable herd. It is straightforward to find herds that are comparable with respect to very general characteristics like herd size, breed, type of ration, or milk production level. To further investigate if these herds are truly comparable, the methods described above or the methods described below can be used to delineate the production systems in sufficient detail to judge whether they are comparable.

The principles of benchmarking used in *stochastic frontier analysis* in which a 'best performance' frontier is estimated to describe the best performance given a specific set of input factors (Kumbhakar and Lovell, 2000). Also *Data Envelopment Analysis* (DEA) describes such a frontier but is driven by actual observations (performance measures), instead of detailed knowledge about production functions. Bramsen and Nielsen (2004) provided an example of DEA in pig production. DEA does not account for uncertainty in the variables. In practical management of Danish dairy production, benchmarking on health indicators so far seems to have used one performance measurement at a time (univariable), which does not account for the correlation between the performance measures.

Correlation between performance measures in essence means that calculating additional performance measures will yield only minor additional information. The negative correlations are the most troublesome because targets often are derived from univariable analyses. In the case of lactation curves, Krogh and Enevoldsen (2012a) addressed this issue in detail. An increasing peak yield is strongly associated with a steeper decreasing slope afterwards, but because the correlation varies from herd to herd, the correlation can be a performance measurement per se.

It is obvious that benchmarking is invalidated if the scale of a measurement differs from herd to herd. Milk yield, fat percentage, and somatic cell counts (SCC) are examples in which the scales are calibrated in central systems. However, for the cattle veterinarian, animal-level conditions like body condition, lameness, and skin lesions are examples in which scoring systems (ratings) are needed. These '*clinical recordings*' obviously must be standardized to be useful for benchmarking. Clinical criteria that are constant within herd (e.g., specific for a single manager or veterinarian) may suffice if performance measurement is restricted to historical data within the herd. Kristensen et al. (2006) demonstrate typical variation in scores and that agreement in clinical scores quite easily can be improved with training. Consequently, before any target health performance measurement (indicator) can be chosen, the quality of available clinical records must be evaluated. The evaluation essentially includes estimation of sources of variation (random, within-herd, between-herd) and identification of systematic errors in data collection.

Even when score values are described in detail in manuals or protocols, they may be used differently by veterinarians or others doing recordings in the herds (Lastein et al., 2009). The veterinarians' perception of the herd health management system could influence the basic clinical recordings. Recordings of disease treatments are also influenced by herd-specific conditions (Vaarst et al., 2002), which will make comparability across herds very poor. Krogh and Enevoldsen have described a concept to detect this type of measurement error (2012b). This approach could be useful in a large veterinary practice that might want to develop a benchmarking system based on recordings from multiple veterinarians in the practice.

Data used for benchmarking are often an aggregation of data for a longer period of time (e.g., a year or a quarter of a year). The same time interval is usually used in routine reports to evaluate the performance of a given concrete herd. In case we have not discovered an important time trend, we may miss a signal or get a misleading signal. Averages, ranges, and histograms all obscure time order, which can be misleading (Wheeler, 2000). If, for instance, performance has improved markedly in our case herd, we might be interested only in the value for the latest month. Consequently, an appropriate time series analysis with as few restrictions as possible should always precede traditional statistical analyses like benchmarking or statistical modeling (Armitage & Berry, 1987).

Planning tools to derive targets for performance

Even if we have identified 'comparable' herds, specific constraints or personal values may persist that make the concrete herd unique. Therefore and ideally, regular and iterative planning processes should produce herd-specific plans that again should have formulated goals for health, fertility, production, etc., based on the system context and the use of the available input factors like feed, medicine, and management. The goals should be specified as targets for the performance measurements that can be derived routinely from the production process (Kristensen et al., 2009). A simple approach to setting herd-specific targets is to take historical results and adjust them for expected results of the planned changes in the next planning period. Enevoldsen (1993) demonstrated this simple approach for a series of health and fertility performance measures. The

expected results (targets) of changes in plans were based on a mix of general theoretical knowledge and context-specific knowledge about the herd and the management.

Numerous advanced tools are available for planning. Major examples include (Kristensen et. al., 2009): expert systems (based on norms and logic), linear programming (widely used to formulate feed rations), dynamic programming and Markov decision processes (e.g., used to select the optimal time to replace cows), Bayesian networks and decision graphs (very complicated development of decision trees that represents uncertainties of decision problems), and simulation (computer model of an entire system; e.g., a herd). Ideally, the targets should be estimated from an optimization of the available resources. This optimization can be obtained by means of some of these tools. For dairy herd health management, a very complicated and scientifically well-documented and commercially available herd model is adapted to the needs of practicing cattle veterinarians (Østergaard et al., 2010; www.simherd.com).

The requirements for performance measurement will depend on the time horizon. In herd management, science decisions about the production system have traditionally been divided into the strategic, tactical, and operational levels. Operational decisions typically relate to day-to-day management routines in the production process. The effects of operational decisions can quite quickly be implemented and evaluated, and the economic impact of the individual decision is often of minor magnitude for the herd as a whole. The tactical decisions are in the month-to-year time frame. The decision could be to increase the amount of labor and change the feed ration. Strategic decisions are long term. The decision could be to build a new stable, increase the number of dairy cows, or convert to organic farming. The needs for and types of performance measurements are very different at these levels.

Wheeler (2000) provides numerous examples of the errors that can occur if the target setting and comparison with an aggregated single-value performance measurement are used alone in some 'Annual Report' without a detailed preceding time series analysis. In fact, his view seems to be that the aggregated report is unnecessary if an appropriate PBC analysis is conducted. The advantage of this graphical approach is that we avoid definition of arbitrary (non-biological) cut-offs between time periods.

Causal analysis supported by Multivariable Statistical Modeling

The application of the tools for time series analysis usually will create a need for further analysis to identify causes of exceptional variation or emerging trends, or options for reduction of routine variation (that is, to re-engineer the system). A possible need for setting targets may also require additional analysis. Well suited for both purposes are *multivariable statistical models* (MSM; e.g., logistic and linear regression or analysis of variance), which have been used for research purposes for many years (e.g., Armitage & Berry, 1987). Implementation of MSM at a larger scale for herd management is described by Markusfeld (1993), Enevoldsen (1997ab), and Nir-Markusfeld (2003). Examples of important information produced by such MSM include: differences in milk production between cows with or without mastitis, differences in chances of pregnancy in cows with or without

previous metritis, and risk of early culling in cows with or without ketosis. If the analyst has context knowledge about the herd, such information can be valid as estimates of predicted effects of management interventions to reduce disease occurrence. A MSM can also be used to estimate a time trend in a performance measurement. Singer and Willett (2003) and Kristensen et al. (2009) suggest a range of approaches for modeling change and event occurrence. Multiple levels (e.g., cow, herd, and veterinary practice) can also be handled (e.g. Krogh and Enevoldsen, 2012c). The advantage compared with the time series analyses described above is that numerous possible confounding factors like parity and stage of lactation can be accounted for in a systematic fashion. Consequently, time trends derived from a MSM may be more valid than time trends derived from the time series analyses. In fact, a MSM may also detect time trends that were not detected by the time series analyses because they were hidden by confounding factors. However, application of MSM relies on several assumptions like distributional properties, independencies of data, or appropriate model specification. Prior application of a PBC may help in identifying situations in which these assumptions are justified. Results of statistical model control may also serve as signals of changes in the process or signals of exceptional variation. Appropriate model control should also detect violation of distributional assumptions.

Quantitative and qualitative methods for a Systems Approach

Andersen and Enevoldsen (2004) give an example of the challenges we can face when a herd health consultant works with the herd manager. Figure 3 represents the synthesis of thorough successive quantitative and qualitative analyses of a single herd conducted at several herd visits and discussions with the herd owner over several months. The *production system* is composed of cows, housing, feeding, and technical equipment. The production process transforms input factors to output (products, milk, meat, and livestock). Measurements from the production system (quantitative data) are used by the farmer to adjust the flow of input (*feedback*). One view on herd management can be that this adjustment is according to simple decision criteria. However, the case behind Figure 3 demonstrated that this particular *farmer's action system* was very complex and dynamic and involved feedback mechanisms. Personal values and views on the role as farmer in the community played some part. Andersen and Enevoldsen (2004) described the entire system as a learning system in which *double-loop learning* took place.

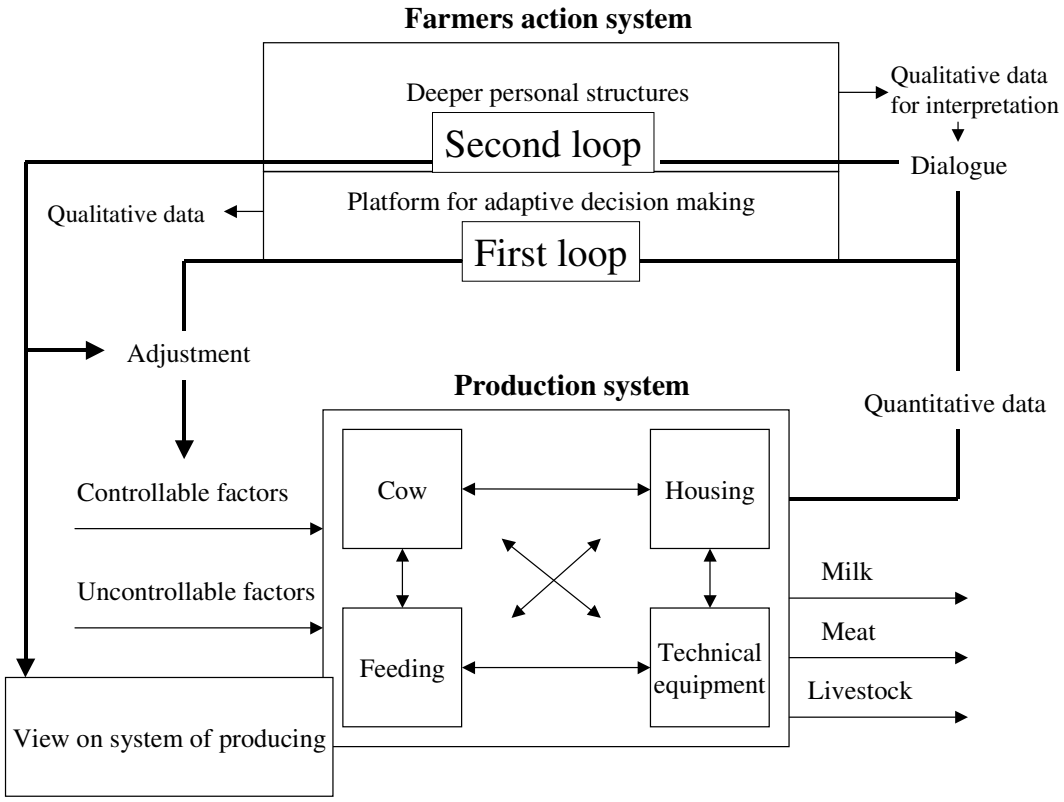

Figure 3 Factors, relationships, feedback, and interactions in a system comprising the production system and the farmer's personal action system (Andersen & Enevoldsen, 2004, with permission)

The joint application of some of the tools described above for performance measurement, including tools for setting targets, is demonstrated by Enevoldsen et al. (1995), where *a systems approach* (Kristensen et al., 2009, pp. 251–252) is applied to a concrete case-herd. This approach allows us to express our prior knowledge of the qualitative and quantitative structures of the system we work with. Complicated computer models usually play a major role in a systems approach. However, essential parts of the information needed for input to the computer model must be derived from the herd manager (cf. Figure 3).

The analysis and subsequent synthesis of a theory about such a system as described in Figure 3 require much more than routinely collected data. A lengthy dialogue is needed to establish a genuine common understanding between the farmer and the researcher. Several qualitative research techniques are useful for such purposes. However, the information obtained with these qualitative methods can also be very useful for specifying and using MSM to analyze the quantitative data. In the particular case demonstrated in Figure 3, advanced quantitative decision-support tools probably would have been of very limited use if applied without the qualitative knowledge obtained. The qualitative knowledge, in contrast, probably would be quite useful alone.

Kristensen et al. (2008) use the term *mixed-methods research* (MMR) to describe the research approach leading to a model like the one in Figure 3. MMR basically is rooted in the social sciences. Kristensen and Enevoldsen (2008) use a so-called Q-Method to obtain more general knowledge about current subjective views like the manager's views indicated in Figure 3. The latter study also showed that the subjective views on consultancy differed markedly between cattle veterinarians and dairy farmers. This factor illustrates the importance of establishing a genuine common understanding of the entire system. From the quantitative perspective, Wheeler (2000) also stresses the importance of context knowledge by specifying a (somewhat provoking) 'first principle for understanding data': *No data have meaning apart from their context.*

Major effects of public management and other organizational constraints on performance data

Figure 4 shows a Process Behavior Chart from a dairy herd during a 4-year period. Limits are empirical and estimated as described for figure 1. The average treatment rate and the natural process limit, based on average moving range, are calculated on the entire time period. The performance measurement is the rate of medical treatment for interdigital phlegmon (IDF) among the cows in the herd. From figure 4 it is evident that there is a clear change in the treatment rate from July 2008. The issues related to proportions and rates are discussed by Wheeler (2000 pp. 140-142). The assignable cause of the marked shift(s) was not a change in the biological processes but a change in the criteria for defining the diagnosis. New legislation introduced some disease categories in which farmers legally could get drugs and others in which they could not. For IDF, a farmer could get prescriptions but could not do so for digital dermatitis (DD). Not surprisingly, the manager had a strong incentive to use IDF instead of DD in cases of foot problems. For the herd presented in figure 4, the herd entered the herd health management program and the new legislation in July 2008.

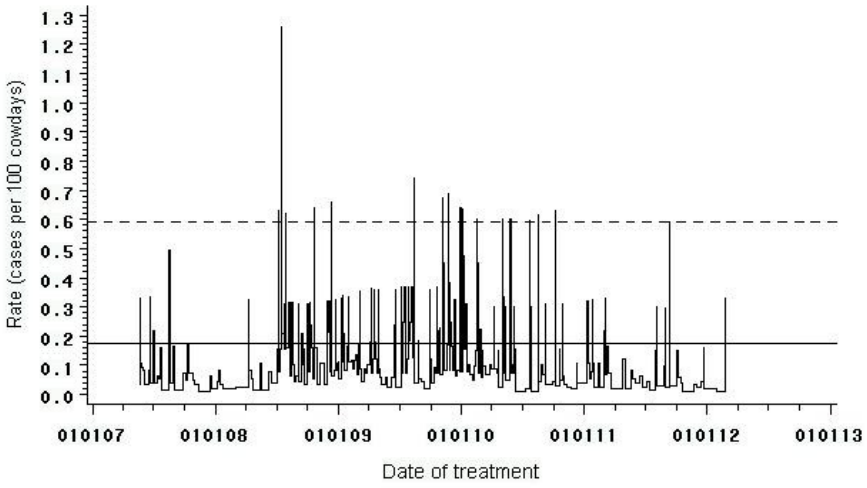

Figure 4. Rate of Interdigital Phlegmon treatment over time in one herd. The average treatment rate (solid line) and natural process limit (dashed line) are calculated on the entire time period. The average moving range is used for calculation of the natural process limit.

Another example is the use of SCC in the milk sold to the milk processor as an indicator of udder health. Because milk payments from the milk processors are reduced in cases of SCC above certain limits, it is quite obvious that farmers have an incentive to discard milk from cows with high SCC values. Consequently, the value of SCC in deliveries as an indicator for the herd's udder health status may be distorted. What happened here is what Wheeler (2000) called the *Voice of the Customer*. That is, the decision takers in the organization attempt to adjust to the needs of the outside world while the process per se is not changed.

Such distortion of the data is not seen as a problem for the manager or the local consultant because they know what goes on in the process. However, an outside observer without sufficient context knowledge (e.g., a statistician working with large data files for research or a veterinary officer doing follow-up on the legal regulations) may draw naive conclusions about the process, which might lead to unjustified political interventions or causal inference. The upshot could be reduced efficiency of the process or even its misdirection.

A misinterpretation of data like the one outlined above is also recognized in the social sciences and basically viewed in the same way as Wheeler (2000), who gives an example (pp. 70–71) and states that "…pressure to meet any arbitrary numerical goal or target will most often result in the distortion of either the system, or the data, or both". Krogstrup (2011) calls such a local distortive management reaction to outside regulation or requirement a '*perverse side effect*' in a thorough discussion of performance measurement, effect evaluation, and evidence in (New) Public Management. As an example, targets for the rate of dead cows and calves are now incorporated into Danish legislation. Despite the fact that the targets are extremely high, the first author has experienced that simply setting the targets has made some farmers change behavior. Some farmers became more reluctant to euthanize chronically ill cows, instead keeping them in the herd, hoping for recovery. The consequence is that in some herds, there is a substantial amount of 'accumulated suffering' – cows kept in the herd suffering from various conditions with poor prospects for recovery. This example represents a perverse side effect because the purpose of setting the target was to improve animal welfare. It is clear that inclusion of these sociological aspects will make even more complicated the rather complicated representation of an organization in Figure 3.

Krogstrup (2011) defines the term 'performance measurement' as the combination of measurements of processes (what goes on in terms of, e.g., types of management routines (actions) like heat detection), output of the processes (in terms of what was actually done in the process-routines; e.g., minutes of heat detection every day), and results (outcome; e.g., pregnancy rate). In our herd context, it is implicit that the process is influenced by some intervention and the context (competencies and capacity). That is, by measuring 'output', we measure the intervention that has taken place. The outcome is the result of the output (process). This outcome (results) is what the recipient experiences. Wheeler (2000) basically uses the same demarcation by distinguishing sharply between *The Voice of the Process* (performance of the process per se) and *The Voice of the Customer* (the quality of the products). A subset of the outcome is the direct or the indirect *effect of the intervention*; that is, the causal effect(s). Management of an organization can be based on

measurements of the outcome; an *evaluation* of whether the results are on target (in new public management terms, a results contract). In this public management context, the term 'evaluation' may seem similar to the term 'control' described above for herd management. However, Krogstrup (2011) gives a broader definition of evaluation: "A systematic retrospective assessment of output (process), outcome (results), administration, and organization of (public) business, which is expected to play a role for practical actions". In this definition, it is essential to note that evaluation includes some judgment that separates important aspects from unimportant aspects. It is also essential that *practical use* is intended. For an intervention to be practically applicable, we need to know how and when it works. This view is similar to the term 'surveillance' used by Schwabe et al. (1977) and Stärk and Salman (2001) in epidemiology. They use surveillance as some active goal-oriented process (Schwabe et al., 1977: 'information for action') in contrast to monitoring as some passive data collection (measurement) without evaluation. If no decision or action is possible, then the measurement does not provide information and is thus worthless for management. Kristensen et al. (2009) do not make a distinction between monitoring and surveillance and simplify the complexity of views, values, interaction, feedback, and learning into a general term like 'utility function' without addressing the problems of identifying this function in practice. To us, the parameterization of a utility function seems to be a big challenge in a veterinary practice context, especially because Figure 3 indicates that the utility seems to be dynamic.

With the increasing public focus on regulation of animal production (e.g., animal welfare promotion and reduced usage of antibiotics), it follows that there will be an increasing need for evaluation of the results of the interventions and ideally the effects of the interventions. In large herds with large personnel, some incentive systems based on obtained results may be used. That is, perverse side effects may be an important issue to consider for both local and public management of data collected from the herds. For the purpose of providing documentation of the state of the production system to public authorities, the manager probably does not see 'perverse side-effects' as perverse.

For obvious reasons, we want to know as much as possible about the causal effects of interventions. In a simple-problem context like assessment of the effects of mechanical changes in an AMS on the frequency of cows' visits to the robot, a quantitative estimation of the effect is straightforward with the numerical methods outlined above, if sufficient context knowledge is available. Krogstrup (2011) calls such a problem a *tame problem*, in contrast to identification or quantification of causal effects (evaluation) in a context like Figure 3. Krogstrup (2011) calls a problem similar to that in Figure 3 a *wild problem*, which mainly is characterized by a vague definition, lack of an optimal solution, unclear causal mechanisms, and interaction between context and mechanisms. Krogstrup (2011) gives a thorough discussion of the possibilities for evaluation of such problems. One prerequisite is to specify an *intervention theory*. Often, the modest ambition will be to explain why some intervention did not work. Basically the formulation of Figure 3 will allow us to identify key elements that can be addressed with a mixed-methods approach. Again, context knowledge is essential. Krogstrup (2011) uses the term Context-Mechanism-Outcome, which means that interventions cause mechanisms that then selectively interact with the case-specific circumstances (context) and result in effects that differ in different contexts. A very complicated system like this

can be considered self-organizing (Rickles et al., 2007). The term *complex responsive processes* (Stacey, 2001) seems applicable, as well. This concept describes organizational knowledge as being in the relationships between people in an organization.

A clear-cut context-specific intervention theory is also needed to reduce the number of potentially relevant performance measurements that otherwise easily becomes large, causing the overview of the system to be lost. Krogstrup (2011) gives an overview of approaches to evaluate evidence of effects of intervention in the spectrum of contexts, from tame to wild, from the randomized controlled trial, which is regarded as the ideal in medicine but is impossible to apply to wild problems, to the everyday evaluation, or an *effect-focused practice*. A systematic use of the simple PBC in a herd (which includes more or less qualitative follow-up to remove effects of exceptional variation) could be seen as an example of an effect-focused practice.

A definition of (herd) health in the context of a Systems Approach

In the preceding text, we have not defined health; we have focused on management of measurements related to disease occurrence. However, our presentation and discussion of these concepts and tools bring us closer to an understanding of health. In standard veterinary textbooks, explicit definitions of health are rare (Gunnarsson, 2006), and Houe et al. (2004, p. 25) also state that health is often defined for a very specific context. Hence, a definition of herd health is at least as problematic. A similar problem exists in humans, for whom the term 'public health' sometimes seems to be defined only as preventive medicine – the science of preventing diseases. However, much broader definitions also have been applied that involve the interaction among society, population, and health, intended to improve the health of the population through education and preventive medicine (e.g., MacQueen et al., 2001).

In a herd health context, the difference between the health of an individual and herd health is that herd health is concerned with the herd as a system, as illustrated in Figure 3; that is, not only the population of animals is of concern but also the 'support' for the population as environment and management. Based on Albrecht et al. (1998, p. 57) and the concepts described above, we propose an analogous definition of herd health, which then can be, *"Animal, environment, and manager together viewed as a dynamic and complex ecosystem. In this context, an ecologically informed or process-view of herd health implies the self-regulation through feedback and maintenance of all relevant systems promoting ongoing physical, mental/emotional, and social well-being. This latter definition gives us a sharper understanding of what poor herd health is. That is, the loss of the ability to self-regulate and the disintegration of support systems leading to the necessity for intervention. In a process-view, intervention is directed towards restoration of all relevant support systems in order for health again to be self-generated and self-regulated"*.

In this definition, it is important to acknowledge that being healthy in a herd health context involves the herd managers' conception of the animals' well-being. Thus, the role of the herd manager (context) is pivotal.

Summary of concepts and tools

It is our experience from several countries that often the only tools for health performance measurement in dairy herds are simple graphical or tabular presentations of data without attempts to address random and systematic variation in the production process. Also, there is limited or no support for systematic evaluation of the performance of the process by means of some type of limits or criteria for intervention. In the following, we suggest to the herd veterinarian for cattle herds a concrete stepwise approach to using the concepts and tools for management of health performance measurement data presented above to develop a systems approach to herd health management in an industrialized dairy herd.

Step 1: Develop process behavior charts like that shown in Figure 1 for the available routine measurements from standard herd management programs. These charts do not require sophisticated software or hard-to-justify assumptions. Use animal-level data directly whenever possible. Do not wait until ideal data are available; there will always be data available that are useful for health performance measurement.

Step 2: Make sure you can answer the following questions concerning the definition of the measurements: For what purpose were data collected? Who collected the data? How, when, and where were data collected? What do values represent? If computed, how were they computed from raw data? Were there changes in formulas over time? Precise knowledge about these topics in the concrete herd will give a very strong and necessary foundation for interpreting the charts. Knowledge about the specific context and the dynamics in the context will increase. Meeting these requirements may be a real challenge for a herd health consultant but also an important learning process.

Step 3: Interpret the patterns in each chart, search for assignable causes of exceptional variation (outside limits or trends), and attempt to remove such causes. This systematic process will add further to your knowledge about the herd context, including the manager's more or less subjective views. The charts and your use of them will document your reasons for suggesting interventions to the herd manager and, if needed, to the public veterinary authorities. You will also be able to distinguish clearly between process-related and results-related measurements and experience the difference between them through the dialogue with the manager.

Step 4: Search for options to reduce the routine variation when the results of the process are unsatisfactory. Some options will be obvious (e.g., repair technical faults in the milking equipment or ensure hoof trimming). However, because of the usually large number of animals and long-time horizon in dairy production, you will profit from some multivariable or multivariate statistical modeling. A range of traditional statistical models and state space models are developed specifically for this purpose (presented and discussed above). Model control of these analyses can also serve as advanced tools to explore causes of exceptional variation. Standard setups are available, and the younger generation of veterinarians has been trained in using simple versions. This process will also add substantially to your context knowledge.

Step 5: Set up targets at the tactical or strategic level. The interventions to reduce the routine variation or simply improve the results by eliminating product out of specifications (e.g., bulk milk cell counts above penalty limit) will often require some investments, which are quite easy to estimate. However, the benefits in terms of increased production or decreased disease-associated losses are more complicated to assess. Models to do such analyses are described above. Some are commercially available, and you can get support for interpretation and use. With the knowledge gained during steps 1 to 4, you will be well equipped to provide relevant and comprehensive input to these models. The models provide predictions of the important health performance measures and potential profit due to the interventions you consider. The discussions of the results with the manager will bring you deep into the topics described in Figure 3, which again will provide knowledge about causes of exceptional variation. The entire process in step 5 will also provide some estimate of the economic value of each health performance measurement.

Step 6: Adjust the measurements and the intervention strategy. Steps 1–5 should initiate an iterative process. Some measurements will be dropped, others added, the quality of the measurements assessed, process limits or targets possibly changed, cost–benefit assessed, etc. In essence, you have established a systems approach to dairy herd (health) management like that outlined above.

Step 7: Develop a framework to support the health performance measurement process at the practice level. This will be particularly useful for establishing a basis for benchmarking because the context knowledge obtained in steps 1 to 6 will allow identification of the most comparable herds. Above, a tool is presented for identifying rater bias in ratings used for health performance measurements that must be corrected prior to benchmarking, or across-herd analyses to, for example, evaluate the effects of various interventions like those discussed above in the case of metritis diagnosis and treatment. The validity and usefulness of across-herd analyses will be greatly improved compared to data from larger data collections from multiple veterinary practices. A homogeneous set of data will also be useful for evaluation of diagnostic tests applied in practice and development of new health performance measures like those demonstrated in the case of lactation curves. The activities in step 7 will almost certainly require expert statistical assistance.

