

FACULTY OF HEALTH AND MEDICAL SCIENCES
UNIVERSITY OF COPENHAGEN



Herd-specific Randomized Trials

– an approach for Effect Evaluation
in a Dairy Herd Health Management Program

Ph.D. Thesis · 2012
DORTE BAY LASTEIN

Herd-specific Randomized Trials

– an approach for Effect Evaluation in
a Dairy Herd Health Management Program

Ph.D. Thesis · 2012

Dorte Bay Lastein

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

Ph.D. Thesis

Author:

Dorte Bay Lastein, DVM
Department of Large Animal Sciences
Faculty of Health and Medical Sciences
University of Copenhagen
Grønnegårdsvej 2
1780 Frederiksberg C
Denmark

Future address:

Nr. Vittrupvej 16
6650 Brørup
Denmark
dorte.bay@gmail.com

Principal supervisor:**Carsten Enevoldsen, Professor**

Department of Large Animal Sciences
Faculty of Health and Medical Sciences
University of Copenhagen
Grønnegårdsvej 2
1780 Frederiksberg C
Denmark

Assessment committee:**Chairman****Ilka C. Klaas, Associate Professor**

Department of Large Animal Sciences
Faculty of Health and Medical Sciences
University of Copenhagen
Grønnegårdsvej 2, DK-1780 Frederiksberg C
Denmark

Egon Noe, Senior Research Scientist

Department of Agroecology - Farming Systems
University of Aarhus
Blichers Allé 20, DK-8830 Tjele
Denmark

Mirjam Nielen, Prof. Dr.

Professor in Evidence- based Veterinary Medicine
Departement Gezondheidszorg Landbouwhuisdieren
Yalelann 7, 3584CL, Utrecht
Netherland

PhD Thesis 2012 © Dorte Bay Lastein

ISBN 978-87-7611-538-8

Printed by SL grafik, Frederiksberg C, Denmark (www.slgrafik.dk)

Preface

This Ph.D. thesis has been carried out to fulfil the requirements for my Ph.D.-education at University of Copenhagen, Faculty of Health and Medical Sciences. The work was initiated in May 2007 and completed in September 2012, interrupted by both practical veterinary work and two maternity leaves.

The process of my Ph.D. education has been like opening a dozen of doors. And not yet I have been forced to close any behind me. I said to myself when I applied for the scholarship: "It is just like opening a door and have a little look on the other side. I can always get back to the world outside." I ended up entering the door of the Ph.D. education, and behind it opened even more doors. Some I have had a quick glance behind. The rooms behind other doors have been examined more thoroughly. One thing I realize now – I cannot go back out the Ph.D. door as the same person I went in. In the future, I can never reclaim total ignorance of what the Ph.D. study has taught me.

I am thankful to have started and finalized my Ph.D. study although I sometimes felt like I was a fish that couldn't swim. I thank my supervisor, Carsten Enevoldsen for letting me float freely around the 'pond of science' and for throwing the fish line out to guide me towards the safer shores. I thank Mette Vaarst, Foulum for all your enthusiasm and inspiration. Thanks to Erling Kristensen and Mogens Krogh for pointing at the Ph.D.-door in the first place and for all the discussions and help along the way. Thanks to Steen Larsen, Veterinary School of Oslo for your enthusiasm and silent inspiration (though you might not know how often I have thought of you). Thanks to Sigge Falkenberg for giving me insight into the world of communication and abstraction. A special thanks to the participating veterinarians and farmers for following my ideas, conducting the trials and waiting patiently on the results.

It took me many years to finalize my studies. Nils, you were there all the time. Andreas and Bertram entered the family on the way. Thankful for their coming though they forced me to change my plans and priorities several time. But as a supervisor said: 'there are important things such as your Ph.D., and there are very important things such as your family and life. Life – and research is all about priorities. I am ready to enter the door to 'outside world' again.

Ved Vejs Ende, Vittrup 2012 - Dorte Bay Lastein

Summary

This Ph.D. thesis is centred on veterinarian and dairy producer interactions in counselling situations in dairy herds. It explores the possibilities of developing knowledge (or evidence) through systematic evaluation principles, specifically randomized controlled trials, which are customized to specific herds and herd problems. The project is based on a case: evaluation of effectiveness of medical treatments for genital disease in a Danish herd health management program. The project shows that the consulting veterinarian's 'tool box' can be extended with practical experimental designs that accommodate both human perceptions and epidemiological methods. Thus, the proposed trial approach can provide specific local evidence that in combination with general evidence, experience, and personal preferences can be implemented under the concept of 'evidence-based veterinary practice'.

First, I introduce the reader to the overall project context, as well as giving a description of the project's specific objectives and the studies performed (Chapter 1). Then, I elaborate briefly on the project approach, the procedures for collecting qualitative and quantitative data, and the analytical methods (Chapter 2). Subsequently, I present the main results of the thesis based on five manuscripts. The manuscripts are included immediately afterwards, and they provide a detailed description of data and methodology (Chapter 3). The five manuscripts are also described briefly below. In the final discussion, conclusion, and perspectives, I combine the results of the entire project to assess future possibilities for implementation of clinical field trials for herd health management.

The purpose of the manuscript '**Evidence-based Veterinary practice for a dairy herd health management context – a tutorial**' is to introduce readers to the theoretical and abstract concepts within the scientific disciplines of trials. The concept of 'hierarchy of evidence' is central. I develop a synthesis of proposals for the implementation of systematic evaluation by means of clinical field trials.

The purpose of the manuscript '**Review of effectiveness of medical treatment for early postpartum bovine genital disease based on vaginal discharge**' is to demonstrate an example of how difficult it can be to find knowledge (evidence) that can be used directly as decision support for solving clinical and management-related herd problems. The reviewed literature is evaluated in relation to the