References

Allaki, FE. 2005. Théorie de la surveillance de la santé des populations. [in French, with English abstract]. Ph.D.-thesis, Université de Montréal, Canada. ISBN: 9780494180303. Available from: https://www.webdepot.umontreal.ca/Usagers/bigraspm/MonDepotPublic/these_theorie_surveillance.pdf

Albrech, G, S Freeman, and N Higginbotham. 1998. Complexity and human health : The case for a transdisiplinary paradigm. Culture Medicine Psychiatry. 22(1):55-92. p. 57.

Andersen, HJ and C Enevoldsen. 2004. Towards a better understanding of the farmer's management practices – the power of combining qualitative and quantitative data. Manuscript in Ph.D.-thesis: Andersen HJ. (2004) Rådgivning – Bevægelse mellem data og dialog. ISBN: 87-89795-81-4

Armitage, P and G. Berry. 1987. Statistical methods in medical research. 2nd Edition. Blackwell. P. 349.

Bramsen, J and K Nielsen. 2004 , Interaktiv benchmarking: med eksempler fra landbruget. [in Danish] Rapport / Fødevareøkonomisk Institut, no. 172, Department of Economics, University of Copenhagen, København.

De Vries, A, JK Reneau. 2010. Application of statistical process control charts to monitor changes in animal production systems. *J Anim Sci.* 88:11-24. doi: 10.2527/jas.2009-2622

Enevoldsen C. 1993. Sundhedsstyring i mælkeproduktionen. PhD-thesis [in Danish]. The Royal Veterinary and Agricultural University, Copenhagen, Denmark

Enevoldsen, C, JT Sørensen, I Thysen, C Guard, and YT Gröhn. 1995. A diagnostic and prognostic tool for epidemiologic and economic analysis of dairy herd health management. J Dairy Sci. 78:947-961.

Enevoldsen, C, J Hindhede, and T Kristensen. 1996. Dairy herd management types assessed from indicators of health, reproduction, replacement, and production. J Dairy Sci. 79:1221-1236.

Enevoldsen C. 1997a. Epidemiological considerations related to within herd multivariable modelling in herd health management. International Symposia on Veterinary Epidemiology and Economics (ISVEE) proceedings, ISVEE 8, Paris France.
Available from: http://www.sciquest.org.nz/node/62258

Enevoldsen, C. 1997b. Det israelske rådgivningskoncept – og en dansk oversættelse. 10 pp [in Danish]. In: Proceedings Danske Kvægfagdyrlægers Årsmøde (Danish Bovine Practitioner Seminar),Hindsgavl Slot, Middelfart, Denmark.

Gunnarsson S. 2006. The conceptualisation of health and disease in veterinary medicine. Acta Vet Scand. 48:20.

Hoerl, RW and AC Palm. 1992. Discussion: Integrating SPC and APC. Technometrics 34, pp. 268–272.

Houe, H, AK Ersbøll and N Toft. 2004. Introduction to veterinary epidemiology. 1st edt. Biofolia. ISBN : 9788791319211. p. 25.

Koutras, MV, S Bersimis and PE Maravelakis. 2007. Statistical process control using Shewhart control charts with supplementary runs rules. Methodol Comput Appl Probab. 9:207–224. Doi: 10.1007/s11009-007-9016-8

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

Kristensen, AR, E Jørgensen and N Toft. 2009. Herd Management Science. II. Advanced topics. University of Copenhagen, Faculty of Life Sciences. SL books. 329 pp. ISBN: 9788763460644

Kristensen, E and C Enevoldsen. 2008. A mixed methods inquiry: How dairy farmers perceive the value(s) of their involvement in an intensive dairy herd health management program. Acta Vet Scand. 50:50.

Krogh MA, Toft N and Enevoldsen C. 2011. Latent class evaluation of a milk test, a urine test, and the fat-to-protein percentage ratio in milk to diagnose ketosis in dairy cows. J Dairy Sci. 94(5):2360-2367.

Krogh MA and C Enevoldsen. 2006. Organizational and educational support to dairy herd health programs. International Symposia on Veterinary Epidemiology and Economics (ISVEE) proceedings, ISVEE 11: Proceedings of the 11th Symposium of the International Society for Veterinary Epidemiology and Economics, Cairns, Australia. p. 986

Krogh MA and C Enevoldsen. 2012a. A framework for integration of benchmarking and within-herd analysis in dairy herd management – analysis of lactation curves as a case. Manuscript

Krogh MA and C Enevoldsen. 2012b. A tool to detect rater-introduced bias in clinical ratings. Manuscript

Krogh, MA and C Enevoldsen. 2012c Evaluation of effects of disease control in a complex dairy herd health management program. Manuscript.

Krogstrup, HK. 2011. Kampen om evidens. Resultatmåling, effektevaluering og evidens [In Danish]. Hans Reitzel Forlag, Copenhagen. 170 pp. ISBN :978-87-412-5516-3

Kumbhakar, SC and CAK Lovell. 2000. Stochatic frotier analysis. Cambridge University Press. UK. ISBN: 0-521-48184-8. pp: 72-93

Lastein, DB, M Vaarst, and C Enevoldsen. 2009. Veterinary decision making in relation to metritis - a qualitative approach to understand the background for variation and bias in veterinary medical records. Acta Vet Scand 51:36. doi:10.1186/1751-0147-51-36

MacQueen, KM, E McLellan, DS Metzger, S Kegeles, RP Strauss, R Scotti, L Blanchard, and RT Trotter II. 2001 What is community? An evidence-based definition for participatory public health. American Journal of Public Health. 91(12):1929-1938. doi: 10.2105/AJPH.91.12.1929

Markusfeld, O. 1993. Epidemiological methods in integrated herd health programs. Acta Vet Scand, 1993; 89:61-67.

Nir-Markusfeld, O. 2003. What are production diseases, and how do we manage them? Acta vet. Scand., suppl. 98, 21-32.

Rickles, D, P Hawe, A Shiell. 2007. A simple guide to chaos and complexity. J Epidemiol Community Health. 61:933-937

SAS Institute Inc. 2011. SAS/QC 9.3 user's guide. Cary, NC: SAS Institute Inc. pp :953-982

SAS Institute Inc. 2010. Getting started with SAS® Text Miner 4.2. Cary, NC: SAS Institute Inc. pp: 1-4

Schwabe, CW, H Riemann and CE Franti. 1977. Epidemiology in veterinary practice. Lea & Febiger, Philadelphia. Pp. 303. ISBN: 0-8121-0573-7

Singer, JD and JB Willett. 2003. Applied longitudinal data analysis. Oxford University Press, New York. ISBN: 0-19-5152964

Stacey, RD. 2011. Complex responsive processes in organizations. Learning and knowledge creation in organizations. Routledge, London and New York. 258 pp. ISBN 0415249198

Stärk, KDC and MD Salman. 2001. Relationships between animal health monitoring and the risk assessment process. Acta Vet Scand, 42(Suppl 1):71-77. doi:10.1186/1751-0147-42-S1-S71

Thysen, I. 1993. Monitoring bulk tank somatic cell counts by a multi-process Kalman filter. Acta Agric Scand Sect A, Animal Sci. 43:58-64

Thysen, I and C Enevoldsen. 1994. Visual monitoring of reproduction in dairy herds. Prev Vet Med. 19: 189-202.

Thysen, I and C Enevoldsen. 2011. HerdView- a PC-program for interactive analysis of reproduction and health in dairy herds. Software homepage.
Available from: http://www.herdview.thysen.dk

Vaarst, M, B Paarup-Laursen, H Houe, C Fossing, and HJ Andersen. 2002. Farmers choice of medical treatment of mastitis in Danish dairy herds based on qualitative research interviews. J Dairy Sci. 85:992-1001

Wheeler, DJ. 2000. Understanding variation. The key to managing chaos. 2nd Edition. SPC Press, Knoxville, Tennessee, USA.

Wheeler, DJ. 2010. The right and wrong ways of computing limits. How does your software measure up? Quality Digest Daily, Jan. 7, 2010. Manuscript No. 205.
Available from: http://www.qualitydigest.com/inside/six-sigma-column/right-and-wrong-ways-computing-limits.html

Wheeler, DJ. 2011. Myths about process behavior charts. How to avoid some common obstacles to good practice. Quality Digest Daily, Sept. 6, 2011. Manuscript 232.
Available from: http://spcpress.com/pdf/DJW232.pdf

Østergaard, S., JF Ettema, AB Kudahl, and JT Sørensen. 2010. Development of a SIMHERD web application for herd health advisors –experiences and perspectives. Proceedings at Farm Animal Health Economics, Nantes, France, 14-15 january.
Available from: http://www.simherd.com/images/stories/simherd/development_of_a_simherd_web.pdf

Woodall, WH. 2000. Controversies and contradictions in statistical process control. Journal of Quality Technology, 32(4):341-350.

3.3 A tool to detect rater-introduced bias in clinical ratings

Mogens A. Krogh & Carsten Enevoldsen

Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Grønnegårdsvej 2, DK-1870 Frederiksberg C, Denmark

Manuscript as 'Short Communication'

A tool to detect rater-introduced bias in clinical ratings

M. A. Krogh[a*] & C. Enevoldsen[a]

[a]*Department of Large Animal Sciences*
Faculty of Health and Medical Sciences, University of Copenhagen,
Groennegaardsvej 2, DK-1870 Frederiksberg C, Denmark

[*] Corresponding author: email: mok@life.ku.dk (Krogh, M.A)

Abstract

Background: We suggest a 'screening test' to examine large data files with clinical ratings for the occurrence of rater-introduced bias prior to using the data for quantitative analyses. The test is based on a statistical model in which a well-standardized interval-scale outcome (for example, milk yield) is related to clinical ratings (for example, body condition scores) obtained from multiple contexts (for example, dairy herds).

Findings: 84,968 calvings from 279 herds, with subsequent body condition scores performed by 117 veterinarians within the first 21 days postpartum were analyzed with a multilevel random coefficient regression model. The model included an independent variable, where body condition score was centered within veterinarian. This is a so-called comparison effect to describe possible rater-introduced bias in the body condition scores. A highly significant comparison effect was found for second and older parities, indicating occurrence of possible rater-introduced bias in this large multi-herd data file.

Conclusions: A within-group centering technique (the comparison effect) appeared to be useful for discriminating between biased and unbiased clinical scores. In some cases, this test for bias should prohibit further analysis of the data and divert the focus of study to the calibration of raters or alternative study designs.

***Key words*:** bias, diagnostic test, large data files, standardization of ratings, body condition scores

Findings

Background

In clinical veterinary medicine, numerous diagnostic measurements are ratings of conditions that cannot be measured using standardized metric tools. It is often relevant to employ collections of ratings from multiple raters (registry data) for benchmarking or statistical analyses. Lastein et al. (2009) describe practicing cattle veterinarians' recording and use of a metritis score [1]. The authors demonstrate that the veterinarians' use of the metritis score (*ratings*) was very different from the intended use, even if detailed rating manuals were disseminated to veterinarians prior to the study. The ratings could be systematically different (*level-shift in scale*), or the rating of a subject could be affected by the subject's context (*relative rating).* Relative rating may occur if other clinical findings are incorporated into the score, or if the score is adjusted to the prognosis (feedback) [1]. Because relative rating will render interpretation across observation contexts (e.g., herds) virtually impossible, we must detect such a measurement error prior to the analysis and use of the data. If a systematic relationship exists between the clinical condition being studied (X) and some other condition measured with a completely objective scale (Y), then level-shift or relative rating caused by rater (R) can be detected by means of an appropriate statistical model. If the effect of X differs among different levels of R, then relative rating is likely. This is also known as a *comparison effect*. A main effect of R indicates level-shift and is not studied further because it is less complicated to detect and adjust for.

The objective of this study was to demonstrate a quantitative screening method to detect occurrence of relative ratings or comparison effects prior to the statistical analysis of large data files containing ratings from multiple raters.

Concepts and terms

To demonstrate our approach to identifying relative ratings, we used the well-established relation between a very well-standardized interval-scale outcome (milk production in energy-corrected milk (ECM)) and a widely used rating, the body condition score (BCS). The BCS is an ordinal-scale rating with symmetrically distributed values. Veterinarians will likely be able to rank cows correctly using the BCS because they typically rate several cows during a single herd visit, and are consequently able to compare the cows directly. However, it is less certain that several veterinarians are able to assign the same BCS to the same cow. This hypothesis is supported by Kristensen et al. (2006) who observed that within-rater agreement of BCS is higher than between-rater agreement [2]. Relative rating could occur if the veterinarian provided 'preferential treatment' according to some implicit characteristics of the cow (e.g., a special feed ration to particularly valuable cows). Vaarst et al. (2002) provides examples of this scenario in an udder health management context [3].

Materials

The data were extracted from the VPR platform [4]. The mean energy corrected milk (ECM, kg) between 9 and 92 days postpartum in individual cows was calculated as a mean of the milk yields from test days within this time period. The final data file consisted of 279 herds with 84,968 calvings, with subsequent BCS rating performed by a veterinarian within the first 21 days

postpartum. A total of 117 veterinarians observed and recorded the BCS of individual cows in the herds. Table 1 shows how veterinarians were distributed with regard to the herds.

Table 1. Distribution of veterinarians among herds and herds among veterinarians

Number of herds scored by one veterinarian (%)		
1-3 herds	4-6 herds	>6 herds
39 (33%)	35 (31%)	43 (36%)
Number of veterinarians in each herd (%)		
1-2 veterinarians	3-4 veterinarians	>4 veterinarians
173 (62%)	81 (29%)	25 (9%)

The mean BCS by veterinarian was in the interval between 2.72 and 3.69. The interquartile range was 0.26, indicating that the veterinarians' BCS means were quite similar in most cases. Similarly, BCS means were calculated at herd level and ranged from 2.57 to 3.75, with an interquartile range of 0.31. The herd-level mean of the daily ECM per cow between 9 and 92 days postpartum had a median value of 33.5 kg ECM. Upper and lower quartiles were 31.5 kg ECM and 35.5 kg ECM, respectively.

Statistical model

To demonstrate rater-introduced bias in our non-documented data, a multilevel random coefficient regression model was used. Consider an ordinary multilevel regression model as model 1, where y_{ij} is the outcome of cow *i* in herd *j*, and x_{ij} is a predictor of y_{ij} measured at cow level. ω_j is a random variable that accounts for herd *j*'s departure from the overall intercept, β_0. ε_{ij} is the random error term.

$$\begin{aligned} &y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij} \\ &\textit{where} \\ &\beta_{0j} = \beta_0 + \omega_j \\ &\omega_j \sim N(0, \sigma_\omega^2) \text{ and } \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \end{aligned} \tag{1}$$

In the following, only the first line of model 1 is presented because the rest of the model does not change. Let x_{ij} symbolize the individual effect of cow i in herd j. Let $\bar{x}_{.k}$ symbolize the effect of rater *k* as the mean of x within rater *k* and $(x_{ij} - \bar{x}_{.k})$ describe the *comparison effect* of rater *k*. We now suggest model 2 as a tool to answer the research question.

$$y_{ijk} = \beta_{0j} + (\beta_1 + \beta_2) x_{ij} - \beta_2 (x_{ij} - \bar{x}_{.k}) + \varepsilon_{ijk} \tag{2}$$

The parameters in model 2 can be interpreted as an effect related to the individual cow ($\beta_1 + \beta_2$), and as an effect that relates to this individual cow's standing as assessed by the rater (β_2). A possible effect of the rater on the clinical score will reveal itself by β_2 differing significantly from

0. Although model 2 can be re-parameterized to answer the question about a level shift in scale between raters, this was not done in the present study.

Statistical analysis

The data file was analyzed using a slightly modified version of model 2. We included the number of days postpartum that BCS was observed, and an interaction between days of observation postpartum and BCS to account for the biological changes due to fat mobilization early postpartum. Separate analyses were conducted for the first lactation, second lactation, and later lactations. BCS was grand-mean centered within parity groups by subtracting the mean from the individual BCS values. This technique eases interpretation of the parameter estimates related to BCS. Grand-mean centering does not influence other parameter estimates or variances [5]. All analyses were performed using SAS® PROC MIXED [6] with Maximum Likelihood estimation. Tests of parameter estimates were performed using the deviance test for tests of fixed effects between nested models.

Results

Table 2 summarizes parameter estimates for significant effects after the removal of non-significant variables from the models. In the analysis of first-parity cows, the comparison effects could be removed from the model ($P = 0.23$). In the analyses of second and later parities, the comparison effects were highly significant ($P < 0.001$). In the analysis of third or later parities, the effect of the interaction between BCS and days of observation postpartum and the main effect of BCS could both be removed ($P = 0.12$ and $P = 0.99$, respectively).

Table 2. Parameter estimates from models of energy-corrected milk from first, second, and later parities

Variable	Parameter estimate		
	Parity 1	Parity 2	Later parities
Intercept	28.0 ***	35.9 ***	36.8 ***
Body condition score (BCS) (centered), 1 to 5 scale	2.19 NT	1.00 NT	0.05 NS
Days postpartum at BCS recording (dpp_obs), interval 5 to 20 dpp_obs.	0.01 NT	0.02 NT	0.03 *
Interaction BCS × dpp_obs	-0.07 ***	-0.08 ***	-0.04 NS
Comparison effect	0.46 NS	2.16 ***	2.53 ***

* $P < 0.05$; *** $P < 0.001$; NT, not tested; NS, not significant ($P > 0.05$)

Discussion

In the analyses of the data file with clinical ratings, a significant comparison effect was observed for the second and later parities, but not for first-parity cows. The interpretation of these results is that

BCS was rater-biased for second and later parities. For parities > 2, the effect of the individual cows' BCS could be removed from the model. The interpretation is that the relative standing of a BCS within a single rater was more important than the absolute BCS. The main effect of BCS could not be removed from the model of the second parity group because of the significant interaction between BCS and the postpartum day at which a cow was rated. We can only guess about the practical reasons for the difference between first and later parities. Knowledge about BCS recordings at drying off (only relevant at second and later calvings) might have been used somehow when the veterinarians recorded BCS after calving. However, to study this hypothesis we obviously require additional data collection, which is beyond the scope of this study. In the BCS setting, some veterinarians could also recommend that cows with high BCS postpartum should be given special attention or special feeding supplements. Such actions, if effective, would also reveal themselves as comparison effects.

In this study, we have deliberately chosen the postpartum BCS instead of the metritis score or lameness score because we believe that it is unlikely that major actions are taken based on the BCS. Actions related to the metritis score, such as medical treatments, may be directly related to the metritis score, and the action taken may be veterinarian-specific [1]. However, if effective actions are taken based on BCS and revealed as a comparison effect, the data will be useless or even misleading for the estimation of relations between the ratings and a given outcome, or between the rating as outcome and some explanatory variable. In other words, focus should be diverted to calibration or the development of alternative study designs.

We could have used a cross-classified design [7] to account for the unbalanced distribution of number of herds per veterinarian; however, this would not correct the underlying problem of rater-specific misclassification of scores. We could also have specified a model that featured the rater as a fixed categorical effect and included the rater in an interaction term with BCS. A significant interaction would imply a relative rating. Although this approach will work when relatively few raters are being considered; it is likely to be problematic when many raters are considered, as in our case. In addition, partial confounding between herds and raters will pose additional problems regarding the interpretation of results from a fixed-effect model.

Based on the results in this study, we suggest that many studies based on non-documented data could benefit from initial investigations of the comparison effect. Burstein (1980) suggested that the comparison effect is an effect related to 'lack of knowledge' [8]. Based on our data, we find the BCS problematic for second and later parities, and we suggest that the comparison effect might be related to rater-introduced bias or rater-specific actions taken based on the clinical score. Hence, we require additional information, such as a standard for calibrating the crude scores or information about rater-specific actions, if we want to study the BCS in detail.

Conflict of interest statement

The authors declare that they have no conflicts of interest to report in this study.

Authors' contributions

MAK and CE contributed equally to the research hypothesis. Data preparation and data analysis and were done by MAK. Writing of the manuscript was done equally by MAK and CE.

Acknowledgements

We would like to thank the 117 Danish veterinarians (who conducted all of the BCS ratings used in this study) for their willingness to share their clinical recordings.

References

1. Lastein DB, Vaarst M, Enevoldsen C: **Veterinary decision making in relation to metritis – a qualitative approach to understand the background for variation and bias in veterinary medical records.** *Acta Vet Scand* 2009, **51**:36
2. Kristensen E, Dueholm L, Vink D, Andersen JE, Jakobsen EB, Illum-Nielsen S, Petersen FA, Enevoldsen C: **Within- and across-person uniformity of body condition scoring in Danish holstein cattle.** *J Dairy Sci* 2006, **89**: 3721-3728
3. Vaarst M, Paarup-Laursen B, Houe H, Fossing C, Andersen HJ: **Farmers' choice of medical treatment of mastitis in Danish dairy herds based on qualitative research interviews.** *J Dairy Sci* 2002, **85**:992-1001
4. Krogh MA, Toft N, Enevoldsen C: **Latent class evaluation of a milk test, a urine test, and the fat-to-protein percentage ratio in milk to diagnose ketosis in dairy cows**. *J Dairy Sci* 2011, **94:**2360-2367
5. Kreft IGG, de Leeuw J, Aiken LS: **The effect of different forms of centering in hierarchical linear models.** *Multivar Behav Res* 1995 **30(1)**:1-21
6. Littell RC, Miliken GA, Stroup WW, Wolfinger RD: *SAS for Mixed Models 2nd ed.* SAS Inst. Inc. Cary, NC. 2006.
7. Fielding A, Goldstein H: **Cross-classified and multiple membership structures in multilevel models: An Introduction and Review.** *Research Report no 791*, University of Birmingham, 2006.
8. Burstein, L: **The analysis of multilevel data in educational research and evaluation.** *Rev Res Educ* 1980. **8:**158-233

3.4 Latent class evaluation of a milk test, a urine test, and the fat percentage to protein percentage ratio in milk to diagnose ketosis in dairy cows

Mogens A. Krogh, Nils Toft and Carsten Enevoldsen

Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Grønnegårdsvej 2, DK-1870 Frederiksberg C, Denmark

Printed in J. Dairy Sci. 94 :2360–2367, 2011
doi: 10.3168/jds.2010-3816

J. Dairy Sci. 94:2360–2367
doi:10.3168/jds.2010-3816

Latent class evaluation of a milk test, a urine test, and the fat-to-protein percentage ratio in milk to diagnose ketosis in dairy cows

M. A. Krogh,[1] N. Toft, and C. Enevoldsen
Department of Large Animal Sciences, Faculty of Life Sciences, University of Copenhagen, Grønnegårdsvej 2, DK-1870 Frederiksberg C, Denmark

ABSTRACT

In this study, 3 commonly used tests to diagnose ketosis were evaluated with a latent class model to avoid the assumption of an available perfect test. The 3 tests were the KetoLac BHB (Sanwa Kagaku Kenkyusho Co. Ltd., Nagoya, Japan) test strip that tests milk for β-hydroxybutyrate, the KetoStix (Bayer Diagnostics Europe Ltd., Dublin, Ireland) test strip that tests urine for acetoacetate, and the fat-to-protein percentage ratio (FPR) in milk. A total of 8,902 cows were included in the analysis. The cows were considered to be a random sample from the population of Danish dairy cattle under intensive management, thus representing a natural spectrum of ketosis as a disease. All cows had a recorded FPR between 7 and 21 d postpartum. The KetoLac BHB recordings were available from 2,257 cows and 6,645 cows had a KetoStix recording. The recordings were analyzed with a modified Hui-Walter model, in a Bayesian framework. The specificity of the KetoLac BHB test and the KetoStix test were both high [0.99 (0.97–0.99)], whereas the specificity of FPR was somewhat lower [0.79 (0.77–0.81)]. The best sensitivity was for the KetoStix test [0.78 (0.55–0.98)], followed by the FPR [0.63 (0.58–0.71)] and KetoLac BHB test [0.58 (0.35–0.93)].
Key words: ketosis, diagnostic test evaluation, latent class model, gold standard

INTRODUCTION

Ketosis is a common disease in adult cattle. It typically occurs in dairy cows in early lactation and is clinically characterized by nonspecific signs such as anorexia, milk loss, central nervous symptoms, and loss of body fat. The pathogenesis of ketosis involves a complex set of manifestations of excessive mobilization of body reserves of stored energy, mostly fat, due to a lack of balance between the cow's energy demand for milk production and her energy intake. The cause of low energy intake can be related to poor quality of feedstuff, including ketogenic substances in the feed, other diseases like milk fever (post-parturient hypocalcaemia), or displaced abomasums. However, increased body condition score at calving may also in itself decrease feed intake (Broster and Broster, 1998).

A range of biochemical tests are available to diagnose ketosis. Measurements of BHBA in serum or milk and acetoacetate (**AcAc**) in urine probably are the most common. Several studies have used serum concentrations of BHBA of 1400 μmol/L as the suggested cut-off or even gold standard to define ketosis, but several cut-offs for the available tests have been suggested (Geishauser et al. 1998; Carrier et al., 2004; Iwersen et al. 2009). Previous studies on the accuracy (sensitivity and specificity) of available tests for ketosis that can be used in the herd (cow-side test) have primarily used serum levels of BHBA as the gold standard. However, BHBA and AcAc measurements vary for several reasons. For example, diurnal variation (Nielsen et al., 2003), high levels of somatic cells in the milk that will give false-positive results of a BHBA test on milk (Jeppesen et al., 2006), temperature (Geishauser et al., 2000), time of reading of urine sticks that influences color changes (Jeppesen et al., 2006), and pure measurement error. Fat-to-protein percentage ratio (**FPR**) in the milk has also been suggested as an indication of ketosis (Duffield et al., 1997; Čejna and Chlàdek, 2005). Excessive mobilization of fat will be reflected in an elevation of fat percentage. Because protein percentage is rather stable and cow-specific, FPR should be an indicator of ketosis that is adjusted for a cow effect. Although neither of the above tests can be considered perfect, they are all relevant tools in the dairy herd from a practical point of view and evaluation of their performance under field conditions is needed. Also, neither level of BHBA in serum, nor any other tests are appropriate to define a gold standard to describe the dynamic level of metabolic stress the cow is exposed to. The term gold standard should only be applied to diagnostic tests that have a sensitivity and specificity of 1. However, as an alternative, the term criterion standard is adopted

Received September 13, 2010.
Accepted February 1, 2011.
[1]Corresponding author: mok@life.ku.dk

by the American Medical Association and defined as a method having an established or widely accepted accuracy for determining a diagnosis, providing a standard to which a new screening or diagnostic test can be compared. Thus, previously obtained estimates of test accuracy for ketosis tests may be biased, due to the misclassification bias, which occurs from using a less than perfect test in a testing scheme to define cases and non-cases (Nielsen and Toft, 2002). Cow-side ketosis tests are ideally tools to detect ketosis at a stage where the possibly diffuse clinical symptoms are not yet present and not just to verify a clinical suspicion. This provides the practitioner the possibility to intervene efficiently. Thus, a practically relevant disease definition relates to cows in all stages of ketosis, rather than one subject to the selection bias imposed by a classification scheme using (e.g., serum as a gold standard or clinical symptoms).

Under certain conditions, latent class analysis (**LCA**) can be used to estimate the sensitivity (**Se**) and specificity (**Sp**) of diagnostic tests without the assumption of one being a gold standard (Toft et al., 2007a). The basic LCA relies on what is generally referred to as the Hui-Walter paradigm (Hui and Walter, 1980): 2 or more tests must be evaluated in 2 or more subpopulations with different prevalence of the disease, the tests must have constant Se and Sp across the populations, and the tests must be conditionally independent, given disease status. The LCA methods have gained increased acceptance as a means of evaluating diagnostic tests for infectious diseases and are now endorsed by the OIE (World Organization for Animal Health) as an alternative to classic test evaluations in the OIE fitness for purpose concept (OIE, 2010).

The objective of this study was to compare the test performance of a milk-based cow-side BHBA test, a urine-based cow-side AcAc test, and the FPR in milk as tools to diagnose ketosis without the assumption of an available gold standard.

MATERIALS AND METHODS

Diagnostic Tests

The KetoLac BHB (Sanwa Kagaku Kenkyusho Co. Ltd., Nagoya, Japan; marketed as KetoTest in the United States) test is a dip-stick that measures the milk contents of BHBA on a semiquantitative scale. In this study we used the cut-off $\geq$200 μmol/L of BHBA in milk, recommended by the manufacturer, to define a positive test. Tests were performed as described by the manufacturer. However, variation in the procedure between veterinarians might exist (such as preceding testing for mastitis).

The KetoStix (Bayer Diagnostics Europe Ltd., Dublin, Ireland) is a dip-stick that measures the AcAc content in the urine on a semiquantitative scale. A concentration of 4 mmol/L (moderate) or above of AcAc in urine was chosen as the cut-off for a positive KetoStix test (Carrier et al., 2004). Tests were performed as described by the manufacturer on urine collected as catheter sampling, spontaneous urination, or urination induced by manual manipulation the distal urethra in the vagina.

The FPR was calculated based on 1 of the 11 annual milk recordings from the Danish milk control program (RYK, 2010). Fat-to-protein percentage ratio values above 1.5 were defined as test positive for ketosis, as suggested by Čejna and Chlàdek (2005).

Data Collection

Data were extracted from the Veterinary Production and Consultancy (**VPR**) database, which is a subset of the National Danish Cattle Database (Krogh and Enevoldsen, 2006), on June 7, 2009. The VPR database consists of herds, where the veterinarians perform systematic clinical recordings on well-defined groups of cows. At least 12% of the Danish dairy herds have their cows routinely tested in early lactation for ketosis by a veterinarian. The veterinarians voluntarily submit cow-level registrations to the VPR database. In most Danish dairy herds, milk yield, fat percentage, and protein percentage are measured at up to 11 annual test days. The primary inclusion criteria for our study were that cows had been tested for ketosis with either the KetoLac BHB or the KetoStix between 7 and 21 d postpartum and had a milk test-date recording 1 d before the day of the ketosis test. The restriction with a minimum of 7 d postpartum was imposed because some cows might still have some colostrum in the milk up to this time point and colostrum will affect the measurement of fat content in the milk. Milk recordings at or after the day of testing for ketone bodies were also excluded because test results might have induced treatments of ketosis (usually glucocorticoids parentally or propylene glycol given orally; Radostits et al., 2000), which might affect FPR. Initial analyses of the data suggested that treatments could affect the milk composition very quickly. Records of ketosis treatment were available but not used, except for the initial data analyses. Out of 141,133 ketosis tests available from individual cows, a total of 8,902 cows were selected based on the milk test date. Of these, 2,257 cows were examined with the KetoLac BHB test and 6,645 cows with the KetoStix test. The KetoLac BHB tests were recorded from February 2004 to April 2009. The KetoStix tests were recorded between February

2003 and April 2009. Both tests were recorded during all seasons of the year.

Statistical Model

To estimate the Se and Sp of the KetoStix, KetoLac BHB, and FPR in the absence of a gold standard, we applied LCA, using a modified version of the Hui-Walter model (Hui and Walter, 1980). As previously mentioned, the Hui-Walter model assumes that prevalences are different across subpopulations, Se and Sp are constant (within tests) across the subpopulations, and tests are conditionally independent, given disease status. These assumptions imply that for each subpopulation a 2 × 2 table of tabulated paired test results is required. For each combination of test results, the probability of that particular result can then be expressed in terms of Se, Sp, and prevalence (**p**). For the ith subpopulation, the joint probability (Pr) of test 1 (T_1) positive and test 2 (T_2) negative can be expressed as $Pr(T_1+,T_2-) = Pr(T_1+,T_2-|D+)Pr(D+) + Pr(T_1+,T_2-|D-)Pr(D-) = Pr(T_1+|D+)Pr(T_2-|D+)Pr(D+) + Pr(T_1+|D-)Pr(T_2-|D-)Pr(D-) = Se_1(1 - Se_2)p_i + (1 - Sp_1)Sp_2(1 - p_i)$, where, for example, $Pr(T_1+,T_2-|D-)$ means the joint probability of test 1 being positive (T_1+) and test 2 being negative (T_2-), conditional on (|), the test subject being truly disease free (D−). The first equation takes into account that, although unobserved, the true state of each animal is either diseased or not diseased. The second equation uses the assumption of conditional independence, given disease status, to express the joint probability as a product. The final equation merely rewrites the expression using the conventional terms of Se, Sp, and p. The above expression uses 5 terms (Se_1, Sp_1, Se_2, Sp_2, and p_i); the other 3 probabilities in the 2 × 2 table may be expressed similarly using the same 5 terms. Thus, each 2 × 2 table of data representing a subpopulation can be expressed using 5 parameters, while providing only 3 degrees of freedom in the data. Hence, identifiability of the original 2-test, 2-population Hui-Walter model was given by the assumptions of conditional independence, given disease status; constant test properties across populations; and the prevalences truly differing between populations. These assumptions ensured that for 2 populations, 2 times 3 (i.e., 6 degrees of freedom) were available and only 6 parameters were used, because adding a population only meant adding 1 new parameter (the prevalence of the population). We shall address the validity of these assumptions for our model further in the discussion.

The original Hui-Walter 2-test, 2-population model was extended as illustrated in Figure 1. Thus, the Se and Sp of the KetoLac BHB test were estimated from subpopulation 1, 2, and 3; the Se and Sp of the KetoStix from subpopulation 4, 5, and 6; and the Se and Sp of the FPR were estimated using all 6 subpopulations. The 6 subpopulations were defined as follows: subpopulation 1 and 4 are first-parity cows, subpopulation 2 and 5 are second-parity cows, and subpopulation 3 and 6 are older than second parity. Essentially, we have a 2-test, 3-population model for each BHBA and FPR test combination, giving us 9 degrees of freedom, while only requiring 7 parameters. Furthermore, because the FPR is assumed to be constant across all populations, the full model uses only 12 parameters with 18 degrees of freedom.

We chose a Bayesian estimation approach (Branscum et al., 2005) implemented in OpenBUGS software (Thomas et al., 2006), which uses a Markov Chain Monte Carlo (MCMC) sampling algorithm to obtain a Monte Carlo (MC) sample from the posterior distribution. The first 10,000 MC samples were discarded as a burn-in to allow convergence, and the following 90,000 iterations were used for posterior inference. Convergence of the MCMC chain after the initial burn-in was assessed by visual inspection of the time-series plots of selected variables as well as Gelman-Rubin diagnostic plots using 3 sample chains with different initial values (Toft et al., 2007b).

In Bayesian analysis, all parameters are modeled using distributions, where prior distributions are provided to reflect what is known about the tests. However, as we used a model capable of estimating all parameters from data alone, we chose priors in the shape of uniform distributions on the interval between 0 and 1, modeled using the Beta(1,1) distribution for all parameters (i.e., Se and Sp of the 3 tests and prevalence of the 6 populations). This implies that posterior estimates are comparable to those obtained from a maximum likelihood analysis. However, we avoided the questionable assumptions about asymptotic normality of the estimates (Toft et al., 2005).

Based upon the estimates of Se and Sp of the 3 tests, we calculated the combined Se and Sp of FPR and the KetoLac BHB/KetoStix tests. An assumption for this calculation is the conditional independence of the tests (Gardner et al. 2000). Due to this issue the combined test performance of the KetoLac BHB and KetoStix test is not calculated. Both a parallel and a serial testing scheme were demonstrated. Parallel testing means that the combined test is only considered negative if, and only if, both tests are negative. The parallel diagnostic Se (Se_{par}) and parallel diagnostic Sp (Sp_{par}) are calculated as $Se_{par} = 1 - (1 - Se_1)(1 - Se_2)$ and $Sp_{par} = Sp_1Sp_2$, where Se_1 and Sp_1 are the diagnostic properties of test 1 and Se_2 and Sp_2 are the diagnostic proper-

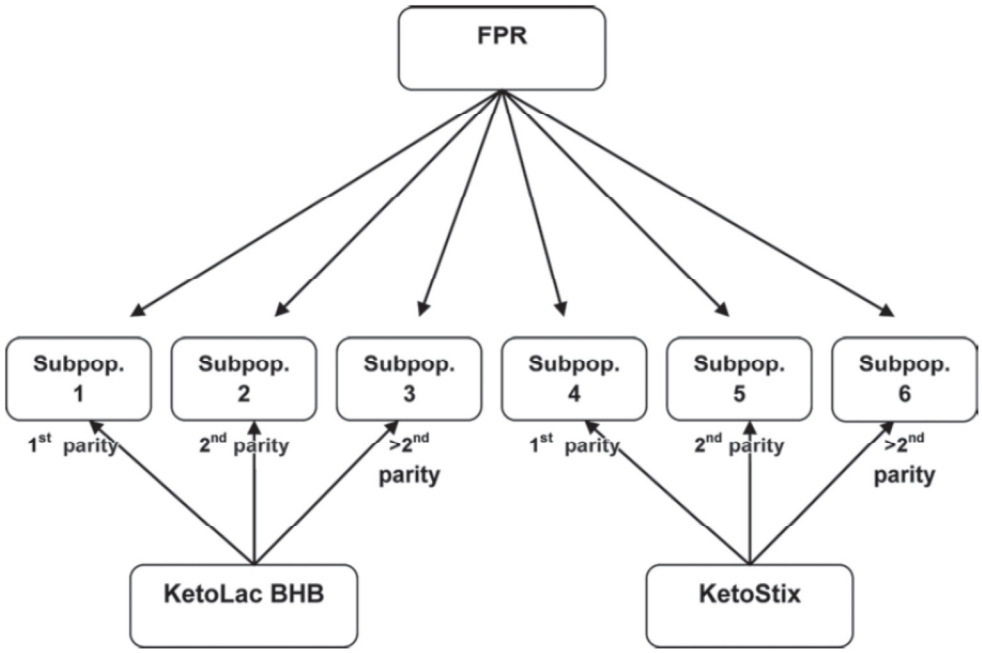

Figure 1. Study design that reflects the applied Bayesian latent class model. The data are divided into population 1, 2, and 3 versus 4, 5, and 6 based on type of cow-side ketosis test [KetoLac BHB (Sanwa Kagaku Kenkyusho Co. Ltd., Nagoya, Japan) or KetoStix (Bayer Diagnostics Europe Ltd., Dublin, Ireland)]. For each type of ketosis test, these recordings are divided into parity groups. Fat-to-protein percentage ratio (FPR) is measured on all of the cows. Subpop. = subpopulation.

ties of test 2. With parallel testing, the Se will often be higher than for any of the tests alone, whereas the Sp often will be substantially lower. In a serial testing scheme, both tests need to be positive for the combined test to be positive. The serial diagnostic Se (Se_{ser}) and parallel diagnostic Sp (Sp_{ser}) are calculated as $Se_{ser} = Se_1Se_2$ and $Sp_{ser} = 1 - (1 - Sp_1)(1 - Sp_2)$, where Se_1 and Sp_1 are the diagnostic properties of test 1 and Se_2 and Sp_2 are the diagnostic properties of test 2.

Posterior inference was based on medians and 95% posterior credibility intervals (PCI, the Bayesian analog of a confidence interval) of the prevalence in the 6 subpopulations, the Se and Sp of the 3 tests, and the combined Se and Sp of the tests.

RESULTS

To satisfy the modeling assumptions, data were divided into the subpopulations defined by Figure 1. Summary statistics of the number of cows, the number of herds, and the median number of cows from each herd, as well as the percentage of breeds in each population, are given in Table 1.

It is apparent that most herds only contributed with few cows to the analysis. The distribution of breeds in-

Table 1. Summary statistics of the sample population: number of cows, number of herds, median number of cows within herds, and breed distribution in parity groups within type of ketone body test (KetoLac BHB[1] or KetoStix[2])

	KetoLac BHB			KetoStix		
Item	First parity	Second parity	>Second parity	First parity	Second parity	>Second parity
Cows (n)	795	620	842	2,552	1,895	2,198
Herds (n)	140	127	150	254	248	255
Cows with ketone tests in herd (median n)	3	3	3	6	5	5
Red Danish cattle (%)	14	12	11	8	7	8
Danish Holstein (%)	60	60	57	73	76	72
Danish Jersey (%)	19	21	25	10	8	10

[1]Sanwa Kagaku Kenkyusho Co. Ltd., Nagoya, Japan.
[2]Bayer Diagnostics Europe Ltd., Dublin, Ireland.

Table 2. Cross-tabulation of the dichotomized test results for fat-to-protein percentage ratio (FPR; cut-off: >1.5), KetoLac BHB[1] (cut-off: ≥200 μmol/L), and KetoStix[2] (cut-off: ≥4 mmol/L), stratified by the use of either KetoLac BHB or KetoStix and, subsequently, into parity groups

Group	Ketone test	FPR +	FPR −
KetoLac BHB			
Subpopulation 1 (first parity)	KetoLac BHB +	21	24
	KetoLac BHB −	202	548
Subpopulation 2 (second parity)	KetoLac BHB +	19	9
	KetoLac BHB −	100	492
Subpopulation 3 (>second parity)	KetoLac BHB +	46	39
	KetoLac BHB −	176	581
KetoStix			
Subpopulation 4 (first parity)	KetoStix +	155	64
	KetoStix −	600	1,733
Subpopulation 5 (second parity)	KetoStix +	67	72
	KetoStix −	357	1,399
Subpopulation 6 (>second parity)	KetoStix +	197	139
	KetoStix −	405	1,457

[1]Sanwa Kagaku Kenkyusho Co. Ltd., Nagoya, Japan.
[2]Bayer Diagnostics Europe Ltd., Dublin, Ireland.

dicates that the KetoStix test is more frequently used to test Danish Holstein cows than the KetoLac BHB test.

The cross-tabulation of the test results dichotomized by the selected cut-offs and stratified by ketone test and parity group are given in Table 2. Based on the data in Table 2, the Se and Sp of the tests and the prevalences in the 6 subpopulations were estimated (Table 3). Slight differences in the parity-specific prevalences of ketosis were present for the populations defined by their use of KetoLac BHB or KetoStix. The overall prevalences of ketosis in the populations defined by the use of KetoLac BHB or KetoStix were 0.10 and 0.12, respectively.

Table 4 shows the results from the combination of tests. Applying the KetoLac BHB test and the FPR test or the KetoStix and FPR in parallel substantially increased the Se, whereas only a minor decrease occurred in the Sp. A serial testing scheme will decrease the Se of the combined test substantially, whereas the Sp here was around 1.

DISCUSSION

The LCA creates a probabilistic disease definition based on the cross-classified test results. Given the tests applied in this study, the definition of ketosis in this study will be a condition between 7 and 21 d postpartum that manifests itself with either elevated levels of ketone bodies in milk or urine, or elevated fat content in milk. These indicators are all very likely to be manifestations of excessive mobilization of fat.

Table 3. Medians and 95% posterior credibility intervals (PCI) from the latent class model using data from Table 2

Parameter[1]	Median	95% PCI
KetoLac BHB		
First parity, prevalence	0.08	0.04–0.17
Second parity, prevalence	0.06	0.03–0.10
>Second parity, prevalence	0.15	0.08–0.23
KetoStix		
First parity, prevalence	0.10	0.07–0.16
Second parity, prevalence	0.08	0.05–0.12
>Second parity, prevalence	0.18	0.14–0.26
$Se_{KetoLac\ BHB}$[2]	0.58	0.35–0.93
$Sp_{KetoLac\ BHB}$[2]	0.99	0.97–0.999
Se_{FPR}[3]	0.63	0.58–0.71
Sp_{FPR}[3]	0.79	0.77–0.81
$Se_{KetoStix}$[4]	0.78	0.55–0.98
$Sp_{KetoStix}$[4]	0.99	0.98–0.999

[1]KetoLac BHB (Sanwa Kagaku Kenkyusho Co. Ltd., Nagoya, Japan); KetoStix (Bayer Diagnostics Europe Ltd., Dublin, Ireland); Se = sensitivity; Sp = specificity; FPR = fat-to-protein percentage ratio.
[2]Cut-off: ≥200 μmol/L.
[3]Cut-off: >1.5.
[4]Cut-off: ≥4 mmol/L.

Table 4. Medians and 95% posterior credibility intervals (PCI) of the combined test performance of KetoLac BHB and KetoStix with fat-to-protein percentage ratio (FPR) in serial (ser) and parallel (par)[1]

Item	Se_{par}	Sp_{par}	Se_{ser}	Sp_{ser}
KetoLac BHB-FPR	0.85 (0.75–0.98)	0.77 (0.75–0.80)	0.37 (0.22–0.60)	0.99 (0.99–1.00)
KetoStix-FPR	0.92 (0.83–0.99)	0.78 (0.76–0.81)	0.49 (0.34–0.64)	0.99 (0.99–1.00)

[1]KetoLac BHB (Sanwa Kagaku Kenkyusho Co. Ltd., Nagoya, Japan); KetoStix (Bayer Diagnostics Europe Ltd., Dublin, Ireland); Se = sensitivity; Sp = specificity.

We used LCA to estimate the sensitivity and specificity of KetoLac BHB, KetoStix, and FPR as tests for ketosis, without the assumption of an available gold standard test. The analysis showed that no real differences existed in median estimates of the Sp of the KetoLac BHB and KetoStix, but both were more specific than FPR. The KetoStix had the best median Se, followed by FPR, and KetoLac BHB, with noticeable differences in median estimates between all 3 tests. Combining KetoLac BHB and FPR or KetoStix and FPR in parallel substantially increased the Se of the test and decreased Sp, whereas the serial test combination substantially decreased Se and increased Sp.

One of the underlying assumptions in the LCA is a direct association between the (latent) disease state and the diagnostic tests included in the analysis. This means that it, to some extent, is the diagnostic tests under evaluation that define our disease. Our data did not allow us to estimate the relation between the latent disease definition derived from this study and the commonly used criterion standard of serum BHBA concentration of 1,400 μmol/L (Radostits et al., 2000). However, it is possible to compare the results of other diagnostic test evaluations to our findings. Using a criterion standard of 1,400 μmol/L in serum and a cut-off of ≥200 μmol/L for the KetoLac BHB test, Carrier et al. (2004) found Se = 0.27 and Sp = 0.99; Geishauser et al. (2000) found Se = 0.59 and Sp = 0.90; and Iwersen et al. (2009) found Se = 0.30 with a 95% CI (0.07; 0.65) and Sp = 0.98. Another study by Geishauser et al. (1998) with the same cut-off but with a criterion standard of 1,200 μmol of BHBA/L in serum, found Se = 0.45 and Sp = 0.97.

Carrier et al. (2004) also studied the KetoStix using serum contents of BHBA of 1,400 μmol/L, as a case definition using our selected cut-off of 4 mmol/L and found Se = 0.49 and Sp = 0.99; Iwersen et al. (2009) found Se = 0.67 and Sp = 1.00.

Duffield et al. (1997) have published Se and Sp of the protein-to-fat ratio (1/FPR) based on a criterion standard of serum BHBA at 1,200 μmol/L: Se = 0.22 and Sp = 0.85. The rather low Se in this study could be due to the study population (entire lactation) and a rather long time span between the 2 tests.

Overall, the LCA results estimated in this study are similar to the ones based on selected cut-offs of BHBA serum levels. However, using LCA, the estimates of test accuracy are not subject to the misclassification bias imposed by using an imperfect test as a gold standard, nor will tests be punished by correctly identifying true positives or negatives, which are misclassified by the imperfect gold standard, and thereby underestimate or overestimate the Se and Sp of the new test. Even though the results we find with the LCA are similar to the ones found in gold standard studies, we do not suggest that the gold standard studies are not subject to misclassification bias. Our findings may only imply that the amount of misclassification bias most likely is not large enough to substantially change the final result. Also, most other studies have been performed in small/medium-scale experiments, which also tend to eliminate important covariates, thereby providing a better result of test performance than can be found under field conditions.

The data used in our study were collected from a large number of different herds, with different management levels, seasons, and years, different breeds, and distribution of parities. Essentially, we obtained a random sample because the cows were selected based on their milk recording data and calving dates. Thus, the cases in our data can be considered a random sample of naturally affected cows, representing ketosis as it is manifest in intensive dairy systems with intensive veterinary services. Data were collected over several years, which should remove possible systematic effects of years and seasons. Thus, the estimates for Se and Sp of the 3 tests obtained in this study reflect the values of these tests when applied under field conditions in the intensive Danish dairy population.

Our motivation for using the FPR was two-fold: FPR linked the KetoLac BHB and the KetoStix because no cows were tested using both, and FPR is freely available. Despite the FPR clearly being less accurate than the other cow-side tests due to a low Sp, it may still have value for establishing and monitoring the herd prevalence of ketosis. Because the milk recordings are already being performed, this prevalence could work as an effective and cheap monitoring tool for ketosis on

herd level, given that the sampling scheme and the Se and Sp of FPR is taken into account. In the near future, in-herd analysis of fat and protein percentage may be available, thus increasing the interest in the FPR as a useful monitor for ketosis on individual cow level. Because the Sp of the FPR test is low (0.79), Se is 0.63, and the prevalence of ketosis is below 20%, numerically more false positives than true positives will occur. Applying an additional test, such as KetoStix, in a parallel testing scheme will not improve this, but more of the true positives will be detected (Se = 0.92). Using KetoStix in a serial testing scheme will improve the overall Sp but the overall Se will drop to 0.49. Application of a routine testing scheme based on FPR alone or in combination with other tests should take into account what the herd manager wants to achieve by using these tests, the cost related to testing, the negative effects of having a false-negative result (diseased but untreated), the cost of treating cows that are not diseased (false negatives), and possible positive or negative effects of treatment.

The overall median prevalences of ketosis in the 2 populations defined by the use of KetoLac BHB or KetoStix were 0.10 and 0.12, respectively. Prevalence estimates from the parity groups tested with the KetoStix were all larger than for the parity groups tested with the KetoLac BHB. However, the same pattern of second-parity cows having the lowest prevalence, followed by first-parity cows, and with old cows having the highest prevalence was found for both sets of subpopulations. Other studies (Dohoo and Martin, 1984; Nielen et al., 1994) have found prevalence estimates of ketosis similar to what we have found here. The difference in overall prevalence may be attributed to sampling variation or systematic differences in measurement protocols. However, it may also indicate that measurements of AcAc in urine and BHBA in milk are associated with different stages of the disease process of ketosis.

It is a requirement of the LCA model assumptions that prevalences differ between subpopulations. The effects on model performance of differences in prevalences in the subpopulations have been studied by Toft et al. (2005) on simulated data. The overall conclusions were that if the differences in prevalences in the subpopulations are small (<10%), then the uncertainty and bias related to the estimates of sensitivity and specificity increase for a given sample size. However, considering the true overall prevalence in our study, it would be difficult to increase the differences between populations while ensuring the other assumptions of the Hui-Walter model. Furthermore, in our model, we use essentially 3 populations for each of the BHBA tests and 6 for the FPR.

We chose to split into subpopulations based on parity. Although this is generally not recommended for infectious diseases, where it most likely will cause differences in test properties, it is not considered a problem for this analysis. The BHBA contents in milk or the urine contents of AcAc will reflect the underlying pathophysiological process within the cow. It is unlikely that this pathophysiological process should be different for different parity groups. It is well known that the milk production increases with increasing parity group. However, there is no reason why (or evidence suggesting) the ratio between fat and protein in milk should be different for different parities and, hence, provide a different substrate for the growth of the calf, which must be considered the basic biological objective of milk production.

The concept of conditional independence between tests, given disease status, can best be understood by considering the opposite: conditional dependence between tests, given disease status. This often occurs when tests are based on the same phenomenon (e.g., 2 ELISA tests to detect antibodies): assuming that the true disease status is known (e.g., the test subject is truly infected), then knowing that test 1 is positive (or negative) will most likely influence the belief in a positive test result for test 2, because now additional information considering the level of antibodies or possible cross-reactions can be incorporated in the belief of a test result of test 2, when applied to the known truly infected animal. From a biological point of view, it is hard to imagine that the KetoStix test for AcAc in urine and the FPR in milk should be conditionally dependent, given disease status, because they measure different substances in different body fluids from the cow. The conditional dependence or independence of the KetoLac BHB test for BHBA in milk and the KetoStix test for AcAc in urine is irrelevant because no cows were tested with both tests. It could be hypothesized that the excretion of BHBA in milk is linked to the fat molecules in the milk, so that higher fat percentage in the milk will give higher concentrations of BHBA in milk. This, however, is not possible to test using the available data.

CONCLUSIONS

Using LCA, we were able to estimate the Se and Sp of 3 tests for ketosis without the assumption of a gold standard. This enabled us to use a large data set with a random sample of naturally occurring stages of ketosis similar to what must be expected in the population and context where the tests are to be used. Thereby, we avoided the misclassification bias often seen in tra-

ditional test evaluations due to the use of an imperfect test or a selection scheme to establish cases and controls. Our findings were that KetoStix had the highest sensitivity and specificity: 0.78 and 0.99, respectively. The KetoLac BHB test had a similarly high specificity (0.99) but the lowest sensitivity (0.58). The FPR had a marginally higher sensitivity (0.63) than the KetoLac BHB and a substantially lower specificity (0.78).

REFERENCES

Branscum, A. J., I. A. Gardner, and W. O. Johnson. 2005. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. Prev. Vet. Med. 68:145–163.

Broster, W. H., and V. J. Broster. 1998. Body score of dairy cows. J. Dairy Res. 65:155–173.

Carrier, J., S. Stewart, S. Godden, J. Fetrow, and P. Rapnicki. 2004. Evaluation and use of three cowside tests for detection of subclinical ketosis in early postpartum cows. J. Dairy Sci. 87:3725–3735.

Čejna, V., and G. Chládek. 2005. The importance of monitoring changes in milk fat to milk protein ratio in Holstein cows during lactation. J. Cent. Eur. Agric. 6:539–546.

Dohoo, I. R., and S. W. Martin. 1984. Subclinical ketosis: Prevalence and associations with production and disease. Can. J. Comp. Med. 48:1–5.

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

Gardner, I. A., H. Stryhn, P. Lind, and M. T. Collins. 2000. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. Prev. Vet. Med. 45:107–122.

Geishauser, T., K. Leslie, D. Kelton, and T. Duffield. 1998. Evaluation of five cowside tests for use with milk to detect subclinical ketosis in dairy cows. J. Dairy Sci. 81:438–443.

Geishauser, T., K. Leslie, J. Tenhag, and A. Bashiri. 2000. Evaluation of eight cow-side ketone tests in milk for detection of subclinical ketosis in dairy cows. J. Dairy Sci. 83:296–299.

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

Iwersen, M., U. Falkenberg, R. Voigtsberger, D. Forderung, and W. Heuwieser. 2009. Evaluation of an electronic cowside test to detect subclinical ketosis in dairy cows. J. Dairy Sci. 92:2618–2624. doi:10.3168/jds.2008-1795.

Jeppesen, R., J. M. D. Enemark, and C. Enevoldsen. 2006. Ketone body measurement in dairy cows. Reference OS43–2 in Proc. 24th World Buiatrics Congress, Nice, France. World Assoc. Buiatrics, Vienna, Austria.

Krogh, M. A., and C. Enevoldsen. 2006. Organizational and educational support to dairy herd health programs. Page 133 in Proc. Int. Soc. Vet. Epidemiol. Econ., Queensland, Australia. Accessed Dec. 12, 2008. http://www.sciquest.org.nz/crusher_download.asp?article=10003577.

Nielen, M., M. G. A. Aarts, A. G. M. Jonkers, T. Wensing, and Y. H. Schukken. 1994. Evaluation of two cowside tests for the detection of subclinical ketosis in dairy cows. Can. Vet. J. 35:229–232.

Nielsen, N. I., K. L. Ingvartsen, and T. Larsen. 2003. Diurnal variation and the effect of feed restriction on plasma and milk metabolites in TMR-fed dairy cows. J. Vet. Med. A Physiol. Pathol. Clin. Med. 50:88–97.

Nielsen, S. S., and N. Toft. 2002. Optimisation of the validity of ELISA and faecal culture tests for paratuberculosis: Selection of population or correction by population characteristics? Pages 400–405 in Proc. 7th Int. Colloq. Paratuberculosis, June 11–14, 2002, Bilbao, Spain. International Association for Paratuberculosis, Inc., Madison, WI.

OIE. 2010. Principles and methods of validation of diagnostic assays for infectious diseases. Chapter 1.1.4/5. Pages 1–18 in OIE Terrestrial Manual 2010. Accessed March 8, 2011. http://www.oie.int/fileadmin/Home/eng/Health_Standards/tahm/1.1.04_VALID.pdf.

Radostits, O. M., C. C. Gay, D. C. Blood, and K. W. Hinchcliff. 2000. Ketosis in ruminants. Pages 1452–1456 in Veterinary Medicine: A Textbook of the Diseases of Cattle, Sheep, Pigs, Goats and Horses. 9th ed. O. M. Radostits, C. C. Gay, D. C. Blood, and K. W. Hinchcliff, ed. Elsevier Health Services. Harcourt Publishers Limited, London, UK.

RYK. 2010. Service Technology Health. RYK–Livestock Registration and Milk Recording. Agro Food Park, Udkaersvej 15, DK-8200 Aarhus N, Denmark. Accessed Sep. 11, 2010. http://www.landbrugsinfo.dk/Kvaeg/RYK/Sider/RYKfolder_english.mht.

Thomas, A., B. O'Hara, U. Ligges, and S. Sturtz. 2006. Making BUGS open. R News 6/1:12–17. ISSN 1609–3631. Accessed March 8, 2010. http://cran.r-project.org/doc/Rnews/Rnews_2006-1.pdf.

Toft, N., J. Åkerstedt, J. Tharaldsen, and P. Hopp. 2007a. Evaluation of three serological tests for diagnosis of Maedi-Visna virus infection using latent class analysis. Vet. Microbiol. 120:77–86.

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

Toft, N., E. Jørgensen, and S. Højsgaard. 2005. Diagnosing diagnostic tests: Evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. Prev. Vet. Med. 68:19–33.

3.5 A framework for integration of benchmarking and within-herd analysis in dairy herd management – analysis of lactation curves as a case

Mogens A. Krogh and Carsten Enevoldsen

Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Grønnegårdsvej 2, DK-1870 Frederiksberg C, Denmark

Manuscript.

A framework for integration of benchmarking and within-herd analysis in dairy herd management – analysis of lactation curves as a case

M. A. Krogh[a*] & C. Enevoldsen[a]

[a]*Department of Large Animal Sciences*
Faculty of Health and Medical Sciences, University of Copenhagen,
Groennegaardsvej 2, DK-1870 Frederiksberg C, Denmark

[*] Corresponding author: email: mok@life.ku.dk (Krogh, M.A)

Abstract

Comparison of livestock herd key performance indicators is a widely used management tool for farmers and herd management consultants (benchmarking). The results of this comparison should lead to thorough within-herd and between-herd analyses to identify causes of exceptional variation (unsatisfactory performance). Milk production is obviously the key output from a dairy herd. However, the distribution of milk production between calvings (the 'shape of the lactation curve') can also be an indicator of production efficiency, including health traits. The objectives of this study were to describe the variability of the shapes of lactation curves within and between dairy herds and suggest frameworks for integration of benchmarking and within-herd analysis in herd management. A total of 51,311 lactations and 345,595 test-days of energy-corrected milk (ECM) yield from 170 Danish Holstein herds were used in the analysis. A random coefficient test-day model that allows a 'break' around 60 days in milk was applied to parity groups 1, 2, and older. A single-herd model was applied for each parity group (3) within each herd (170), and a multi-herd model was applied once for each parity group. Internal validity of the models was acceptable. The multi-herd model lactation curve estimates were closer to the overall mean, but the numeric differences in lactation curve estimates between the models were marginal. To demonstrate a principle for within-herd identification of causes of unsatisfactory performance, age at first calving was included in the multi-herd model as a random variable at herd level and a fixed variable at cow level. The result was that the impact of age at first calving on herd lactation curves was highly herd-specific. In addition, age at first calving modified the persistency of the lactation curves through an interaction. For a 1-month increase in age at first calving from 26.6 months, herd-level estimates ranged from 24 kg ECM to 126 kg ECM per 295 d per cow. The study conclusions were that 1) both the single-herd and multi-herd models provided similar and valid estimates of the major lactation curve parameters at the herd and cow levels; 2) the estimates of the lactation curves derived from the models are unbiased and useful for benchmarking of herds; 3) inclusion of other determinants of milk production is possible and provides herd-specific estimates about the relationship between the

determinant and the milk production; and 4) important information about how the determinant influences the lactation curve is obtained.

Keywords: Benchmarking, herd management, within-herd analysis, between-herd analysis, random coefficient regression model, effect modification

Introduction

Ranking the performance of herds and cows and comparing them with the best in a comparable group is a widely used principle in dairy herd management. This benchmarking is a tool to identify best management practices and set targets for key performance indicators (KPIs). Benchmarking or identification of best practices is regarded as a challenging factor for many farmers (Nir-Markusfeld, 2003). Herd management consultants, including veterinarians, often work in groups that have started to organize data collection; usually these data are organized hierarchically. For the consultants, information that is useful for improvement of the production process may be derived from systematic statistical analyses of the differences in performance measurements among groups, herds, and cows within herds.

A huge number of performance indicators can be derived from routinely collected records in dairy herds. Some are very complex. Milk production is obviously the key output from a dairy herd and could simply be measured as an average per cow. However, the distribution of milk production between calvings (usually about 1-year intervals) can be quite complex (the 'shape of the lactation curve') and can be an indicator of production efficiency, including health traits. An efficient analysis, including benchmarking, of a complex performance indicator like the lactation curve within the consultants' hierarchical contexts is not a trivial task. Consequently, development of a proper framework for analysis of lactation curves may be seen as a useful model for other traits (i.e., KPIs).

For management purposes, it is relevant to separate the milk production into milk production before 'peak' ('acceleration'), around 'peak lactation', and slope from peak lactation to 305 days in milk (persistency). These parameters together describe the shape of the lactation curve, an important cow-level trait or response for the following major reasons: Flat (persistent) curves may indicate a more efficient production process because the cows can be or have been fed with a lower concentrate-to-roughage ratio (Sölkner and Fuchs, 1987). Increasing peak milk yield is associated with a higher risk of disease (Østergaard & Gröhn, 1999) and poorer fertility (Loeffler et al., 1999) because of a lack of energy. However, a higher peak will often be associated with a higher total milk production during lactation although this relationship obviously depends on the magnitude of the correlation between peak and persistency. Correlations between peak and persistency have been estimated but with a primary focus on genetics (Cobuci et al., 2005). There is huge variability in the acceleration part of the lactation curve where some cows experience a marked drop in milk production from calving to peak lactation (Macciotta et al., 2006). This shape is not physiological. Obviously, the shape of the lactation curve and the variation in lactation curves within a herd cannot be expressed in one single number, which makes benchmarking complicated.

Many traits (determinants) are associated with milk production (e.g., age at calving, dry period length, and diseases), and quantification of these associations is important for efficient

management. However, these associations seem to differ among herds (Nir-Markusfeld, 2003). Consequently, within-herd analyses are relevant for herd management purposes. However, within-herd analysis of risk factors may be complicated because of low numbers of study units in some herds. Multilevel analysis based on multiple herds should make it possible to derive more precise estimates of the influence of a trait on the individual herd's milk production instead of providing overall across-herd estimates that do not reflect the specific herd, or imprecise herd-specific estimates from single-herd analyses. Knowledge is lacking about the effects of choice of modelling principle (multi-herd or single-herd) on the parameter estimates.

The overall objective of this study was to suggest a framework for examining the variability of a complex KPI within and between dairy herds and for integration of benchmarking and within-herd analysis in herd management. The specific objectives were to 1) compare estimates of lactation curve characteristics (as a model for a complex KPI) derived from herd-specific models with estimates from a multi-herd model and 2) provide estimates of the between-herd variation in the effect of a determinant of milk production (a KPI).

Material and methods

Data collection

Data were extracted from the VPR platform (Krogh & Enevoldsen, 2006), which provides a subset of the national Danish Cattle Database. The subset is homogeneous in the sense that the herds work intensively with herd health management. Data were extracted on September 22, 2008. We included herds for which the last milk yield recording at herd level was later than December 31, 2007, and we included lactations within these herds if the calving was less than 2 years prior to the last milk yield recording in the herd. We chose the 2-year period so that we safely could ignore the possible increase in milk production attributable to genetic progress within the herds. Herds with fewer than 11 milk yield recordings each year or herds with less than 90 percent of the calvings derived from Danish Holstein cows were excluded. Based on these criteria, 170 herds were included. We excluded milk yield recordings after 400 Day in Milk (DIM) because previous exploratory studies indicated that these values can be quite variable. The last milk yield recording date within a lactation was also excluded from the analyses because this recording could be influenced by non-recorded drying-off procedures like reduced energy supply or milking frequency. Test-day results that had missing or zero values for fat percentage, protein percentage, or kilograms of milk were excluded from the data file. First-parity cows with an age at first calving above 42 mo or below 18 mo (21 cows) were excluded from the data file. The resulting dataset consisted of 51,311 lactations and 345,595 milk recording test-dates equal to 6.7 test-dates per lactation. The test-day energy-corrected milk (ECMT) yield was calculated using formula 1:

$$\text{kg ECMT} = (\text{kg milk} \times (383 \times \text{fat\%} + 242 \times \text{protein\%} + 780.8))/3140 \qquad (1)$$

Descriptive statistics

Table 1 provides an overview of the data file. Distributions of herd-averages of test-dates, ECM, and age at first calving indicate a considerable herd-level clustering. As an example, the median ages at first calving within herds ranged from 23.5 mo in one herd to 32 mo in another herd.

Table 2. Distributions of median number of calvings, test-days, energy-corrected milk yield (ECM) per day, and age at first calving at herd level by parity group

	1st parity	2nd parity	3rd + parity
N herds	170	170	170
N calvings	19,720	14,955	16,636
N test-dates	137,584	100,456	106,555
Average ECM per test-day	28.5	33.4	34.3
Median no. of calvings (min–max) per herd	102 (22–449)	78 (17–293)	88 (28–341)
Median no of test-days (min–max) per herd (median within herd)	8 (3–10)	7 (3.5–10)	7 (4–10)
Median ECM per test-date (min–max) per herd (median within herd)	28.5 (19.8–35.8)	33.5 (22.4–41.3)	34.9 (23.5–44.5)
Median age at first calving in months (min–max) per herd (median within herd)	26.0 (23.5–32.0)	-	-

Statistical analysis

The lactation profiles of ECMT were created as a piecewise linear function of DIM, which allowed a break at 60 DIM for individual cows. This calculation was accomplished by creating two variables derived from DIM: 1) DIMun60, which takes the value of (DIM-60)/60 when DIM is less than 60, or otherwise the value is 0; and 2) DIM60, which takes the value (DIM-60)/245 when DIM is greater than 60, or otherwise DIM60 equals 0. DIMun60 then describes the change in ECM in the period from calving to 60 DIM, and DIM60 describes the change in ECM from 60 DIM to 305 DIM.

This model type has been used in other studies in recent years (Bennedsgaard et al., 2003; Nielsen et al., 2009). The advantages are the minimized number of parameter estimates and ease of interpretation. These features are essential when the aim is to provide support for herd-specific management in populations with limited numbers of observations (compared to sire-evaluations). The statistical analysis was conducted in three steps. First, a 3-level piecewise linear random coefficient model was fitted. The levels were herd, cow, and test-date. We fitted separate models to parity groups 1, 2, and >2 because we expected the shapes to be substantially different at both cow and herd levels. The following model 2 (M2) was applied to the 3-level data:

$$ECMT_{ijk} = \beta_{0jk}\text{DIMun60}_{ijk} + \beta_{1jk} + \beta_{2jk}\text{DIM60}_{ijk} + \varepsilon_{ijk}$$

where

$$\beta_{0jk} \sim \beta_0 + \nu_{0k} + \mu_{0jk}$$

$$\beta_{1jk} \sim \beta_1 + \nu_{1k} + \mu_{1jk}$$

$$\beta_{2jk} \sim \beta_2 + \nu_{2k} + \mu_{2jk}$$

$$\begin{pmatrix} \nu_{0k} \\ \nu_{1k} \\ \nu_{2k} \end{pmatrix} \sim N(0, \Omega_\nu) : \Omega_\nu = \begin{pmatrix} \sigma^2_{\nu 0} & & \\ \sigma^2_{\nu 01} & \sigma^2_{\nu 1} & \\ \sigma^2_{\nu 02} & \sigma^2_{\nu 12} & \sigma^2_{\nu 2} \end{pmatrix}$$

$$\begin{pmatrix} \mu_{0jk} \\ \mu_{1jk} \\ \mu_{2jk} \end{pmatrix} \sim N(0, \Omega_\mu) : \Omega_\mu = \begin{pmatrix} \sigma^2_{\mu 0} & & \\ \sigma^2_{\mu 01} & \sigma^2_{\mu 1} & \\ \sigma^2_{\mu 02} & \sigma^2_{\mu 12} & \sigma^2_{\mu 2} \end{pmatrix} \quad (2)$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2_\varepsilon)$$

where i = test-date, j = cow, and k = herd.

Based on parameter estimates from M2, milk yields were predicted at DIM 10, 60, and 305 for each cow. The change in milk yield from 10 to 60 DIM ($\beta_{0jk,}$ acceleration) and the change in milk yield from 60 to 305 DIM (β_{2jk}, persistency) were estimated. The point estimate at DIM 60 (the intercept) is an estimate of β_{1jk}, which could be labelled 'peak', but because the acceleration may be negative, this designation can cause confusion. However, in the following we use the terms acceleration, peak, and persistency (β_{0jk}, $\beta_{1jk,}$ and β_{2jk}, respectively). Similarly, a two-level model (within herd) was specified and analysed for each herd (170) and each parity group. Model 3 (M3) is given below.

$$ECMT_{ij} = \beta_{0j}\text{DIMun60}_{ij} + \beta_{1j} + \beta_{2j}\text{DIM60}_{ij} + \varepsilon_{ij}$$

where

$$\beta_{0j} \sim \beta_0 + \mu_{0j}$$

$$\beta_{1j} \sim \beta_1 + \mu_{1j}$$

$$\beta_{2j} \sim \beta_2 + \mu_{2j} \quad (3)$$

$$\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \\ \mu_{2j} \end{pmatrix} \sim N(0, \Omega_\mu) : \Omega_\mu = \begin{pmatrix} \sigma^2_{\mu 0} & & \\ \sigma^2_{\mu 01} & \sigma^2_{\mu 1} & \\ \sigma^2_{\mu 02} & \sigma^2_{\mu 12} & \sigma^2_{\mu 2} \end{pmatrix}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$$

where i = test date and j = cow.

The cow-level estimates of acceleration, peak, and persistency (β_{0jk}, $\beta_{1jk,}$ β_{2jk}), which were cow-level estimates derived from model 3, were aggregated to herd level. The parameter estimates

for cows within a herd were averaged, and the variances and covariances were calculated. In this way, they are comparable to the parameter estimates from model 2.

Finally, age at first calving was added to model 2 for first-parity cows. Age at first calving was centred to the average across all cows (26.5 mo). Linear and quadratic terms of age at first calving, interactions of age at first calving, and DIMun60 and DIM60 were added to the fixed part of the model. The linear term of age at first calving was also included as a herd-level random effect, and the covariances with DIMun60, peak, and DIM60 were all included. At cow level, this model will allow the shape of the individual cow's lactation curves to vary with age at first calving because of the interactions with the variables describing the lactation curve. The herd-level age at first calving will modify the effect of age at first calving of the individual cows within each herd. This initial model was subsequently reduced based on deviance tests for fixed effects (maximum likelihood) and covariances (restricted maximum likelihood) to a final 3-level model (model 4). Based on the 3-level model selection, a similar 2-level model was specified, and estimates of the effects of age at first calving on milk production were compared for the two models.

$$ECMT_{ijk} = \beta_{0jk} + \beta_{1jk}\text{DIMun60}_{ijk} + \beta_{2jk}\text{DIM60}_{ijk} + \beta_{3k}\text{Age}_j + \beta_4\text{Age}_j^2 + \beta_5\text{Age*DIM60}_j + \varepsilon_{ijk}$$

where

$$\beta_{0jk} \sim \beta_0 + \nu_{0k} + \mu_{0jk}$$

$$\beta_{1jk} \sim \beta_1 + \nu_{1k} + \mu_{1jk}$$

$$\beta_{2jk} \sim \beta_2 + \nu_{2k} + \mu_{2jk}$$

$$\beta_{3k} \sim \beta_3 + \nu_{3k}$$

$$\begin{pmatrix} \nu_{0k} \\ \nu_{1k} \\ \nu_{2k} \\ \nu_{3k} \end{pmatrix} \sim N(0, \Omega_\nu) : \Omega_\nu = \begin{pmatrix} \sigma^2_{\nu 0} & & & \\ \sigma^2_{\nu 01} & \sigma^2_{\nu 1} & & \\ 0 & 0 & \sigma^2_{\nu 2} & \\ 0 & 0 & 0 & \sigma^2_{\nu 3} \end{pmatrix}$$

$$\begin{pmatrix} \mu_{0jk} \\ \mu_{1jk} \\ \mu_{2jk} \end{pmatrix} \sim N(0, \Omega_\mu) : \Omega_\mu = \begin{pmatrix} \sigma^2_{\mu 0} & & \\ \sigma^2_{\mu 01} & \sigma^2_{\mu 1} & \\ \sigma^2_{\mu 02} & \sigma^2_{\mu 12} & \sigma^2_{\mu 2} \end{pmatrix} \quad (4)$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2_\varepsilon)$$

where i = test-date, j = cow, and k = herd.

Statistical analysis for the 3-level model was conducted with the software programme MlWin 2.0 (Rasbash, 2004). Prediction and analysis of the 2-level models were conducted with SAS® PROC MIXED (Littell et al., 2006). All analyses were performed with restricted maximum likelihood estimation. Model control was performed by examining leverage, influence, and standardized residuals graphically for each level in the models.

Results

Three-level model

Parameter estimates and standard errors of the mean for the fixed effects derived from model (2) are in Table 2. These parameter estimates describe the overall average shapes of the parity-specific lacation curves. It is evident that lactation curves from 1st-parity cows differ in shape from 2nd and older parities with an increase in milk yield from DIM 10 to DIM 60, a much lower peak milk yield, and a much better persistency. Second and older parities had only minor changes in milk yield from DIM 10 to DIM 60 compared to first parity: a 7–9 kg ECM higher production at peak lactation and a persistency of -11.5 kg ECM/245 d to -14.1 kg ECM/245 d. Regardless of parity, the milk production at 305 DIM was around 25 kg ECM. The intercept for 2nd parity was lower than for older cows, and the average slope before peak was negative.

Table 2. Fixed-effects parameter estimates from a multi-herd model of test-day energy-corrected milk yield (ECM). DIMun60 is the change in ECM from 10 to 60 DIM, Intercept is the point estimate ECM at 60 DIM, and DIM60 is the change in ECM from 60 to 305 DIM. Standard error of the mean (SEM) is in parentheses.

	1st parity	2nd parity	3rd + parity
DIMun60 β_1 (SEM)	2.16 (0.14)	-0.26 (0.18)	0.46 (0.18)
Intercept β_0 (SEM)	29.97 (0.18)	37.01 (0.25)	38.91 (0.27)
DIM60 β_2 (SEM)	-3.85 (0.11)	-11.58 (0.18)	-14.17 (0.17)

Table 3 gives the parameter estimates of the variance–covariance matrices Ω_v and Ω_μ for each parity. All the variance components increased with increasing parity. The residual variances, which are the variation within cows within herds for parity groups 1, 2, and >2, were 8.5, 12.5, and 15.0, respectively. These values represent about half of the cow-level (between cow) variance components at peak milk production (16.5, 29.9, and 37.9, respectively). Variance components at herd level (Ω_v) are the variation between herds, whereas variance components at cow level (Ω_μ) are the variation between cows within a herd. The total between-cow variation is, consequently, the sum of the cow-level and herd-level variance components. The herd-level variance components related to peak production were 3–10 times smaller than the corresponding cow-level variance components. It is clear that the herd level accounted for only minor differences in the shapes of the lactation curve between cows. The herd-level variance component related to peak lactation (σ^2_{v1}) is of special interest because this component probably will account for most of herd-level differences in level of total production.

Covariance parameter estimates from different levels are difficult to interpret directly. Therefore, correlations are calculated instead according to formula 5:

$$\rho_{xy} = \frac{\mathrm{cov}(x,y)}{S_x * S_y} \quad (5)$$

The herd-level covariance estimates between acceleration and persistency (σ^2_{v02}) and between peak lactation and persistency (σ^2_{v12}) in first parity were statistically non-significant (P>0.20 and P=0.15, respectively) and consequently set to 0. Based on the correlations, it is evident that even though the

variances are different between parity groups, the correlations are very similar both at herd level and cow level for all parities (approximately -0.5). That is, an increase in peak milk yield is consistently associated with a markedly poorer persistency.

Table 3. Estimates of variance–covariance matrices Ω_ν, Ω_μ, and residual variances derived from a multi-herd model of test-day measurements of energy-corrected milk yield (ECM). Correlations (ρ) for the covariances are in parentheses.

	1st parity			2nd parity			3rd + parity		
	ν_{ok}	ν_{1k}	ν_{2k}	ν_{ok}	ν_{1k}	ν_{2k}	ν_{ok}	ν_{1k}	ν_{2k}
$\widehat{\Omega_\nu}$ (ρ)	2.8[a]	-	-	4.2[a]	-	-	4.3[a]	-	-
	1.3[d] (0.3)	5.1[b]	-	1.5[d] (0.3)	9.6[b]	-	1.9[d] (0.3)	11.6[b]	-
	0[e] (0.0)	0[f] (0.0)	1.7[c]	-1.2[e] (-0.3)	-3.1[f] (-0.5)	4.17[c]	-1.1[e] (-0.3)	-3.4[f] (-0.5)	3.9[c]
	μ_{ojk}	μ_{1jk}	μ_{2jk}	μ_{ojk}	μ_{1jk}	μ_{2jk}	μ_{ojk}	μ_{1jk}	μ_{2jk}
$\widehat{\Omega_\mu}$ (ρ)	22.0[g]	-	-	41.5[g]	-	-	54.3[g]	-	-
	3.8[j] (0.2)	16.5[h]	-	6.25[j] (0.2)	29.9[h]	-	9.5[j] (0.2)	37.9[h]	-
	-1.3[k] (-0.1)	-6.1[l] (-0.4)	13.6[i]	-2.65[k] (-0.1)	-15.6[l] (-0.5)	32.45[i]	-3.3[k] (-0.1)	-21.0[l] (-0.5)	39.3[i]
Residual		8.50			12.54			14.96	

[a] Herd-level variance component related to acceleration
[b] Herd-level variance component related to peak milk production
[c] Herd-level variance component related to persistency
[d] Herd-level covariance between acceleration and peak
[e] Herd-level covariance between acceleration and persistency
[f] Herd-level covariance between peak and persistency
[g] Cow-level variance component related to acceleration
[h] Cow-level variance component related to peak milk production
[i] Cow-level variance component related to persistency
[j] Cow-level covariance between acceleration and peak
[k] Cow-level covariance between acceleration and persistency
[l] Cow-level covariance between peak and persistency

Model control (results not shown) for the 1st parity indicated that three herds had high leverage values for the intercept (DIM=60 d) at herd level. The three herds could also be found in the residual plots for the intercept. However, we retained them in the study because excluding these herds from analysis did not influence the parameter estimates. Leverage at herd level for DIMun60 and DIM60 did not cause concern. At cow level, there were no apparent problems with leverage, influence, and residuals. Plots of standardized residuals at test-day level against DIMun60, DIM60, and DIM gave no indication of a fanning pattern (results not shown). Similar diagnostic graphs were created for 2nd- and 3rd- parity groups (results not shown). As with the 1st-parity group, one and three herds showed high leverage values for the 2nd- and 3rd-parity groups, respectively, for the intercept term. We could not identify obvious reasons for these potential problems. Leaving these herds out of the analyses did not change the parameter estimates. For the 2nd-parity group, there was a tendency to an excess of small negative residuals for the intercept term, making the residual distribution for the intercept a little skewed to the left. For the 3rd-parity group, the distributions of residuals were acceptable at all levels.

Two-level model

Each of the 170 herds for each of the three parity groups was analysed using model 3 (510 analyses). In total, eight parity groups within herds did not converge successfully to give unique estimates of the parameters included in the model. One of these herd-parity group combinations had a low number of cows (n=22), but for the other groups, we cannot suggest possible reasons. Table 4 gives fixed-effect parameter estimates as means and the minimum and maximum values of herds to describe the variation in parameter estimates. It is evident that the means are very similar to the parameter estimates derived from model 2. The means of the variance components and the means of correlations between variance components are all very similar to the parameter estimates derived from model 2.

Table 4. Means of parameter estimates from model 3. DIMun60 is the change in energy-corrected milk (ECM) production from 10 to 60 DIM, Intercept is the point estimate of ECM at 60 DIM, and DIM60 is the change in ECM from 60 to 305 DIM. The number of herds indicates the number of herds where the model converged. Parameter estimates are averaged over the number of herds where the model converged (N herds). Minimum and maximum estimates are given in parentheses.

	1st parity	2nd parity	3rd + parity
DIMun60 β_1	2.08 (-2.84; 7.22)	-0.38 (-5.48; 7.65)	0.34 (-4.66; 8.30)
Intercept β_0	29.94 (20.00; 39.09)	36.98 (24.01; 45.29)	38.89 (26.07; 51.5)
DIM60 β_2	-3.83 (-6.60; -1.31)	-11.54 (-18.12; -5.67)	-14.09 (-19.59; -7.98)
N herds	165	168	169

Model control (results not shown) was conducted the same way as for the 3-level model except herd level was not included in these analyses. Plots of standardized residuals against DIMun60, DIM60, and DIM did not cause major concern (results not shown). There were no apparent fanning patterns in the residuals (results not shown), but for some herds (~10%), the distribution of residuals was a little wider than for the normal curve.

Comparison of the 2-level and 3-level models

The 2-level model was compared graphically with the 3-level model. The parameter estimates of model 2 were ranked according to the peak milk production (β_{1j}), and the first, the last, and every 10th herd were selected to allow a visual inspection. The same 18 herds were selected from the 3-level model. Figure 1 shows the 18 selected herds for 2nd parity. The dotted horizontal reference line symbolizes the grand mean of the peak milk production from the 2-level model. It is evident that the 3-level model will give estimates closer to the grand mean of peak milk production than the 2-level model.

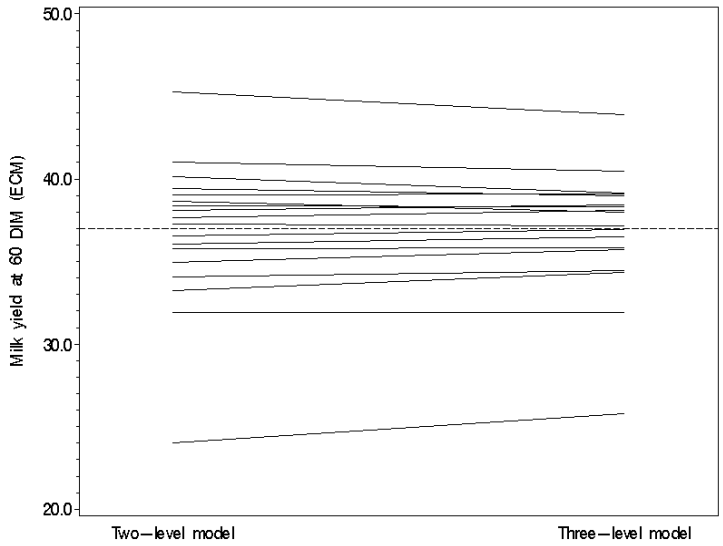

Figure 1. Ranking of 18 representative herds (2nd parity) based on the 2-level (single-herd) and 3-level (multi-herd) models of peak milk yield at 60 DIM. Dashed line is grand mean for all herds.

Figure 1 shows that some herds will change rank based on model type. The two models were subsequently evaluated with 2×2 contingency tables to allow comparison of the results of the two models. Based on model 2, the 25th and the 75th percentiles were identified for the herd-level estimates for acceleration, peak, and persistency. These percentiles were then used as cut-off values for the parameter estimates from model 3. The conditional probability that the model 3 (M3) estimates were outside the interquartile range given that model 2 (M2) estimates were within the interquartile range was denoted as P(M3-| M2+). The conditional probability that a model 3 estimate is within the interquartile range given that a model 2 estimate was not within the interquartile range was denoted as P(M3+|M2-). For all parities, these two probabilities were very similar. For the acceleration, P(M3-|M2+) was between 9.5% and 11.3%, and P(M3+|M2-) was between 0% and 0.6%. For peak lactation, P(M3-|M2+) was between 2.3% and 3.6%, and P(M3+|M2-) was between 0% and 0.6%. For persistency, P(M3-|M2+) was between 10.7% and 15.2%, and P(M3+|M2-) was between 0% and 1.8%. Consequently, there was a much higher probability that model 3 gave estimates outside the interquartile range when model 2 gave estimates within the interquartile range than the reverse probability.

Herd-level estimates of peak production, persistency, and total milk yield from DIM10 to DIM305 were compared for the two models to assess the numerical differences between the models. Estimates of peak milk production and persistency were rounded to 0.5 kg ECM, and total milk yield estimates were rounded to 50 kg ECM. For the estimates of peak milk production, 83%, 66%, and 66% of the herds gave a complete match of absolute values between the models for 1st, 2nd, and 3rd parity, respectively. The largest difference between the two models was 1.5 kg ECM for 2nd parity. For persistency, there was a complete match in absolute values between the models in 54%, 33%, and 35% for 1st, 2nd, and 3rd parity, respectively. The largest persistency difference was 2 kg ECM, and there was a slight tendency for the 3-level models to give a poorer persistency than the 2-level model. The estimates of kg ECM produced between 10 DIM and 305 DIM from the two models showed that for all parity groups, 75% of the herds were considered a match. For the 1st-

parity cows, the largest difference was 200 kg ECM. For 2nd- and 3rd-parity cows, the largest difference between the models was 300 kg ECM. There was no tendency for any one of the models to give systematically larger estimates than the other for total milk production.

Similarly, herd-level prediction intervals were calculated, and the differences were minor. In Figure 2, predicted lactation curves and the corresponding 95% prediction intervals are given for 2nd parity in herd no. 36. The gray lines are the 3-level model, and the black lines are the 2-level model. Herd no. 36 was chosen because it had the lowest number of cows in 2nd parity (17 cows) and because it had the largest deviations between the lactation curves derived from models 2 and 3.

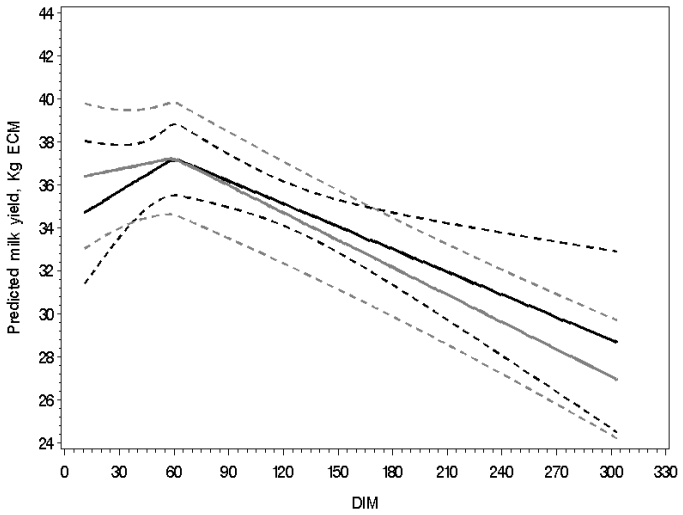

Figure 2. Lactation curves for 2nd lactation cows in herd no. 36. Black lines are derived from the 2-level (single-herd) model; gray lines are derived from the 3-level (multi-herd) model. Dashed lines represent 95 percent prediction intervals.

Figure 2 shows that the 2-level model for herd no. 36 until 210 DIM has a narrower prediction interval than the 3-level model. After 210 DIM, the 3-level model gave the narrowest prediction interval. Note here that the prediction interval for the 3-level model consists of two components: one for the fixed-effects equation and one for the prediction of the random effects. The standard error of the mean related to the fixed part of the prediction equation is low because this is based on the entire datafile. However, because the number of cows in herd no. 36 was low (17 cows), the standard error of the mean related to the random prediction is large. In herd number no. 36, the ratio between standard error of the mean for the fixed part and standard error of the mean for the random part in the prediction is 1:6 whereas in a herd with more cows, the ratio is around 1:3.

Age at first calving

The 3-level model with inclusion of age at first calving was reduced to model 4 from the modified model 2 with age at first calving included. The interaction between age at first calving and DIMun60 could be removed (P>0.2) together with all the herd-level covariances between the random component related to age and the random components related to the shape of the lactation curve, $\sigma^2_{v0}, \sigma^2_{v1}, \sigma^2_{v2}$ (P>0.2). The interaction between age at first calving and persistency was highly

significant (P<0.001), as were the quadratic age term and random components related to age (P<0.001).

Table 5. Key parameter estimates for first-parity cows derived from model 5. DIMun60 is the change in energy-corrected milk yield (ECM) from 10 to 60 DIM, Intercept is the point estimate of ECM at 60 DIM, and DIM60 is the change in milk yield from 60 to 305 DIM. Standard errors of the mean (SEM) are given after the estimates in parentheses.

	Fixed estimate	Random estimate herd-level	Random estimate cow-level
DIMun60	2.16 (0.14)	2.81 (0.36)	22.04 (0.46)
Intercept	30.05 (0.19)	5.67 (0.64)	15.80 (0.21)
DIM60	-3.84 (0.11)	1.70 (0.22)	13.46 (0.32)
Age	0.34 (0.02)	0.02 (0.00)	
Age^2	-0.015 (0.00)		
Age*Dim60	-0.13 (0.02)		

Empty cells indicate that the variable is not included as random in the model.

Parameter estimates from model (4) are provided in Table 5. The additional milk produced between 10 and 305 DIM if an average cow calves one month later than the overall average (26.5 mo) can be calculated from the function 0.323 kg/d × 50 d + 245 d × (0.323 kg/d–0.127 kg/d) + ½ × (245 d × 0.127 kg/d) = 79.7 kg/295 d of lactation, which also includes the interaction between age at first calving and persistency. If no interaction was present, the additional milk yield would simply have been 295 d × (0.34 kg/d–0.015 kg/d) = 98.5 kg/295 d lactation, based on parameter estimates from a model without the interaction term (not shown). For an average cow, the parameter estimates in Table 5 can be used to extrapolate to any changes in age at calving. However, including herd information gives estimates for the same change in age (26.5 mo + 1 mo) at first calving in milk production from 24 kg ECM as the lowest in one herd to a maximum at 126 kg ECM in another. Figure 3 shows predicted ECM at 60 DIM as a function of age at first calving. Lines describe the variation in age at calving. That is, herds clearly have different shapes.

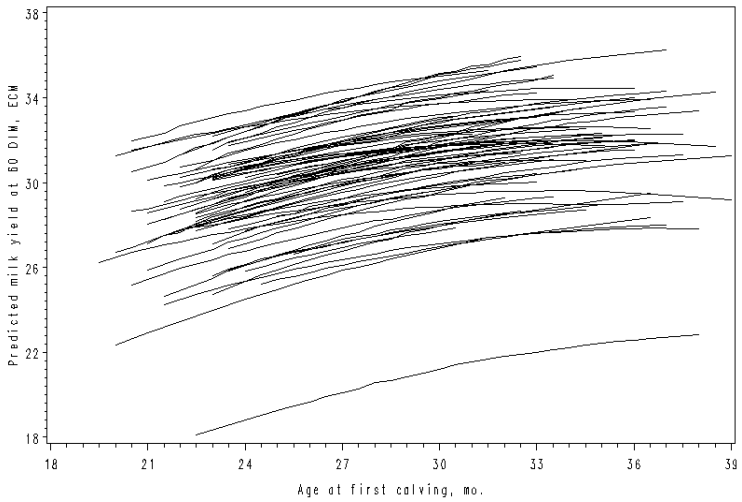

Figure 3. Predicted milk yield at 60 DIM at different ages at first calving. Lines represent 70 randomly selected herds.

Time to reach convergence for the most complex (3-level) model was achieved with three iterations and a runtime of less than a minute with an ordinary Pentium 4 3.00 GHz PC with 2 MB RAM. This time frame indicates a very robust model specification. Model control concerning the random component of age at first calving did not cause any concerns about the normality assumption (results not shown).

Discussion

Organizational framework for herd analysis and choice of statistical model

The organizational framework for a production or health management consultant's datafiles will typically be hierarchical with groups of herds, animals within herds, and some measurements within individual animals. Benchmarking and other decision support are relevant at all levels. The multilevel random coefficient model presented in this study provides a formal framework that allows benchmarking and both within-herd and between-herd analyses of effects of relevant determinants. More complex organizational frameworks may occur. For example, several veterinarians may provide service to the same herds, and one farmer may own several herds. In such cases, the structure is no longer strictly hierarchical. However, such cross-classification of levels can be handled with modifications of model 2 (Fielding and Goldstein, 2006). Addition of one or more levels is also straightforward, although computer power may be a constraint with an increasing number of levels and complexity of cross-classification.

In cattle breeding programs, very advanced statistical models have been developed ('random coefficient test day animal models'). Obviously, the primary objective is to provide efficient indices for selection of the best sires and dams. In recent years, attempts have been made to supplement these animal models with specifications that provide estimates of relevance for herd management, such as estimates of herd-level persistency of lactation curves (Caccamo et al., 2008). However, the estimates from this type of analysis may be complicated for consultants to use and interpret (e.g., Legendre polynomials). Because the breeding models obviously must include the entire population, they may also be too rigid to serve the diverse needs of the consultants. For instance, a local network of veterinarians may have implemented a new system for clinical examinations. The resulting records should be included in a statistical model including the participating herds. Meeting this diversity of needs for specification of effects will probably not be possible in a national breeding program. For management purposes, examination of numerous interactions (effect modification) are also of major interest (cf. the interactions including age at calving demonstrated in this study). Herd-level estimates from breeding models are problematic for herd-specific use because several potentially important determinants of milk yield (e.g., age at calving) are adjusted for. That is, cow-effects are standardized (Koivula et al., 2007). However, the model we suggest in this study probably often will provide more precise estimates if we include genetic information.

Data Envelopment Analysis (DEA) is well developed for benchmarking (e.g., Førsund and Sarafoglou, 2002). An advantage of DEA is that it combines numerous KPIs in a joint analysis. However, the uncertainty associated with the ranking based on the aggregation of measurements into a KPI is not accounted for. In the model we suggest, uncertainty at all levels is explicitly estimated. The components of variance per se are also of interest for management. DEA or factor analysis (e.g., as demonstrated by Enevoldsen et al., 1996) might be useful tools to combine the

estimates from a model like ours in this paper, including the variance components, with estimates from similar statistical analyses of other KPIs.

Model validity

In this study, we focused on a test-day model of lactation curves that is similar to the well-known and commonly applied Wilmink function (Wilmink, 1987). There are two reasons why we did not use the Wilmink function directly. First, an estimate of the acceleration part as a straight line is easier to interpret than the corresponding exponential function in the Wilmink function; especially, interpretation of the variance is easier. Second, the Wilmink function probably has been developed to analyse milk, fat, and protein separately. Lactation curves of milk, fat, and protein likely are poorly described by straight lines compared to lactation curves based on ECM, which directly reflect the energy output from the cow. The use of ECM is also a more biological focus instead of other measures of milk production that are more related to varying market conditions.

In this study, we have chosen to treat 2^{nd} and older parities separately, primarily because most 2^{nd} lactation cows need to grow during 2^{nd} lactations. Also, the disease pattern of 2^{nd} lactation is different from older parities; for example, the risk of milk fever is much lower.

Standardized residuals, influence diagnostics, and leverage points indicate that the models fitted the test-date data acceptably well. The parameter estimates derived from our baseline models—models 2 and 3 that describe our lactation curve—meet our expectations of a detectable peak of the lactation curves for 1^{st} parity, decreasing persistency with increasing parity, increasing residual variation with increasing parity, and increasing between-cow variation with increasing parity group (Tables 2 and 3). The variance components and fixed-effect estimates related to the lactation curve from the two models are very similar (Tables 2, 3, and 4). Inclusion of a new variable in the model—age at first calving—did not cause marked changes in the parameter estimates of the lactation curve. Parameter estimates that did not change in different analyses and fast convergence of the models indicate that the models are properly specified and robust.

Interpretation of model output

The actual value of a KPI must be compared to predetermined limits for “good” or “acceptable” KPI values. Ideally, such objectives (or goals, targets, or reference values) should reflect the individual herd’s combination of resources. That is, the manager needs to know what is possible to achieve within the production system. However, predicting the future performance because of the complexity of the system may be a difficult task.

Benchmarking is one obvious way to select targets. The principle of benchmarking is to identify other herds with a similar combination of resources and compare the actual KPI with the range of results in these reference herds. Historical results from the actual herd can also provide very useful target values. The selected target KPI can also be considered a prognosis for the future.

A major advantage of random coefficient models is that estimation of the components of the lactation curve at cow level is not entirely based only on the individual cow’s performance but also on other cows’ performance within that herd. In plain language, each level in the model ‘borrows’ information from the other levels. This sharing of information from other levels could be seen as a ‘filtering’ or ‘smoothing’ technique. The apparent shrinkage shown in Figure 1 demonstrates this

phenomenon. The modelling approach allows prediction of lactation curve components at cow level for cows with incomplete lactations. Our model allows a relative ranking of herds or cows based on a KPI that is adjusted for various determinants of the KPI. This ranking can be done most effectively using a multilevel model. In addition, the random coefficient model provides parameter estimates concerning the variation in lactation curve components within herds, which could be of interest (Kristensen et al., 2008). However, these issues have apparently not been addressed in detail in the scientific literature and are related to the inherent problem with interpretation of rankings of variables estimated with uncertainty. Goldstein and Spiegelhalter (1996) pointed out the need for care with such rankings given that they may often be over-interpreted because they frequently have too much imprecision (very wide confidence intervals) for fine comparisons. In our case, the uncertainty associated with the herd-level estimates may permit only distinction of herds separated by one quartile. Caterpillar diagrams with confidence intervals may be a useful tool for identifying proper ranking intervals.

A key task in rational management is to predict future performance. The models presented here do allow prediction of the milk yield for the rest of the lactation for each cow. Consequently, short-term outlying milk yield records (e.g., caused by acute disease) can be identified for the individual cow. A framework is also provided for objective identification of individual cows with marked deviations in shapes of lactation curves (e.g., caused by chronic disease). At the herd level, knowledge about variability of the input parameters in the chosen prediction model (e.g., Kristensen et al., 2008) is essential. In this study, we have provided readily interpretable and unbiased estimates about the shape of the lactation curve at both cow and herd levels. The correlations between the shape parameters are estimated and accounted for. To our knowledge, information about the simultaneous variation in shape of the lactation curve within and between herds has not been studied in detail before.

In addition, in the current work, we compared a single-herd model with a multi-herd model. The conclusion is that if the primary goal of the analysis is to provide estimates of the shape of the lactation curve, there is little difference between the two models. However, there were convergence problems with 8 herd/parity groups out of 510 in the single-herd analyses. In Denmark, farmers can lower the number of milk yield recordings from 11 times a year to 6 times a year. This change could lead to increasing convergence problems for the single-herd model. In contrast, the herd sizes keep increasing (as of late 2011, approx. 150 cows per herd in Denmark), which also increases the number of study subjects in the single-herd model. For management purposes, we prefer the study period to be as recent as possible to allow prompt management reactions to changes in input factors. This approach will favor decreasing sample size, which will increase variability and increase estimation problems. These problems can be counteracted by some type of exponential smoothing or filtering of the key-figures (Thysen, 1993). However, the 3-level multi-herd model may be a better option that does not require a choice of arbitrary smoothing factors. The multi-herd model could also be expanded with estimates of general and herd-specific time trends like the harmonic seasonal trends suggested by Koivula et al. (2007). Such expansions probably will be robust because of the qualities of the mixed model. If the multi-herd model is preferred, the shrinkage phenomenon should be considered. It is well known that multilevel models experience shrinkage towards the mean of parameter estimates (Hox, 2002). The amount of shrinkage is large for herds

that are far from the overall mean and for herds with a low number of study units. That is, shrinkage is similar to smoothing. Based on the result shown in Figure 2, we would perfer the multi-herd model if the desired estimate is persistency, due to the narrower prediction interval late in lactation.

We have estimated the effects of age at first calving on the shape of the lactation curve in 1st-parity cows in this study as an example of an analysis of a determinant of milk production. In common production reports, we can find KPIs such as average age at first calving and average total milk production calculated at the herd level. A typically applied target figure for age at first calving is that it should be around 24 mo. Such recommendations are often based on herd averages like the ones in the production reports. However, relationships between such herd averages may be prone to ecological fallacy (Woodward, 2005). Ecological fallacy is the assumption that an observed relationship in aggregated data at the herd level will hold at cow level. This assumption can easily be violated when using recommendations or target figures. It is important to know if the target figure is based on cow-level or aggregated cow-level data. Even if no ecological fallacy exists, there could often be herd-specific conditions such as reduced feeding space for the cows that make these target figures inappropriate in some herds. One obvious solution is to base recommendations about the effects of a cow-level determinant of the output parameter of interest on the individual herd's own previous performance. Traditionally, this basis has been aggregated milk yield recordings (e.g., milk in early lactation or 305 d) at the cow level. The average milk yield during the first 3 mo will be a reasonable estimate of peak milk production, but information about the rest of the lactation is lost, and possibly imprecise conclusions can be expected. If total lactation is chosen as the outcome in the analyses, the outcome must be based on predicted yield to avoid selection bias due to incomplete lactations.

The single-herd model we have used in this study will provide reasonable estimates of the entire lactation and handle the selection bias related to incomplete lactations. Additionally, we have shown that the effect of age at first calving clearly is herd specific. It is not within the objectives of this study to derive general recommendations about age at first calving, and because of the strong herd effects, a general recommendation is of only minor interest. The multi-herd model could, if applied, improve the robustness of the herd-specific estimates of the effects of age at calving. In addition, it provides valid estimates about herd-to-herd variation that would be interesting in a sensitivity analysis about age at first calving (e.g., Kristensen et al., 2008).

A common problem, however, is that even though we have information about the individual herd, we may want to extrapolate. For example, we might want to predict milk yield if age at first calving is lowered from 30 mo to 24 mo in a herd without previous calvings at the 24-mo age. In this situation, the fixed effects from a multi-herd model probably would be the best choice, especially if the model is based on herds in which the context is homogeneous, i.e., feeding and housing are similar. If the multi-herd model is applied, the prediction will still be somewhat uncertain and close to the mean of other herds that have calvings around 24 mo (shrinkage). The biological relationship between age at first calving and milk production will not be violated as could happen if the single-herd model is applied. In that case, the quadratic effect might cause unrealistic predictions, and data would be insufficient to add additional polynomial effects.

In this study, age at first calving was chosen as an example of a factor that modifies the shape of the herd lactation curve. We have demonstrated how the herd lactation curve parameters (fixed and

random) can be influenced directly and through interactions. Our test-day model can easily be expanded to include other (potential) determinants of milk production. Age at first calving is a determinant that is collected and analysed at cow level. A similar and relevant factor is length of dry period. Factors sampled at other levels can also be included. In addition, test-day information like somatic cell count and herd-level information like grazing strategy, milking frequency, or milking system can be included. Obviously, herd-level factors can be added only in the multi-herd model. The major advantage is that multiple factors at different levels in the model provide us the ability to specify interaction and thereby handle correlations between the factors. An example could be inclusion of age at first calving, score of calving ease, and metritis handled in one model. So far, we have not experienced technical problems, but it is evident that when model complexity increases, problems with convergence are more likely to appear and computation time will increase.

For benchmarking purposes, a multi-herd model could be applied to (selected) herds within an advisory unit like an extension service or a veterinary practice. The benefits of such an organizational model would be that the consultants within that unit would have first-hand knowledge about the herd, which would provide the best possible foundation to select herds for benchmarking. Recording of diseases and applied disease codes are often practiced in a way that is specific to the practice. A within-practice analysis could, consequently, decrease random and systematic variation and thereby enhance the power of the analysis and increase validity of conclusions. Knowledge about the herds included makes it possible to incorporate herd-level information into the model that is not a part of central databases. Examples could be the drying-off strategy or milking system. Finally, ranking of herds that are all known to the veterinarian makes the ranking procedure far more transparent and useful for consultancy.

Conclusion

We suggest a multi-herd modeling framework for integration of benchmarking and within-herd analysis in dairy herd management. The framework is demonstrated with an analysis of lactation curves that represents a complex key performance indicator. Information for decision support is produced within cow, between cows, and at the herd level. Internal validity of the models was acceptable. Age at first calving was included in the models to assess the potential of the model framework for analysis of determinants of milk production or other KPIs. The conclusions of the study were as follows: 1) Both the single-herd and multi-herd models provided similar and valid estimates of the major lactation curve parameters at the herd and cow levels; 2) the estimates of the lactation curves derived from the models are unbiased and useful for benchmarking of herds; 3) inclusion of other determinants of milk production is possible and provides herd-specific estimates about the relationship between the determinant and the milk production; and 4) important information about how the determinant influences the lactation curve is obtained.

References

Bennedsgaard, T. W., Enevoldsen, C., Thamsborg, S. M., Vaarst, M. 2003. Effect of mastitis treatment and somatic cell counts on milk yield in Danish organic dairy cows. J Dairy Sci 86:3174-3183.

Cobusi, J.A., Euclydes, R.F., Lopes, P.S., Costa, C.N., Torres, R.A., Pereira, C.S. 2005. Estimation of genetic parameters for test-day milk yield in Holstein cows using a random regression model. Genet Mol Bio, 28(1):75-83.

Enevoldsen, C., Hindhede, J. & Kristensen, T. 1996. Dairy herd management types assessed from indicators of health, reproduction, replacement, and production. J Dairy Sci, 79:1221-1236.

Hill, P.W. and Goldstein, H. 1998. Multilevel modeling of educational data with cross-classification and missing identification for units. J Educ Behav Stat 23(2):117-128

Hox, J.J. 2002. Multilevel Analysis: Techniques and Applications. Lawrence Erlbaum Associates Inc., Mahwah, NJ. ISBN:0805832181 pp: 27-30

Jakobsen, J. H., Rekaya, R., Jensen, J., Sorensen, D. A., Madsen, P., Gianola, D., Christensen, L. G., Pedersen, J. 2003. Bayesian estimates of covariance components between lactation curve parameters and disease liability in Danish Holstein cows. J Dairy Sci 86:3000-3007

Koivula, M., Nousiainen, J. I., Nousiainen, J., Mäntysaari, E. A. 2007. Use of herd solutions from a random regression test-day model for diagnostic dairy herd management. J Dairy Sci 90:2563–2568. doi:10.3168/jds.2006-517

Kristensen, E., Østergaard, S., Krogh, M. A., Enevoldsen, C. 2008. Technical indicators of financial performance in the dairy herd. J Dairy Sci 91:620–631

Krogh, M. A., Enevoldsen, C. 2006. Organizational and educational support to dairy herd health programs. Page 133 in Proc. Int. Soc. Vet. Epidemiol. Econ., Queensland, Australia. http://www.sciquest.org.nz/crusher_download.asp?article=10003577 Accessed Dec. 12, 2008.

Littell, R. C., Miliken, G. A., Stroup, W. W., Wolfinger, R. D. 2006. SAS for Mixed Models 2nd ed. SAS Inst. Inc. Cary, NC. ISBN:1590475003

Loeffler, S. H., de Vries, M. J., Schukken, Y. H. 1999. The effect of time of disease occurence, milk yield and body condition on fertility of dairy cows. J Dairy Sci, 82: 2589- 2604.

Macciotta, N. P. P, Dimauro, C., Catillo, C. G., Coletta, A., Cappio-Borlina, A. 2006, Factors affecting individual lactation curve shape in Italian river buffaloes. Livest Sci, 104:33-37.

Mayeres, P., Stoll, J., Bormann, J., Reents, R., Gengler, N. 2004. Prediction of daily milk, fat, and protein production by a random regression test-day model. J Dairy Sci 87:1925–1933

Nielsen, S.S., Krogh, M. A., Enevoldsen, C. 2009. Time to occurrence of a decline in milk production in cows with various paratuberculosis antibody profiles. J Dairy Sci 92:149-155. doi:10.3168/jds.2008-1488

Nir-Markusfeld, O. 2003. What are production diseases, and how do we manage them? Acta vet scand, Suppl. 98:21-31.

Rasbash, J., Steele, F., Browne, W., Prosser, B. 2004. *A user's guide to MLwiN version 2.0*. London, Institute of Education

Sölkner, J., Fuchs, W. 1987. A comparison of different measures of persistency with special respect to variations of test day milk yields. Livest. Prod. Sci., 16: 305-319.

Thysen, I. 1993. Monitoring bulk milk somatic cell count by a multi-process Kalman filter. Acta Agr Scand Anim Sci, 43:58-63

Wilmink, J. B. M. 1987. Adjustment of test-day milk, fat and protein yield for age, season and stage of lactation. Livest. Prod. Sci., 16:335-348

Woodward, M. (2004) Epidemiology: study design and data analysis 2nd ed. CRC Press, 2000 N.W Corporate Blvd., Boca Raton, Florida. p:21

Østergaard, S., Gröhn, Y. T. 1999. Effect of diseases on test day milk yield and body weigth of dairy cows on Danish research herds. J Dairy Sci 82:1182-1201.

3.6 Evaluation of effects of disease control in a complex dairy herd health management program

Mogens A. Krogh & Carsten Enevoldsen

Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Grønnegårdsvej 2, DK-1870 Frederiksberg C, Denmark

Manuscript.

Evaluation of effects of disease control in a complex dairy herd health management program

M. A. Krogh*[1] and C. Enevoldsen[*]
[*]Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Grønnegårdsvej 2, DK-1870 Frederiksberg C, Denmark
[1] Corresponding author: Mogens A. Krogh, email: mok@life.ku.dk

ABSTRACT

Evaluating the effects of all interventions in a dairy herd, including the effects of various herd health management programs (**HHMP**), is highly relevant. A traditional randomized controlled trial is the gold standard but is likely practically impossible or prohibitively expensive to use for a general evaluation of a HHMP. Generalizability may also be poor because of the dynamics of the production contexts. In this study, we demonstrate an approach for evaluating the effects of a HHMP in the field, specifying an intervention theory for an ongoing HHMP in the context of the Danish dairy industry. As an example, we suggest one statistical model for studying the possible effects on milk production of systematic post-partum examinations of vaginal discharge, which is supposed to improve detection and treatment of **metritis or endometritis. This routine** is one component of the HHMP. The data consisted of 121 herds and 76,953 lactations over a 15-year period. For parity group 1, the negative effects of metritis (with treatment) on 305-d milk production after a normal calving were reduced by 17.3% after enrollment in the HHMP. For parity group 2 and parity group >2, enrollment in the HHMP resulted in a 129 kg and an 80 kg energy-corrected milk yield increase in milk production, respectively. There was some indication that the effect of the HHMP was mediated through improved metritis detection. This study demonstrates the importance of a clear-cut intervention theory although even with a theory, the research question can be too context (herd) specific. In such a case, a within-herd randomized controlled trial study design seems to be the only way to achieve a valid result for a given herd, and acquiring valid results from an observational multi-herd study will be very difficult.

Key words: metritis, herd health management, evaluation

INTRODUCTION

Because of the apparently decreasing profit margin in dairy production, there is an increasing need to evaluate the effects of services offered to and used by the producer. Evaluations obviously must include the effects of herd health management programs (**HHMP**), including very commonly applied disease treatment regimes. It is straightforward to use a randomized controlled trial (**RCT**) approach to assess the effects at the animal level of changing from one drug or feed ration to another, and RCT is the gold standard method in medicine and animal science. However, a HHMP is usually composed of multiple procedures, dynamic performance measurements, and feedback mechanisms, and the herd will usually be the unit of interest. The RCT approach can be practically impossible or prohibitively expensive to use for a general evaluation of a HHMP, and generalizability may also be poor because of the dynamics of the production context.

The estimation of effects of a HHMP is analogous to providing evidence of effects of various programs (interventions) in social systems, which is a large discipline per se. For example, The American Evaluation Association is an international professional association of evaluators 'devoted to the application and exploration of program evaluation to improve their effectiveness' (www.eval.org). Public education or health programs are typical examples. Krogstrup (2011) reviewed and discussed the vast literature on the topic from a new public management perspective and suggested categories of interventions and approaches to evaluate the effects of programs. According to this categorization, a HHMP would be a 'wild problem', which mainly is characterized by a vague definition, lack of an optimal solution, unclear causal mechanisms, and interaction between context and mechanisms. Krogstrup uses the term 'Context Mechanism Outcome' (**CMO**), which means that interventions cause mechanisms, which selectively interact with the case-specific circumstances (the context) and result in effects that differ in different contexts. The consequence of CMO is that each intervention must be evaluated separately, and the applied tools for evaluation are essential. Therefore, the generalizability of the results probably will be poor. Often the modest ambition of such an evaluation will be to explain why some intervention did not work. Ideally, we want to identify causal effects in the intervention to permit prediction. According to Krogstrup (2011), one prerequisite to providing evidence of causal effects is to specify an intervention theory. Based on this theory, we may be able to deduce which components and paths within the intervention program can be used to evaluate causal effects of one or more maybe minor components of the intervention.

The purposes of this study were to demonstrate one approach to specifying an intervention theory for an ongoing HHMP in the context of the Danish dairy industry and suggest one statistical model for estimating possible causal effects given this intervention theory and the context.

MATERIALS AND METHODS

Study Context

During the 1990s, several Danish cattle veterinarians were inspired to implement the key components of an Israeli HHMP described by Nir-Markusfeld (1993). Unless otherwise stated, the following description of the implementation and applications are based on the authors' involvement as teachers, researchers, or users in various stages of this process. One key component of the Danish adaptation of the Israeli program was the local veterinarian's systematic clinical examination of

well-defined groups of cows with an expected high risk of health problems. The clinical examination was performed by a local veterinarian every 7 or 14 days. Fresh cows between 5 and 21 DIM and cows about 10 weeks prior to expected calving were always examined. The veterinarians had access to a wide range of tools for digitizing the data, merging the results of the examinations with standard production and fertility data, and subsequent statistical analyses to support the advisory services (Enevoldsen, 2006).

The HHMP was a service offered to the dairy producer by the veterinarian on a commercial basis (no subsidies). Around 2002, the Danish public veterinary authorities wanted to explore the consequences of lifting the very strict regulation of dairy farmers' access to antibiotics that until then required veterinary presence for a prescription. As a pilot study initiated in 2004, farmers who entered the HHMP described above were permitted to initiate treatment with antibiotics for a limited number of diagnoses without the presence of a veterinarian. The link to the HHMP was expected to ensure sufficient health surveillance and prudent use of antibiotics. Within the pilot project period, requirements were quite strict for delivery of registrations from both farmers and veterinarians to a central database. Similar to the findings of Bennedsgaard (2003), the pilot project revealed numerous errors in the veterinarians' management systems used to collect treatment records and usage of antibiotics and transfer the records into the project database. In 2006, the link between more liberal access to antibiotics and the HHMP was formalized in a new law concerning herd health. In 2009, the legislation was further liberalized so that farmers could treat almost any disease, including parturient paresis and retained placenta, without veterinary assistance. In addition, the very restricted requirements for registration of disease treatments and results from clinical examinations were abandoned. However, participation was still voluntary and unsubsidized.

When the HHMP was implemented in the late 1990s, the farmers' incentives were purely personal or professional interests. In the later stages, the link between HHMP and liberal access to antibiotics and cost reduction probably was an incentive to many farmers. However, the farmers' backgrounds for joining the program were diverse and often different from what the veterinarians expected (Kristensen and Enevoldsen, 2008). Quite detailed guides and manuals for the clinical work in the herds were available, such as guides for scoring body condition or vaginal discharge. Every cow was examined gynecologically by the veterinarian between day 5 post-partum to day 21 post-partum. The examination was performed to estimate the degree of metritis based on an evaluation of the vaginal discharge. The results were given on an ordinal scale from 0 (no metritis) to 9 (severe metritis). However, qualitative studies showed that the use of scores, examination methods, and treatment protocols differed among veterinarians (Kristensen et al., 2008; Lastein et al., 2009). Records of diagnoses and medical treatments of disease symptoms were collected in all herds by means of a limited number of diagnosis codes. The use of these codes was highly herd specific, as described in detail by Vaarst et al. (2002) in a study related to mastitis treatment, but we expect that similar mechanisms occurred in the case of other diseases.

The Intervention Theory

Because we can regard the dairy HMMP described above as a wild problem in the CMO framework, the evaluation of the effects of the program is very complicated. Issues related to the production system (herd) and the basic production unit (cow) pose particularly difficult problems

for program evaluation in dairy production. Often, a HHMP will affect all the animals within a herd or a certain group of animals, leading to the obvious problem that there are no naturally occurring control groups of animals within the herd. Comparing the results with animals in another herd may be impossible or at least problematic for the following reasons: First, herd size and production facilities are rather easy to account for, but differences in management or threshold of detection of disease are virtually impossible to quantify. Second, because the use of the program was voluntary, participants likely had different attitudes and skills than non-participants. Third, a rather simplistic approach of comparing animals before and after intervention within a herd also yields serious obstacles. General or herd-specific seasonal effects must be accounted for. For example, a general increasing trend in milk yield over time must be expected because of genetic progress. However, one of the most important issues to address is the change of recordings because of the initiation of the HHMP (e.g., number of treated cases of mastitis because of change in intensity of monitoring or treatment threshold). In essence, it is extremely difficult to find truly comparable herds.

Instead of that approach, an option for evaluating a HHMP may be to study the possible causal effects of a single component of the HHMP for which we can specify rather simple mechanisms. In this work, we chose to study the possible effects on milk production of systematic post-partum examinations of vaginal discharge, which is supposed to improve detection of **metritis or endometritis. Because these diagnoses are indistinguishable in our setting, we have designated them collectively as 'metritis' in this paper**. We based the study on the following expectations concerning metritis diagnosis and treatments (**MDT**), and we specifically focus on the relationship between metritis treatment and milk yield because milk yield is a very important performance indicator that can be defined objectively. The relationship between MDT and subsequent milk production has been estimated several times; e.g., Fourichon et al. (1999), Bar and Ezra (2005), and Goshen and Shpigel (2006). In our context, it is not sufficient to make a simple comparison of the relationship between milk and MDT before and after the start of the intervention program (HHMP) because the threshold for detections inevitably will change over time. That is, the number of undetected MDT cases in the non-treated group may be reduced. We must also take into account that numerous studies have found clear relationships between stillbirth, dystocia, twin calving, retained placenta, and metritis (Deluyker et al., 1991; Correa et al., 1993; Emanuelson et al., 1993). Metritis with or without previous calving complications is likely to constitute different entities with different effects on milk yield and recovery rates from medical treatment (Pugh et al., 1994). **We also assume that the way the risk factors for metritis are recorded did not change when the examinations began. The prevalences of the risk factors are, however, likely to change with the beginning of examinations because the dairy producer may take action (hopefully) to prevent metritis by reducing the occurrence of these major determinants of metritis.**

In this study, neither the ordinal scale of the metritis score nor the specific threshold for treatment were used for the following reasons: First, the results from the examinations for metritis were not available before the initiation of the HHMP. Second, the examination method changed from farmer-based perception of disease to the veterinarians' systematic examinations of all cows. In diagnostic test terminology, this change will mean that at the time point of initiation of the HHMP, the sensitivity and specificity of metritis detection changed, most likely towards a higher sensitivity and lower specificity. How much it changed is impossible to estimate but is expected to

be dependent on the farmer's previous conception of metritis (herd specific) and the veterinarian's conception of metritis and attitude towards the HHMP (Lastein et al., 2009). Finally, work by Lastein (2009) suggests that the score values of metritis in some instances will be modified if the veterinarian believes that the cow deserves treatment for particular reasons (preferential treatment).

Given the study context and intervention theory described above, we could reduce the wild problem to a statistically tractable problem as formulated with the following specific objective: We want to demonstrate an approach to evaluating the effect on milk production of early post-partum gynecological examinations and associated medical treatments given the context of the Danish HHMP and the mechanisms suggested above. The evaluation is based on a statistical model that estimates the effects on milk yield of time of program initiation, occurrence of calving complications, and MDT.

Data Collection

In December 2009, herds were selected from the VPR database (Enevoldsen, 2006) based on the following criteria: 1) There should be at least 2 yr of registrations of routinely conducted clinical examinations in the herd. 2) The herd should be enrolled in the milk recording program with the standard number of annual test dates (at least 11). 3) There should be valid data 2 yr before the initiation of the herd health program and 2 yr after the initiation of the program. 4) Within this 4-yr period, a minimum of 75% of the calvings should be Danish Holsteins. Based on these criteria, 121 herds and 76,953 calvings were selected.

Data Preparation

On the 11 annual test dates, the amount of milk (in kg), the fat percentage, and the protein percentage were recorded. Based on these values, the test-day energy-corrected milk yield (**ECM**) was calculated using formula 1.

$$\text{kg ECM} = (\text{kg milk} \times (383\ \text{fat\%} + 242 \times \text{protein\%} + 780.8))/3140 \quad (1)$$

The ECM on a given test day was subsequently adjusted with a lactation curve model. The lactation curve model is based on a straight line from calving to 60 DIM and a subsequent straight line decrease in production throughout lactation. This lactation curve model has been applied by Bennedsgaard et al. (2003) and Nielsen et al. (2009) and was run separately for each herd and each parity group. The lactation curve model is given in model 1.

$$ECM_{ij} = \beta_0 + \beta_1 DIMun60_{ij} + \beta_2 DIM60_{ij} + \varepsilon_{ij} \quad (1)$$

where $\beta_0 = \beta_{00} + \mu_{0j}$, $\beta_1 = \beta_{10} + \mu_{1j}$, $\beta_2 = \beta_{20} + \mu_{2j}$

and $\mu_{0j} \sim N(0, \tau_{0j})$, $\mu_{1j} \sim N(0, \tau_{1j})$, $\mu_{2j} \sim N(0, \tau_{2j})$ and $\varepsilon_{ijk} \sim N(0, \sigma_{ij})$.

ECM_{ij} was the kg of ECM on the i^{th} test-day of the j^{th} cow; $DIMun60_{ijk}$ was the i^{th} DIM of the j^{th} cow for DIM 1 to 60. For DIM larger than 60, DIM60 takes the value 0; $DIM60_{ijk}$ was the i^{th} (DIM-60)/245 of the j^{th} cow for 60 DIM to 305 DIM. For values less than 60 DIM, DIM60 takes the value 0. β_0 can be separated into an overall mean (β_{00}) that represents the average milk yield at 60 DIM for that herd and parity, and a contribution from the individual cows (τ_{0j}). β_1 and β_2 were the fixed linear regression coefficients of DIM60 and DIMun60; τ_{1j} the random linear regression coefficient of $DIMun60_{ij}$; τ_{2j} the random linear regression coefficient of $DIM60_{ij}$; and ε_{ij} the random residual component.

Based on model 1, the predicted milk production at 10, 60 and 305 DIM was estimated for each cow. These three predictions allowed us to estimate the total 305-d milk production in lactation (**MPL**) as the area under the curve. The advantage of this approach is that herd-specific culling decisions are accounted for because MPL is calculated even though the lactation is terminated early after calving. Consequently, MPL is not a measurement of the actual milk production but a prediction of the milk production if the cow remained in the herd throughout lactation. That is, this approach makes milk production from individual cows comparable.

Table 1. Proportions of calvings with metritis diagnosis and treatment (MDT) before and after enrollment in the herd health management program split on the dystocia variable (combination of calf condition and calving ease) across the 121 herds

	Parity group 1		Parity group 2		Parity group >2	
	Before enrollment (n = 14,186)	After enrollment (n = 15,264)	Before enrollment (n = 10,178)	After enrollment (n = 11,162)	Before enrollment (n = 12,541)	After enrollment (n = 13,622)
Dead calf	12.7	27.6	14.9	21.9	13.5	26.9
Twins –alive[1]	10.8	31.2	12.8	34.9	20.3	37.4
Calf –alive	6.4	14.5	4.8	9.6	5.4	11.0
Calf alive – difficult calving[1]	12.5	27.0	14.7	17.9	5.6	17.7
Calf alive – veterinary assisted[1]	7.5	28.4	8.7	2.4	10.3	20.3
Calf alive – missing code[1]	7.6	14.4	7.1	9.7	7.3	11.2

[1]These 4 groups together account for 10% of the total numbers of calving and less than 10% of the total MDT cases.

In Denmark, it is mandatory to record the sex of the calf and the condition of the calf/calves after calving using an 8-point categorical scale. In essence, two values describe a living calf and six describe variations of a dead calf. Calving ease can be recorded but is not mandatory, using a 5-point categorical scale in which 1 is an unassisted calving and 5 is a caesarean section. Based on this information, a new variable was created to describe dystocia: 81.2% of the calvings were normal with a living calf, and 9.6% of the calvings resulted in calving with a dead calf.

Cows that had a recorded diagnosis code of metritis with medical treatment within the first 21 d post-partum were classified as cows with MDT. The frequency of MDT is given in Table 1 to verify the assumption that the detection and treatment threshold for metritis will be lower after enrollment in the HHMP (see intervention theory above). Table 1 also shows the distribution of MDT within parity groups, and before and after enrollment in the HHMP.

Table 1 shows that calvings with a normal calf before enrollment in the HHMP had a risk of MDT that was approximately half the risk of MDT at calvings with a dead calf. The risk of MDT approximately doubled after enrollment in the HHMP for both calving with a dead calf and having a normal calving, except for calving with a dead calf in parity group 2.

Statistical Models

A multilevel random regression model was specified for parity groups 1, 2, and >2 separately. The dependent variable was MPL. The independent variables were year of the calving (YEAR), season as quarter of the year of the calving (SEASON), a variable to describe whether the calving occurred when the herd was enrolled in the HHMP, the dystocia variable (**DYS**), and MDT. All two-way interactions and the three-way interaction among HHMP, DYS, and MDT were included. The baseline model is given in model 2.

$$MPL_{ij} = \beta_0 + \beta_1 SEASON_{ij} + \beta_2 YEAR_{ij} + \beta_3 MDT + \beta_4 DYS_{ij} + \beta_5 HHMP_{ij} + \beta_6 MDT \times HHMP_{ij} + \beta_7 MDT \times DYS_{ij} + \beta_8 DYS \times HHMP_{ij} + \beta_9 MDT \times HHMP \times DYS_{ij} + \varepsilon_{ij} \quad (2)$$

where $\beta_0 = \beta_{00} + \mu_{0j}$ and $\varepsilon_{ijk} \sim N(0, \sigma_{ij})$.

MPL_{ij} was the predicted 305-d kg of ECM produced in the lactation of the i^{th} cow of the j^{th} herd; $SEASON_{ij}$ was season of calving for the i^{th} cow of the j^{th} herd (4 categories). $YEAR_{ij}$ was the year of the calving for the i^{th} cow of the j^{th} herd (15 categories). MDT_{ij} was treatment of metritis at the calving of the i^{th} cow in the j^{th} herd (2 categories). DYS_{ij} was dystocia at the calving of the i^{th} cow in the j^{th} herd (6 categories). $HHMP_{ij}$ was whether the calving of the i^{th} cow in the j^{th} herd was included in the herd health program (2 categories). MDT×HHMP×DYS, MDT×HHMP, HHMP×DYS, and MT×DYS are the three-way and two-way interactions between the main effects. β_0 can be separated into an overall mean (β_{00}) that represents the average MPL across herds and a contribution from the individual herds (τ_{0j}). β_1 to β_9 were the fixed class coefficients of SEASON, YEAR, MDT, DYS, HHMP, MDT×HHMP, MDT×DYS, DYS×HHMP, and MDT×HHMP×DYS, respectively, and ε_{ij} the random residual component.

The analyses were performed with the mixed procedure (Little et al., 2006) in SAS using maximum likelihood estimation. The deviance test was used to test for fixed effects in the model. The baseline model was reduced with backward elimination. Model assumptions were investigated by histograms, Q-Q plots, and residuals vs. predicted values. Variance homogeneity across herds was evaluated by boxplots of the conditional studentized residuals.

RESULTS

Descriptive Statistics

Figure 1 shows the number of calvings included in the analysis for each year. It is obvious that the years 1995 to 1997 were entirely estimated from calvings that had not been included in the herd health program and that the years 2007 to 2009 almost entirely were from cows that had been enrolled in the herd health program. Note that any one herd only will contribute to four calendar years. Figure 1 shows that a large proportion of new herds were enrolled in the program in 2005.

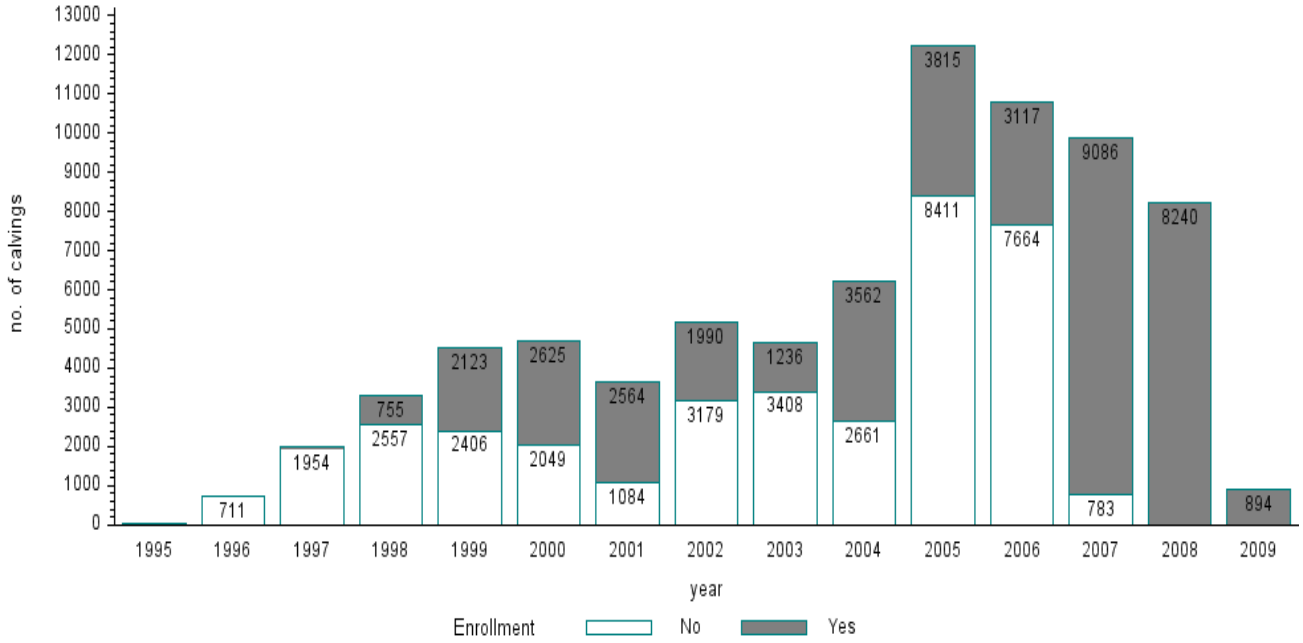

Figure 3: Distribution of the 76,953 calvings from 121 herds during the study period

Table 2 gives key characteristics of what happened in these 121 herds when comparing the 2-yr period before enrollment in the HHMP with the 2-yr follow-up period. In 25% of the herds, the number of calvings increased with more than 38 calvings every 2 years. Thus, there was an increase in herd size in a substantial proportion of herds during the study period. Because the average number of calvings/year before enrollment in the HHMP was around 150, it is unlikely that improved reproduction alone can account for this increase. Table 2 also shows that the average difference in days to first insemination hardly changed.

Table 2. Differences in key characteristics at herd level (N = 121) before and after enrollment in the herd health management program

	Lower quartile	Median	Upper quartile
Difference in number of calvings	-1	13	38
Difference in avg. 305-d milk production (energy-corrected milk), kg	-16	214	532
Difference in days to first insemination	-2	1	4
Odds of receiving a metritis treatment after vs. before enrollment in the Herd Health Management Program	1.5	2.6	5.6

For half of the herds, the milk production increased by more than 200 kg ECM, whereas 25% of herds had no increase in average milk production in lactation. Although this description is crude and to some extent may be related to differences in herd demographics as with distribution of parity groups, it indicates that most herds experienced an increase in MPL over a 4-year period.

As expected, the odds of receiving a MDT after versus before enrollment indicated a substantially higher risk of receiving a MDT in almost every herd (lower quartile is 1.5). This finding supports the expectations concerning increased detection intensity.

Statistical Analyses

In the analyses of all three parity groups, the effects of year, season, and dystocia were highly significant ($P < 0.001$). For parity group 1, the three-way interaction of MDT×HHMP×DYS was significant ($P = 0.0015$). Hence, the baseline model could not be reduced. For parity group 2, the three-way interaction MDT×HHMP×DYS ($P = 0.29$) and MDT×HHMP ($P = 0.43$) were removed. The two-way interactions MDT×DYS and HHMP×DYS were both significant ($P = 0.008$ and $P = 0.045$, respectively). The final reduced model for parity group 2 then contained the interactions MDT×DYS and HHMP×DYS, their main effects, and YEAR and SEASON. For >2 parity group, the three-way interaction MDT×HHMP×DYS ($P = 0.94$) and the interactions HHMP×MDT ($P = 0.94$) and HHMP×DYS ($P = 0.85$) were removed from the model. The final reduced model for parity group >2 consisted of the interaction DYS×MDT ($P < 0.0001$), the main effects of the interaction, the main effect of HHMP ($P = 0.03$), and SEASON and YEAR. Model control indicated variance homogeneity across herds.

Figure 2 illustrates the parity-specific effect of year of calving adjusted for enrollment in the HHMP, season, and dystocia and indicates a clear increase in milk production from 1998 to 2004. However, it is important to note that the herds that contributed to year 2000 were not the same as those that contributed to year 2005.

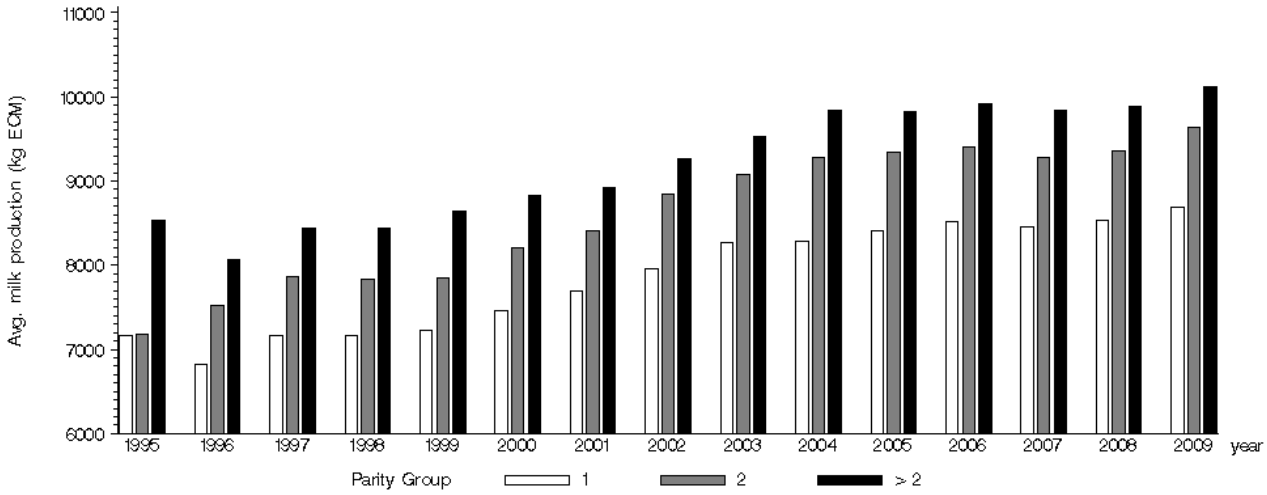

Figure 4: Parity-specific effect of year of calving on milk production of 76,953 calvings in 121 herds adjusted for enrollment in the herd health management program, dystocia, and metritis treatment.

Table 3 shows least squares mean differences for selected effects. Only effects related to the dystocia categories of 'Normal Calving' and 'Dead Calf' are shown because these two categories accounted for more than 90% of the observations.

For parity group 1 with a normal calving, a MDT cow produced 192 kg ECM less than a non-MDT cow if the latter cow was not enrolled in the HHMP. This difference was reduced to 69 kg ECM in the HMMP. Non-MDT parity group 1 cows produced the same amount of milk (7 kg ECM difference) regardless of whether the cows were enrolled in the HHMP or not. MDT cows produced

116 kg ECM more if the treatment took place after enrollment in the HHMP. All differences were non-significant (P > 0.11) for parity group 1 cows that calved with a dead calf.

Among cows from parity group 2 that had a normal calving, MDT cows produced 91 kg ECM less than non-MDT cows. Cows enrolled in the HHMP that had a normal calving produced 127 kg ECM more than cows not enrolled. For parity group 2 cows that had a dead calf, MDT cows produced 348 kg ECM less than non-MDT cows. There was no effect of being enrolled in the HHMP if the calving resulted in a dead calf.

For cows in parity group >2 with a normal calving, MDT cows produced 247 kg ECM less than non-MDT cows. A calving that resulted in a dead calf in parity group >2 cows with MDT produced 194 kg ECM less than non-MDT cows. In parity group >2, the effect of being enrolled in the HHMP was 80 kg ECM.

Table 3. Least squares mean differences in energy-corrected 305-d milk production for selected categories of Herd Health Management Program (HHMP), dystocia, and metritis treatment[1]

	HMMP	Dystocia	Metritis Treatment	Effect (kg ECM)
Parity Group 1	Not enrolled	Normal calving	Treated vs. untreated	-192***
	Enrolled	Normal calving	Treated vs. untreated	-69***
	Not enrolled vs. enrolled	Normal calving	Untreated	-7, NS[2]
	Not enrolled vs. enrolled	Normal calving	Treated	116*
	Not enrolled	Dead calf	Treated vs. untreated	7, NS
	Enrolled	Dead calf	Treated vs. untreated	-80, NS
	Not enrolled vs. enrolled	Dead calf	Untreated	-59, NS
	Not enrolled vs. enrolled	Dead calf	Treated	-146, NS
Parity Group 2		Normal calving	Treated vs. untreated	-91*
	Not enrolled vs. enrolled	Normal calving		127***
		Dead calf	Treated vs. untreated	-348**
	Not enrolled vs. enrolled	Dead calf		-18, NS
Parity Group >2		Normal calving	Treated vs. untreated	-247***
		Dead calf	Treated vs. untreated	-192*
	Not enrolled vs. enrolled			80*

[1]The estimated differences are for each line highlighted with bold. The other categories are fixed. Empty cells mean that the category of the variable did not influence the estimated difference.

[2]NS: Non-significant; *P < 0.05; **P < 0.01; ***P < 0.001.

DISCUSSION

The highly significant three-way interaction among HHMP, dystocia, and MDT in parity group 1 shows that the effect of the HHMP depended not only on the dystocia occurrence but also on whether the cows had MDT. In parity group 1 with cows having a normal calving, the negative effect of MDT was 192 kg ECM before enrollment in the HHMP and 69 kg ECM after enrollment. The negative effect of 192 kg ECM can be considered a baseline milk loss of a case of MDT. This loss associated with a MDT case was then reduced after enrollment in HHMP. However, as hypothesized above, it is very likely that MDT cases before and after enrollment do not describe the same clinical condition. Because of a reduced detection threshold, MDT cases before enrollment in the HHMP probably were more severe with a larger milk loss than MDT cases after enrollment in the HHMP. Thus, it is possible that what we found here was merely a dilution effect arising from

inclusion of cows that had experienced minor or no milk production loss regardless of treatment in the MDT group.

We know from Table 1 that the average risk of MDT before enrollment in the HHMP was 6.4% and 14.5% after enrollment in parity group 1 having a normal calving. Based on these figures, the total milk production loss in an average herd can be calculated before and after enrollment in the HHMP. Such a calculation shows that the total milk production loss related to MDT was reduced by 17.3% after enrollment, indicating that non-MDT cases existed before enrollment and that these cases benefitted from being diagnosed and treated after enrollment in the HHMP. Looking at the non-MDT cases before and after enrollment in the HHMP, we find that there was no difference in milk production (7 kg ECM). A difference was expected because the population of cows before enrollment included non-MDT cases that could have experienced some milk production loss, whereas the population after enrollment did not include this group of cows. We expect that the milk production loss of these additionally diagnosed and treated cows was somewhat less than for cows treated before enrollment in the HHMP. Because the magnitude of the production loss was smaller, it is likely that the approximately 8% (14.5% after enrollment vs. 6.4% before enrollment) of cows additionally diagnosed and treated after enrollment in the HHMP was too small to be detected when diluted in the >80% normal cows. As for the MDT cases, there was a positive effect of enrollment in the HHMP (116 kg ECM), which corresponded nicely with the difference between MDT cases and non-MDT cases before and after enrollment in the HHMP. The possible explanation for this effect has been discussed above.

If the calving resulted in a dead calf in parity group 1, none of the effects were significant. The risk of being a MDT case was 12.7% before enrollment in the HHMP among calvings that resulted in a dead calf. After enrollment, the corresponding risk was 27.6%. Possible reasons for the lack of effects despite this increased detection intensity could be that farmers were aware of problems with heifers that experienced a calving with a dead calf. The consequence would be that the proportion of non-MDT cows that had experienced a milk production loss was small compared with calvings with a normal calf. In addition, there could be other diseases of the genital tract (injuries in the vulva or vagina) that caused pain or infections that were treated and recorded. The treatment of these other infections could have some preventive effect on the risk of subsequent metritis. Hence, it is possible that a cow had a case of metritis that was never recorded. Routine examinations of cows will tend to focus on metritis detection, and the probability of having a diseased cow recorded as metritis will be greater despite the possible co-occurrence of other diseases.

In parity group 2, the negative effect of MDT was 91 kg ECM if the calving was normal regardless of enrollment in HHMP. This effect is about half of the corresponding negative effect for parity group 1 cows. The positive effect of being enrolled in the HHMP was 127 kg ECM regardless of MDT if calving was normal. This finding can be interpreted as prevention of a production loss of 127 kg ECM if the parity group 2 cows are enrolled in the HHMP and the calving was normal. Because the interaction between dystocia and HHMP was significant, the effect of the HHMP is essentially modified by dystocia, which provides support for our hypothesis that the effect of the HHMP on 305-d ECM is related to the gynecological examination and metritis detection.

If the calving of a parity 2 group cow resulted in a dead calf, the negative effect was 348 kg ECM regardless of enrollment in the HHMP. This effect is much greater than for parity 1 group cows, which could indicate that the mechanisms that led to a dead calf in parity group 2 were quite different from the mechanisms that led to a dead calf in parity group 1. Dystocia caused by problems with the relative size of the cow and the calf is highly associated with heifers. The risk of MDT before enrollment in the HHMP was 14.9% and 21.9% after enrollment if the parity group 2 cow had a dead calf. This increase was rather modest compared to the other parity groups (Table 1) and suggests that the type of cow with MDT before and after enrollment in the HHMP was similar or that treatments had virtually no effect.

Parity group >2 cows with MDT after a normal calving produced 247 kg ECM less than non-MDT cows regardless of enrollment in the HHMP. This milk production loss is in the same range as for parity groups 1 and 2, especially if the milk production loss is seen relative to the entire production of the parity groups. After a calving that resulted in a dead calf, the milk loss was 192 kg ECM, which is substantially smaller than for parity group 2. A possible reason could be that other diseases like parturient paresis were the cause of the dead calf. These diseases may not cause pain or infections to the same extent as diseases in younger cows. For parity group >2, there was a positive main effect of enrollment in the HHMP (80 kg ECM), but this could not be related to MDT.

The milk loss associated with MDT as identified in this study can be compared to the work of Goshen and Shpigel (2006). In a RCT setting, they found that milk production losses from metritis could be totally avoided by treatment. Actually, there was an increased milk production in 305-d lactation for the treated group. They also estimated a 300–500 kg milk (not ECM as in this study) production loss associated with untreated metritis. That value represents a substantially larger effect than what we found, but because we did not have animals diagnosed and left untreated in our study, this 'true' milk loss resulting from metritis cannot be estimated here. However, we still identified an approximately 200 kg ECM production loss related to the diagnosed and treated cases of metritis. A possible explanation could be differences in treatment protocols. A common metritis treatment in Denmark is 1/5 the amount of oxytetracycline used in Israel. In addition, some Danish veterinarians treat mild metritis cases with prostaglandins alone. The effect of this treatment at this stage of lactation is debatable (e.g., Archbald et al., 1990). The consequence of metritis on milk production has also been estimated by Østergaard and Gröhn (1999). They found no effect of metritis on first-parity cows and a 239 kg ECM production loss in the first 182 d of lactation in second and older parities compared to 'healthy' cows. In our study, we found a substantial effect of MDT on 305-d milk production in parity group 1 in contrast to the findings of Østergaard and Gröhn (1999), whereas their estimate of second and older parities is comparable to our findings. One possible reason could be that their results are based on three research dairy herds that may differ from ours with respect to environment and management.

Our presentation of the intervention program (the HHMP), its context, our intervention theory, and the discussion of our results demonstrate that evaluation of a HHMP in dairy herds is complex. Despite a relatively clear intervention theory about the relationship between dystocia and detection and treatment of metritis before and after enrollment in the HHMP, the problem of evaluation of this minor part of the HHMP still remains a wild problem, or at least complicated. In terms of the

CMO concept described in the introduction, we believe we had a good understanding of the mechanisms and outcome. However, it is also clear that we did not have sufficient information about the context (herds) to fully interpret the results. This raises the question of whether or not it is possible to evaluate a complex HHMP in a multi-herd observational study. The answer will be that it is possible if the effects of the HHMP are strictly founded in biological or pathological mechanisms that manifest themselves equally across herds and cows. If the contexts are very important and there are many management factors that can influence the outcome of the results, then multi-herd analysis will most likely be of limited relevance. If context is important, a within-herd RCT study design seems to be the only way to achieve a valid result for a given herd.

REFERENCES

Archbald, L.F., T. Tran, P.G.A. Thomas, and S.K. Lyle. 1990. Apparent failure of prostaglandin F2a to improve the reproductive efficiency of postpartum dairy cows that had experienced dystocia and/or retained fetal membranes. Theriogenology 34(6): 1025-1034

Bar, D. and Ezra S. 2005. Effect of common calving diseases on milk production in high yielding dairy cows. Isr. J. Vet. Med. 60:106-111.

Bennedsgaard, T. W., C. Enevoldsen, S. M. Thamsborg, and M. Vaarst. 2003. Effect of mastitis treatment and somatic cell counts on milk yield in Danish organic dairy cows. J. Dairy Sci.86:3174-3183.

Bennedsgaard, T. W. 2003. Reduced use of veterinary drugs in organic herds. Ph.D dissertation. The Royal Veterinary and Agricultural University, Frederiksberg, Denmark. Pages 11-19.

Correa M. T., H. Erb, and J. Scarlett. 1993. Path analysis for seven postpartum disorders of holstein cows. J. Dairy Sci. 76:1305-1312

Deluyker, H. A.,J. M. Gay, L. D. Weaver, and A.S. Azari. 1991. Change of milk yield with clinical diseases for a high producing dairy herd. J. Dairy Sci. 74:436.

Emanuelson, U., P. A. Oltenacu, and Y. T. Grohn. 1993. Nonlinear mixed model analyses of five production disorders of dairy cattle. J. Dairy Sci. 76:2765-2772.

Enevoldsen**,** C. 2006. Epidemiological tool for herd diagnosis. Page 376-383 in *Proceedings of XXIV World Buiatrics Congress: 15–19 October 2006; Nice*. Edited by Navetat H, Schelcher F.

Fourichon C., H. Seegers, N. Bareille, and F. Beaudeau. 1999. Effects of disease on milk production in the dairy cow: a review. Prev. Vet. Med. 41(1):1–35. doi:10.1016/S0167-5877(99)00035-5

Goshen T., and N. Y. Shpigel. 2006. Evaluation of intrauterine antibiotic treatment of clinical metritis and retained fetal membranes in dairy cows. *Theriogenology* 66: 2210-2218

Kristensen, E., D. B. Nielsen, L. N. Jensen, M. Vaarst, and C. Enevoldsen. 2008. A mixed methods inquiry into the validity of data. Acta. Vet. Scand. 50:30

Lastein, D. B., M. Vaarst, and Enevoldsen C. 2009. Veterinary decision making in relation to metritis - a qualitative approach to understand the background for variation and bias in veterinary medical records. Acta. Vet. Scand. 51:36 doi:10.1186/1751-0147-51-36

Littell, R. C., G. A. Miliken, W. W. Stroup, and R. D. Wolfinger. 2006. SAS for Mixed Models 2nd ed. SAS Inst. Inc. Cary, NC. ISBN:1590475003

Nielsen, S. S., Krogh M. A., and Enevoldsen C. 2009. Time to occurrence of drop in milk production in cows with various paratuberculosis antibody profiles. J. Dairy Sci. 91:149-155

Pugh, D. G., M. Q. Lowder and J. G. W. Wenzel, Retrospective analysis of the management of 78 cases of postpartum metritis in the cow. Theriogenology 42(3):455-463

Vaarst, M., B. Paarup-Laursen, H. Houe, C. Fossing, and H. J. Andersen. 2002. Farmers' choice of medical treatment of mastitis in Danish dairy herds based on qualitative research interviews. J. Dairy Sci. 85:992–1001

4 Discussion

The overall objective of this thesis is to suggest a coherent concept for management of data for health performance measurement that is suitable and sufficient for the diverse contexts of industrialized Danish dairy herds and associated veterinary practices. Such a concept is outlined in section 3.1 with details in sections 3.2 to 3.6.

In the veterinary literature, the term 'surveillance' is widely used and defined as some active use of information in contrast to 'monitoring', which implies a simple collection of data (Stärk and Salman, 2001). This use of surveillance seems to have the same meaning as the term 'health performance measurement' in this work. In herd management science, monitoring appears to have the same meaning as surveillance in veterinary science. Because performance measurement seems to be widely used in social science and business management (Krogstrup, 2011) and the term intuitively signals some active process including evaluation, I suggest it is preferable to use the term health performance measurement.

The work in chapter 3 gives examples of health performance indicators that demonstrate the key principles of health performance measurement. However, I have made no attempt to provide a comprehensive list of suggestions for a minimum or prioritized list of health performance indicators. Kristensen et al. (2008) provides a prioritized list of eight key performance indicators based on financial criteria. Nir-Markusfeld (2003) also suggests reducing the number of performance measurements to a minimum. At first, it seems completely rational to restrict health performance measurement to a few indicators. However, my experience from numerous herd analyses during the years with some of the tools presented in this thesis shows that a large number of health indicators is needed. For planning or economic analysis, few indicators or major results indicators may suffice. For the herd veterinarian's diagnostic approach to exploring exceptional variations ('identification of causes of signs of herd health problems'), numerous indicators are needed. As an example, if ketosis is suspected by clinical manifestations, it is relevant to explore time series of risk factors like components of lactation curves in previous lactation, body condition scores at various stages of lactation, dry period length, other metabolic disorders, and metritis. The exploration may show that some or many factors are irrelevant for problem-solving in that particular herd at that particular point in time. However, at another time in the same herd, other indicators may be relevant. Consequently, we need a large number of indicators for diagnostic work. A continuous screening of numerous process behavior charts with health performance measurements probably will reveal virtually all emerging health problems and provide a solid platform for identification of removable causes by means of the tools described in chapter 3. This process can be considered as an effect-focused practice as defined by Krogstrup (2011).

In this thesis I claim it is a major obligation for the practicing veterinarian to continuously perform health performance measurement of production and health in the individual herds. The veterinarian can make two errors: 1) Interpret noise (routine variation) as if it were a signal of exceptional variation. 2) Fail to detect exceptional variation when it is present. The challenge is to strike the

best balance between these two mistakes. The criteria for 'best balance' depend on the type of problem that is addressed. In some cases, detecting every problem is essential (e.g., a sign of a disastrous disease like Foot and Mouth Disease); in other situations, false-positive reactions should be few because too many reactions on non-cases can be very costly. De Vries and Reneau (2010) discuss this issue in some detail based on a review of applications of the control charts in animal production. Their main conclusion is that an actual search for the true causes of exceptional is very seldom done. Virtually all studies are based on simulations, which may be problematic because simulations usually are based assumptions about distributions, which we do not know in a real life setting.

I will argue that the importance of false positive or false negative signals from a PBC depends on whether we work in the phase 1 or the phase 2 settings described by Woodall (2000). In phase 1, the purpose of using the charts is to learn about the process, including statistical issues like distributions. That is, phase 1 is a retrospective evaluation that integrates detecting 'atypical patterns' and searching for explanations of exceptional variation (primarily inductive reasoning). The derived explanation may support action to remove the cause(s) of the problem. This also implies that false positives probably will be detected. Maybe the intervention in terms of continuous searching for causes of exceptional variation will reveal false negatives as well.

In phase 2 the form of the distribution is assumed to be known along with values of the in-control parameters, and the process closely resembles repeated hypothesis testing based on a planned prospective and sequential sampling over time to detect changes from an in-control process (primarily deductive reasoning). Much work, process understanding, and process improvement is often required in the transition from phase 1 to phase 2 (Woodall, 2000). If conducted appropriately, a phase 2 process will allow correcting-action to be taken more or less directly based on the derived signals. In addition, false positive risk and average run length until an out of control situation can be estimated. Such estimates are essential for estimating the financially optimal decision criteria, including control limits. The requirements for a phase 2 process may be met by the measurement systems in advanced AMS but usually the algorithms in such systems are unknown (business secrets). Sequential randomized controlled trials in a herd could be seen as an ideal phase 2 process.

Continuous exploration of numerous process behaviour charts may seem practically hopeless. Fortunately, the initial screening of such charts can be automated because the criteria for defining signals are clear and probably robust. Then, interpretation of signals can be restricted to a limited number of charts with clear signals. Setting up a hierarchy of indicators defined by relatively few major problem complexes (e.g., ketosis or poor fertility) may also facilitate the diagnostic process. Work remains to be done to organize automation of this diagnostic screening. The inherent problem concerning multiple 'testing' (increased risk of false positives) must also be addressed. For the context described in this thesis, the practice level seems to be a suitable organizational level for establishment of a framework for these development tasks.

A continuous screening of numerous process behaviour charts with health performance measurements probably will reveal virtually all emerging health problems and provide a solid platform for identification of removable causes by means of the tools described in chapter 3. This process can be considered as an effect-focused practice as defined by Krogstrup (2011). When the process is conducted within practice combined with deep insight into the herd's context, including the manager's values and attitudes, possible wild problems may be reduced to tame problems.

Because of the constant development in computer and communication technology, the number of variables recorded in dairy herds and stored in large databases will continue to increase. This development in data availability gives opportunities for more advanced statistical modeling. My own experience from about 10 years of work with development and administration of herd health programs in the field and experience from plant production management (McCrown et al., 2002) show that sophisticated and potentially powerful methods are rarely applied. Lack of transparency could be one reason why proper principles and techniques are not implemented in herd health programs despite advances in the epidemiological techniques. Multilevel random regressions and various Bayesian methods are now commonly applied in scientific articles and also applied in some software for management support. Herd managers and most advisors may not be able to understand the information derived from these tools. Consequently, they may be reluctant to use the information. The gap may be too big between those who develop the tools for the herd health programs and the decision makers in the herds who have to take actions accordingly. Other reasons for possible failure of herd health programs may be: they are often not used correctly; herd managers may lose interest in them; what the managers see within their herd does not correspond with the assumptions of the program (Sørensen, 1990). Based on the insight gained through this Ph.D. study, I suggest that more researchers should focus intensively on the intervention theory or the socio-biological hypotheses that should drive the subsequent data analysis. In-depth understanding of the nature of the basic recordings and potential biases is a prerequisite. For instance, an insemination is not just a concrete event; it is based on the farmer's ability to detect the heat and the active decision to breed the cow. The decision to breed is affected by more or less person-specific perceptions of the cow's condition and the actual situation in the herd. It is a challenge for the veterinarian to understand processes like these (e.g., Lastein et al. 2009 and Vaarst et al. 2002), and even with extremely advanced statistical techniques, such complex relationships are difficult to handle. A practical alternative to seeking a statistical solution to this kind of wild problem will be to explore the values and constraints in management to identify a simpler solution.

The preceding examples demonstrate that even if theoretically very efficient systems for health performance measurements based on numerical data are established, the herd veterinarian also needs to continuously evaluate non-structured (qualitative) information about the herd and its management. Possible tools are briefly covered in section 3.2. Ordinary but systematic dialogue at the regular herd visits probably will reveal the major issues. However, in the future, there likely will be a need for more professional approaches to this aspect as the herds grow and the human relationships become increasingly important. The recent increase in legal regulation probably will

further intensify the need for knowing the effects on management decisions of the positive or negative incentives arising from legal constraints.

The veterinary authorities use legislation to change farmer behavior. Sometimes, the legislation can lead to side-effects that counteract the purpose of the legislation. In addition, the changes in farmer behavior can interfere with the nature of (health) recordings. There is a serious risk that the authorities' attempt to measure, for instance, welfare based on farm-based recordings can distort the recordings making them useless or less useful for herd management. For example, farmers can delay the recording of the calving until day 5-7 post partum. If the calf dies within this time period it can then be recorded as a stillbirth, despite the fact that the calf was alive at birth. Farmers would have an incentive to do so because stillbirths are not included in public authorities' welfare measurements but early deaths of calves are. If such actions are unknown for the herd veterinarian, actions related to improvement of the situation will be unsuccessful. A deep understanding of the context of the herd can prevent the latter. There is a need to make veterinary authorities understand these mechanisms prior to imposing legal restraints on herd management.

Kristensen & Jakobsen (2010) suggest the term 'social epidemiology' as an approach to integrating techniques from social sciences and traditional 'number-focused' epidemiology. Vaarst et al. (2002) and Ellis-Iversen et al. (2010) also emphasize the importance of knowledge about farmer attitudes and behavior in health management. Social techniques like the Q-method have been applied to reveal and classify attitudes and values among dairy farmers and veterinarians (Kristensen & Enevoldsen, 2008). Additional advantages could be had by subsequently incorporating the results from the classifications into traditional statistical methods like analysis of variance or regression. Hopefully, the social science results will explain a significant proportion of variance in production outcomes at herd level.

The work with and in the context described in section 3.2 has shown that the relationship between a signal from a tool (such as a statistical model) and an underlying observation (e.g., an individual cow) must not be lost in the process. Maintenance of this relationship strengthens farmer compliance and context knowledge, which is important for revealing causes of exceptional variation and opportunities for their removal. More work is needed to further develop tools linking signals with individuals in the underlying process. Such investigation and development will require formal comparison and integration of the available tools for health performance measurement.

I argue above, that context-knowledge probably will eliminate most false positive signals. However, across-herd analyses may still be relevant to evaluating more general interventions, as described in section 3.6. With increasing size of practices, standardization of the veterinarians' clinical routines may be poorer, or at least, an estimation of the degree of standardization is warranted. In that situation, the problem entity may best be designated as a wild problem, which also can be addressed by approaches as described in sections 3.2, 3.3, and 3.6. Thorough across-herd analyses with these tools can serve to improve the quality of the services provided by the veterinarians in practice. These tasks require employment of people with statistical expertise. This expertise may also be

useful for designing randomized controlled trials, which can be the best option for evaluating the effects of the substantial number of medical interventions available to veterinarians.

RCT is essential in the concept of 'Evidence-based Medicine' (EBM), but a study like demonstrated in section 3.6 is also considered as a contribution to EBM. EBM has recently attracted growing interest in veterinary practice (Ruegg 2010). However, in real life, the EBM concept is not that easy to apply. As shown in this thesis, some (most?) situations concerning evaluation of disease treatment are often highly context-specific. Hence, the generalizability of the published results (peer reviewed or not) can be questionable. Within-herd trials, including RCT, can be a useful approach in the future. However, the design, analysis, and interpretation of the results of these trials require people trained in epidemiology and statistics with an additional thorough understanding of the production systems.

It is my experience that most dairy herd management systems in Denmark and elsewhere merely present collections of recordings without convincing attempts to support evaluation of the performance. Consequently, more consistent use of the methods described in this thesis should provide options for improving efficiency. However, there are few studies that attempt to evaluate the added value of improved data recording, data management, statistical analysis, and follow-up. This thesis demonstrates the inherent problems associated with evaluation studies.

The work initiated with the analysis of the lactation curve shows perspective in terms of benchmarking. It has been established by simulation that the milk production is the single most economical important characteristic of the dairy herd (Kristensen et al., 2008). As it can be considered straight forward to increase the entire milk production level of the herd, much less emphasis have been given on reducing cow to cow variation in milk production and/or improving the persistency. The concept presented in section 3.5 will definitely mean an improvement of the information available for herd health management. However, at least one criticism can be made about the proposed lactation curve model: Is it reasonable that the peak of the lactation curve is set at 60 days in milk for all herds and all parities? Because the purpose of the lactation curve model is to provide estimates for benchmarking, the model must handle each herd the same way. Also, it is a constraint that there currently is only 8-10 data points to describe more than 300 days of lactation in which we must allow for a highly variable early stage of lactation (increasing or even decreasing phase), a 'peak', and long decreasing phase of milk production that is less variable than the early phase. Because within-lactation residual variance is up to 15 (Krogh and Enevoldsen, 2012), the location of the actual peak is also highly variable. However, by using a systematic examination of the residuals for the first 3 to 4 months of lactation, a herd-level indicator of peak-location could be estimated. Such an 'average days to peak' variable might be valuable as an additional variable for benchmarking.

In addition to milk production and ketosis, there are numerous areas of dairy herd health management where we have multiple diagnostic tests to describe the same (or almost same) condition. For example, to describe mastitis the following diagnostic tests are commonly applied:

Milk culture, California Mastitis Test at quarter level, SCC at cow level, PCR of a number of udder pathogens at cow-level or in bulk milk (Grauber et al., 2007; Studer et al., 2008). In near future, additional mastitis indicators at cow level will be recorded automatically. In AMS, recordings of milk production at every milking, various changes in the milk composition, milk temperature, and, in specialized cases, inflammatory enzymes are already recorded. The multivariate methods described in section 3.2 and 3.4 are obviously relevant to provide indicators of latent variables like, for instance, a '*Staph. aureus*'-cow.

5 Conclusions

I use the term health performance measurement because the combination of the words measurement and performance directly signals an evaluation of the current state of the system. That is, an evaluation of how the process of interest is functioning at any time, which does not make sense without some criteria for distinguishing between acceptable or unacceptable. This is in contrast to systems that merely present collections of recordings without any attempt to evaluate.

The context for health performance measurement in the dairy herd is characterized with the following, which is essential for the choice of methods for measurement:

- Numerous indicators are needed for sufficient health performance measurement in a dairy herd because we need to be able to diagnose numerous unpredictable disease complexes or other signs of ill-health, including suboptimal productivity.
- Measurements are taken at time intervals from milliseconds to years and measured at levels from udder quarter to herd or veterinary practice level.
- Most of the measurements and the criteria for evaluation can be very context-dependent, and contexts may be variable; in particular if measurements are related to legal regulations or other incentives affecting the farm owner's or the personnel's attitudes and management actions.

Consequently, it is necessary for each herd and the associated veterinary practice to have measurement concepts and a data analysis setup that can adapt to each specific context and organizational level. Overall, it is very unlikely that one specific type of measurement setup fits all contexts.

My suggestion for an applicable and adaptive overall strategy is an initial non-parametric explorative phase 1 and a subsequent parametric phase 2:

Phase 1 – exploration of the context:

- Quantitative time series analysis (numerical data): The process behavior chart (PBC) described and discussed in detail in section 3.2 is well suited for the initial analysis of the structured (numerical) data produced in the complex and highly variable context described above. A context like the one described in this thesis requires a flexible, adaptive and comprehensive analysis with very few restrictions including a minimum of assumptions about statistical distributions. Sufficiently sensitive intervention criteria (that define signals) can be specified although the risk of false positive signals may be considerable.
- Qualitative time series analysis (non-structured data): It is essential that a signal from the PBC is followed immediately by a qualitative search for and evaluation of possible causes. Sufficient context-knowledge and the transparency of the PBC probably limit the number of possible causes and reduce false positive signals to a minimum. The sufficient cause(s) may be identified. If not, at least an approach for further causal analysis in a phase 2 should be specified.

Phase 2 – parametric analysis of data:

- The information derived from phase 1 will provide a solid platform for specification and validation of one or more of the statistical models described in section 3.2. The approaches demonstrated in sections 3.3, 3.4, 3.5, and 3.6 are particularly useful for the Danish dairy herd contexts. An iterative supplementary use of the methods from phase 1 may be needed for refining model specification and model control in phase 2. Phase 2 requires statistical expertise with access to context-specific knowledge.

As an operationalization of the above, I suggest to the herd veterinarian in cattle herds the concrete stepwise approach outlined below. This approach uses the concepts and tools for management of data for health performance measurement presented in this thesis to develop a systems approach to herd health management in an industrialized dairy herd. Note that health performance measurement is a component of herd health management.

Step 1: Develop process behavior charts like that shown in Figure 1 for the available routine measurements from standard herd management programs. These charts do not require sophisticated software or hard-to-justify assumptions. Use animal-level data directly whenever possible. Do not wait until ideal data are available; there will always be data available that are useful for health performance measurement.

Step 2: Make sure you can answer the following questions concerning the definition of the measurements: For what purpose were data collected? Who collected the data? How, when, and where were data collected? What do values represent? If computed, how were they computed from raw data? Were there changes in formulas over time? Precise knowledge about these topics in the concrete herd will give a very strong and necessary foundation for interpreting the charts. Knowledge about the specific context and the dynamics in the context will increase. Meeting these requirements may be a real challenge for a herd health consultant but also an important learning process.

Step 3: Interpret the patterns in each chart, search for assignable causes of exceptional variation (data points outside limits, level shifts or trends), and attempt to remove such causes. This systematic process will add further to your knowledge about the herd context, including the manager's more or less subjective views. The charts and your use of them will document your reasons for suggesting interventions to the herd manager and, if needed, to the public veterinary authorities. You will also be able to distinguish clearly between process-related and results-related measurements and experience the difference between them through the dialogue with the manager.

Step 4: Search for options to reduce the routine variation when the results of the process are unsatisfactory. Some options will be obvious (e.g., repair technical faults in the milking equipment or ensure hoof trimming). However, because of the usually large number of animals and long-time horizon in dairy production, you will profit from some multivariable or multivariate statistical modeling. A range of traditional statistical models, including state space models, are developed

specifically for this purpose (presented and discussed in section 3.2). Model control of these analyses can also serve as advanced tools to explore causes of exceptional variation. Standard setups are available, and the younger generation of veterinarians has been trained in using simple versions. This process will also add substantially to your context knowledge.

Step 5: Set up targets at the tactical or strategic level. The interventions to reduce the routine variation or simply improve the results by eliminating product out of specifications (e.g., high cell counts) will often require some investments, which are quite easy to estimate. However, the benefits in terms of increased production or decreased disease-associated losses are more complicated to assess. Models to do such analyses are described above. Some are commercially available, and you can get support for interpretation and use. With the knowledge gained during steps 1 to 4, you will be well equipped to provide relevant and comprehensive input to these models. The models provide predictions of the important health performance measures and potential profit due to the interventions you consider (targets). The discussions of the results with the manager will bring you deep into the topics described in Figure 3 (section 3.2), which again will provide knowledge about causes of exceptional variation. The entire process in step 5 will also provide estimates of the economic value of each health performance measurement.

Step 6: Adjust the measurements and the intervention strategy. Steps 1–5 should initiate an iterative process. Some measurements will be dropped, others added, the quality of the measurements assessed, process limits or targets possibly changed, cost–benefit assessed, etc. In essence, you have established a systems approach to dairy herd (health) management like that outlined in section 3.2.

Step 7: Develop a framework to support the health performance measurement process at the practice level. This will be particularly useful for establishing a basis for benchmarking because the context knowledge obtained in steps 1 to 6 will allow identification of the most comparable herds. In the section 3.3, a tool is presented for identifying rater bias in ratings used for health performance measurements that must be corrected prior to benchmarking, or across-herd analyses to, for example, evaluate the effects of various interventions like those discussed above in the case of metritis diagnosis and treatment. The validity and usefulness of across-herd analyses will be greatly improved compared to data from larger data collections from multiple veterinary practices. A homogeneous set of data will also be useful for evaluation of diagnostic tests applied in practice and development of new health performance measures like those demonstrated in the case of lactation curves (section 3.4). The activities in step 7 will almost certainly require expert statistical assistance.

6 Perspectives

It is my experience that most computer systems for dairy herd management in Denmark and elsewhere merely produce collections of recordings without convincing attempts to support an evaluation of the herd's performance. Consequently, more consistent use of the methods described in this thesis should provide options for improving efficiency. There are only few studies that attempt to evaluate the added value of improved data recording, data management, statistical analysis, and qualitative follow-up. This thesis demonstrates the inherent problems associated with evaluation studies. Even if we cannot provide convincing evidence of major financial effects of improved management of data for health performance measurement, the improved knowledge about the production system gained from an effect-focused practice may make the production system more robust or adaptive to the future challenges.

In the discussion section, a series of topics for further development were raised. I suggest that the following are of major interest in the near future:

- More work is needed to further develop tools linking signals from process behavior charts or similar with individual animals in the underlying process. Such investigation will require formal comparisons of the transparency of available tools for health performance measurement.
- Initial screening of process behaviour charts should be automated. Then, interpretation of signals can be restricted to a limited number of charts with clear signals.
- Hierarchies of indicators defined by relatively few major problem complexes (e.g., ketosis or poor fertility) should be further developed and evaluated to facilitate the diagnostic process and reduce the inherent problem concerning multiple 'testing' (increased risk of false positives). As of now, I assume it only is a problem for phase 2 analyses.
- The infrastructure for organization of automation of the diagnostic screening (phase 1) needs development. The practice level will be a suitable organizational level.
- Veterinary practice needs some sort of research and development sector that can assist in day-to-day practical implementation of research methods and findings. Such an effort also would involve continuing education of practicing veterinarians because a basic understanding of the principles behind both the statistical and qualitative analyses is needed to establish efficient communication about performance measurement between herd manager and veterinarian.
- Cross-disciplinary research on integrating traditional (herd health) management sciences with social science methods should be established.

References

Andersen, HJ and C Enevoldsen. 2004. Towards a better understanding of the farmer's management practices – the power of combining qualitative and quantitative data. Manuscript in Ph.D.-thesis: Andersen HJ. (2004) Rådgivning – Bevægelse mellem data og dialog. ISBN: 87-89795-81-4

Danish Cattle Federation. 2012. Kvægbrug i verdensklasse. [in Danish] p. 5. Available from: http://www.vfl.dk/NR/rdonlyres/AB2D78FA-5338-4F53-BF1C-AB583AE33C95/0/Kvaegbrugiverdensklasse2012.pdf [Accessed 05/04/2012].

Danish Veterinary and Food Administration. 2011. Guidelines on herd health contracts for cattle herds. (In Danish: Vejledning om sundhedsrådgivningsaftaler for kvægbesætninger).
Available from:
http://www.foedevarestyrelsen.dk/SiteCollectionDocuments/25_PDF_word_filer%20til%20download/05kontor/SRA%20vejledninger%202011/Vejl%20SRA_Kvaeg_290911.pdf
[Accessed 26/03/2012].

De Vries, A, JK Reneau. 2010. Application of statistical process control charts to monitor changes in animal production systems. *J Anim Sci.* 88:11-24. doi: 10.2527/jas.2009-2622

Ellis-Iversen J, AJ Cook, E Watson, M Nielen, L Larkin, M Wooldridge, H Hogeveen. 2010. Perceptions, circumstances and motivators that influence implementation of zoonotic control programs on cattle farms. Prev Vet Med. 93(4):276-85. Erratum in: Prev Vet Med. 2010 94(3-4):318.

Enevoldsen C. 1993. Sundhedsstyring i mælkeproduktionen. PhD-thesis [in Danish]. The Royal Veterinary and Agricultural University, Copenhagen, Denmark

Enevoldsen C. 1997a. Epidemiological considerations related to within herd multivariable modelling in herd health management. International Symposia on Veterinary Epidemiology and Economics (ISVEE) proceedings, ISVEE 8, Paris France.
Available from: http://www.sciquest.org.nz/node/62258

Enevoldsen, C. 1997b. Det israelske rådgivningskoncept – og en dansk oversættelse. 10 pp [in Danish]. In: Proceedings Danske Kvægfagdyrlægers Årsmøde (Danish Bovine Practitioner Seminar),Hindsgavl Slot, Middelfart, Denmark.

Graber, HU, MG Casey, J Naskova, A Steiner, and W Schaeren. 2007. Development of a highly sensitive and specific assay to detect Staphylococcus aureus in bovine mastitic milk. J Dairy Sci. 90:4661–4669. doi:10.3168/jds.2006-902

Hill, A. 2005. Analysis of data from the New Herd Health Program Pilot Project -An evaluation of the effects of the pilot project on productivity, medicine consumption, herd health, and indicators of animal welfare. Available from: www.svkv.dk/sfk/sfk.nsf/.../Slutevaluering%20Pilotprojektet.doc
[Accessed 26/03/2012].

Jensen, NP. 1997. En dansk oversættelse af det Israelske rådgivningssystem. [In Danish]. 6 pp. In: Proceedings Danske Kvægfagdyrlægers Årsmøde (Danish Bovine Practitioner Seminar),Hindsgavl Slot, Middelfart, Denmark.

Kristensen, E and C Enevoldsen. 2008. A mixed methods inquiry: How dairy farmers perceive the value(s) of their involvement in an intensive dairy herd health management program. Acta Vet Scand. 50:50.

Kristensen, E, S Østergaard, MA Krogh, and C Enevoldsen. 2008. Technical indicators of financial performance in the dairy herd. J Dairy Sci 91: 620-631.

Kristensen, E and EB Jakobsen, 2010. (E-)valuation of dairy herd health management. Pp 53-63 in Proceedings World Buiatrics Congress XXVI, Santiago, Chile

Krogh, MA and C Enevoldsen. 2012. A framework for integration of benchmarking and within-herd analysis in dairy herd management – analysis of lactation curves as a case. Manuscript in Ph.D. thesis: Management of data for health performance measurement in the dairy herd, University of Copenhagen, Faculty of Health and Medical Sciences. pp. 55-73

Krogstrup, HK. 2011. Kampen om evidens. Resultatmåling, effektevaluering og evidens. Hans Reitzel Forlag, Copenhagen. 170 pp. ISBN :978-87-412-5516-3

Lastein, DB, M Vaarst, and C Enevoldsen. 2009. Veterinary decision making in relation to metritis - a qualitative approach to understand the background for variation and bias in veterinary medical records. Acta Vet Scand 51:36. doi:10.1186/1751-0147-51-36

McCrown, RL. 2002. Changing systems for supporting farmers' decisions: problems, paradigms and prospects. Agricultural Systems. 74:179-220

Nir-Markusfeld, O. 2003. What are production diseases, and how do we manage them? Acta vet. Scand., suppl. 98, 21-32.

Ruegg, PL. 2010. The application of evidence based medicine to mastitis therapy. World Buiatrics Congress, Santiago Chile. November 14-18, 2010.

SAS Institute Inc. 2008. SAS/IntrNet 9.2: Application Dispatcher. Cary, NC: SAS Institute Inc.

Schwabe, CW, H Riemann and CE Franti. 1977. Epidemiology in veterinary practice. Lea & Febiger, Philadelphia. Pp. 303. ISBN: 0-8121-0573-7

Stärk, KDC and MD Salman. 2001. Relationships between animal health monitoring and the risk assessment process. Acta Vet Scand, 42(Suppl 1):71-77. doi:10.1186/1751-0147-42-S1-S71

Studer, E, W Schaeren, J Naskova, H Pfaeffli, T Kaufmann, M Kirchhofer, A Steiner, and H U Graber. 2008. A longitudinal field study to evaluate the diagnostic properties of a quantitative real-time polymerase chain reaction–based assay to detect Staphylococcus aureus in milk. J. Dairy Sci. 91:1893–1902. doi:10.3168/jds.2007-0485

Sørensen, JT. 1990. Validation of livestock herd simulation models: a review. Livestock Prod Sci. 26(2):79-90

Vaarst, M, B Paarup-Laursen, H Houe, C Fossing, and HJ Andersen. 2002. Farmers choice of medical treatment of mastitis in Danish dairy herds based on qualitative research interviews. J Dairy Sci. 85:992-1001

VFL. 2012. Kvægnyt –Tema:Malkerobotten, no 6. [In Danish]
Available from http://www.vfl.dk/NR/rdonlyres/C5467A1A-89FF-4BCA-988A-7462588094B8/0/KvaegNYT0612.pdf [Accessed 26/03/2012].

Woodall, WH. 2000. Controversies and contradictions in statistical process control. Journal of Quality Technology, 32(4):341-350.

Appendix: Terminology related to health performance measurement

Terms like monitoring, surveillance, control, benchmarking, epidemiological or business intelligence, performance measurement or management, evaluation, evidence, statistical process control, and quality control are widely used. However, the definitions and distinctions between them seem to differ among disciplines, the objectives for application are often vague, and the interpretation can be complicated. This appendix summarizes most of the terms used or discussed in this thesis. In cases where there is doubt about the translation to Danish, I suggest a Danish translation marked with 'Da' in parenthesis. The list below is organized according to increased complexity or abstraction level.

Raw data (Da: *Rådata*)
The basic recordings (Da: ***Registreringer***) or variables available to a user. They may be the direct output from some measuring device (e.g., body weight from a scale). They may also be aggregations of several measurements (e.g., energy corrected milk without access to the underlying information, kg milk, fat percentage, and protein percentage), where the quality has not been evaluated (see about evaluation below). The data can be structured or unstructured:

- Structured data: The values are represented as numbers (interval (continuous or discrete) or ordinal scales, or counts) or concrete nominal categories (unordered categories, including dichotomous)
- Unstructured data: Data can for example be text, pictures, or sound recordings.

To completely and fully describe the data, the user needs to know: Who collected the data? How, when, and where were data collected? What do values represent? If computed, how were they computed from raw data? Were there changes in formulas over time? I will add that sometimes it is crucial to know for what purpose the data are collected to understand why data can be misleading.

Indicator (Da: *Indikator*)
An indicator can be a variable that measures (Da: ***Måler***) the state of a trait (condition) of interest (e.g., mastitis or milk production capacity); similar to a diagnostic test. The diagnostic value of the test is evaluated (see below).

Measurement or measurements (Da: *Måling eller målinger*)
Measurements can be either raw data or indicators; not to be confused with 'a measure to handle something' like 'a mastitis control measure' (Da: ***Foranstaltning***).

The Process Behavior Chart (Da: *Procesovervågning*)
Wheeler (2000) uses the term 'method of continual improvement' to describe the graphical Process Behavior Charts (PBC) and its intended uses. A PBC is a non-parametric (no assumptions of normality or independence over time need to be made) time series graph showing the measurements of interest. So-called Natural Process Limits (NPL) are added to separate the routine variation of the

process (the natural process) from the exceptional variation. With all points inside the limits, the process will also be predictable (within limits). The NPL are estimated from moving ranges (mR), which directly measure the cow-to-cow variation. The use of PBC is integrated with a more or less qualitative follow-up to reveal causes and remove effects of exceptional variation. In case there are no signs of exceptional variation or trends, intervention based on single ('extreme') data points is not warranted. In fact, such intervention may distort the process.

Statistical Process Control (Da: *Statistisk proceskontrol*)
The concept is similar to PBC except for Statistical Process Control (SPC) being based on assumptions about various statistical distributions (parametric). Some SPC also account for autocorrelation and incorporate filtering or smoothing to better reveal patterns in the data. The filtering is achieved by calculating one of several types of moving averages. The typical terminology of SPC differs from PBC. NPL are called 'control limits'. Routine variation is called 'in control'. SPC has been a core tool for quality control in various manufacturing industries. Quality control (Da: ***Kvalitetskontrol***) probably aims most at maintaining the products within specifications, which are more or less arbitrary criteria for quality of the outcome (e.g. sales).

Phase 1 – numerical exploration of context (Da: *Fase 1 numerisk undersøgelse af kontekst*)
The PBC is well suited for the initial analysis of the structured (numerical) data produced in the complex and highly variable dairy herd context. For that context, we require a flexible, adaptive and comprehensive analysis with very few restrictions including a minimum of assumptions about statistical distributions. Sufficiently sensitive intervention criteria (that define signals) can be specified although the risk of false positive signals may be considerable.

Phase 1 – qualitative exploration of context (Da: *Fase 1 kvalitativ undersøgelse af kontekst*)
It is essential that a signal from the PBC is followed immediately by and integrated with a qualitative search for and an evaluation of possible causes. Sufficient context-knowledge and the transparency of the PBC probably limit the number of possible causes and reduce false positive signals to a minimum. The sufficient cause(s) may be identified. If not, at least an approach for further causal analysis in a phase 2 should be specified.

Phase 2 – statistical herd-context modeling (Da: *Fase 2 statistisk modellering af besætningskontekst)*
An initial purely explorative time series analysis (phase 1) may provide justifications for assuming homogeneous processes or certain distributions (e.g., normal or binomial) that permit application of a series of parametric analytical techniques that may be used for prediction and quantification (phase 2).

State Space Models (Da: *Multiproces modeller*)
Statistical analyses that provide predictions of a series of possible future developments of various states (e.g. bulk milk somatic cell counts). A Bayesian approach is often applied, which means that a priori knowledge can be combined with new information in a systematic fashion. A state space

model is in principle any model that includes an observation process and a state process. Kalman-filter models are a subset of the state space models.

Multivariate Statistical Process Control (Da: *Multivariate SPC*)
By 'multivariate analysis', we mean that several variables are analyzed jointly by creating a new Y-variable (response variable) that is defined by the correlations between the original variables. The new indicator may represent an unobservable (latent) condition that has an interpretation or simply a hidden data structure. The calculations are usually based on so-called principal components. Unstructured data (text etc.) can also be classified by means of similar techniques (called text mining).

Diagnostic Test Evaluation (Da: *Evaluering af diagnostiske test*)
Diagnostic tests (including performance measurements) will be used for decision support. Consequently, it is necessary to evaluate the quality of the tests in terms of sensitivity and specificity. Virtually all diagnostic tests are imperfect. A method called latent class analysis (LCA) is suitable for evaluating imperfect tests.

Performance measurement (Da: *Præstationsmåling*)
Measuring (business) performance by means of indicators; using 'Key performance indicators' (Da: ***Nøgletal***) to evaluate performance is a widely used component of business management. Monitoring, surveillance, benchmarking, epidemiological or business intelligence, and control are used with the same meaning in various disciplines (definitions below). Inspired by Krogstrup (2011), I chose to subdivide performance measurement into process measurement and results measurement:

- Process measurement (Da: ***Procesmåling***): Related to the quality of activities in the system (what goes on in terms of, e.g., types of management routines (actions) like heat detection). The activities (intervention) will be based on some more or less explicit formulation of needs and objectives. The measurements are primarily focused on the means (the capacity; Da: ***Midler***), not the final results (products). The quality of human activities can also be called competencies or qualifications. The process measurements provide knowledge about the output of the intervention/the processes (in terms of what was actually done in the process-routines; e.g., minutes of heat detection every day).
- Results measurement (Da: ***Resultatmåling***): Related to the end results of the transformations in the system. That is, what is of interest to the end-user (the costumer, the recipient), which is called the outcome of the performance of the process (e.g., pregnancy rate or the quality of milk deliveries). Because the outcome may be affected by interaction between intervention and context, it is useful to estimate the subset of the outcome that is directly related to (caused by) the intervention: Intervention effect (Da: ***Kausalvirkning***). Estimation of effects (causal and other) requires evaluation.

Monitoring (Da: *Monitorering eller monitering*)
In some areas of veterinary medicine, monitoring may be defined as merely some more or less systematic collection of raw data without specific objectives. In public management monitoring is characterized as a continuous and systematic function that by means of indicators provides knowledge about the development in the organization (e.g. progression, meeting goals, and responsibility in the use of resources); this process is also defined as auditing (Krogstrup, 2011).

Surveillance (Da: *Overvågning*)
In veterinary public health, surveillance is defined as some active objective-oriented activity based on indicators in contrast to monitoring, which is seen merely as some more or less systematic collection of raw data. Due to the diverse uses in different disciplines, it is important to specify the objectives and the methods exactly in case of collaboration across disciplines.

Control (Da: *Kontrol*)
Control is an evaluation of the importance of a deviation between an obtained result and a target. In broader terms, 'to control' means to keep performance within certain limits. In veterinary public health actions of what should occur if limits are crossed are predetermined which is in contrast to surveillance.

Intervention (Da: *Korrigerende handling*)
An intervention is the actions initiated to achieve a specific outcome. For an intervention to be practically applicable, we need to know how and when it works. This is known as an intervention theory (Da: ***Interventionsteori***) and is at least a socio-biological hypothesis of why the intervention should be working.

Performance management (Da: *Præstationsledelse*)
Herd health management (Da: ***Sundhedsstyring***) is a specific aspect of (business) performance management. Performance measurement is one of several tools for management. Other examples of tools for dairy herd management are optimization of feeding and formulation of operating procedures. The term 'business intelligence' can be seen as similar to performance management.

Benchmark (Da: *Sammenligningsgrundlag*) Benchmarking (Da: *Foretage sammenligning*)
Comparison of indicators' values in comparative herds is one obvious way to select targets. It will indicate performance level at best practice. The selected target performance measures can also be considered a prognosis for the future or a budget. If the scale of a measurement differs, benchmarking becomes invalid. 'Clinical recordings' obviously must be standardized to be useful for benchmarking. Clinical criteria that are constant within herd (e.g., specific for a single manager or veterinarian) may suffice if performance measurement is restricted to historical data within the herd. A simple approach to setting herd-specific targets is to take historical results and adjust them for expected results of the planned changes in the next planning period.

Evaluation (Da: *Evaluering*)
Krogstrup (2011) gives a broad definition of evaluation for public management: "A systematic retrospective assessment of output (process), outcome (results), administration, and organization of (public) business, which is expected to play a role for practical actions". In this definition, it is essential to note that evaluation includes some judgment that separates important aspects from unimportant aspects. It is also essential that practical use is intended.

Effect evaluation (Da: Effektevaluering; kausal virkning)
Summarized from Krogstrup (2011), a broad definition for public management is: To provide knowledge about causal effects of interventions and the mechanisms of the intervention-effect relations.

Evidence (Da: *Evidens*)
The outcome of an effect evaluation. In a hierarchy of evidence suggested by Krogstrup (2011), the randomized controlled trial ranks at the top, action research around the middle, and user-assessments at the bottom.

A tame problem (Da: *Simpelt problem*)
A tame problem is a simple problem concerning a well-understood context with concrete solutions.

A wild problem (Da: *Komplekst problem*)
A tame problem is in contrast to identification or quantification of causal effects (evaluation) in a context like Figure 3 (section 3.2). Krogstrup (2011) calls a problem similar to that in Figure 3 a wild problem, which mainly is characterized by a vague definition, lack of an optimal solution, unclear causal mechanisms, and interaction between context and mechanisms

An effect-focused practice (Da: *Effect-fokuseret praksis*).
A systematic use of the simple PBC in a herd (which includes more or less qualitative follow-up to remove effects of exceptional variation) could be seen as an example of an effect-focused practice.

Herd Health (Da: *Besætningssundhed*)
Herd health can be defined as (section 3.2), "Animal, environment, and manager together viewed as a dynamic and complex ecosystem. In this context, an ecologically informed or process-view of herd health implies the self-regulation through feedback and maintenance of all relevant support systems promoting ongoing physical, mental/emotional, and social well-being. This latter definition gives us a sharper understanding of what poor herd health is. That is, the loss of the ability to self-regulate and the disintegration of support systems leading to the necessity for intervention. In a process-view, intervention is directed towards restoration of all relevant support systems in order for health again to be self-generated and self-regulated".

References

Krogstrup, HK. 2011. Kampen om evidens. Resultatmåling, effektevaluering og evidens. Hans Reitzel Forlag, Copenhagen. 170 pp. ISBN :978-87-412-5516-3

Wheeler, DJ. 2000. Understanding variation. The key to managing chaos. 2nd Edition. SPC Press, Knoxville, Tennessee, USA.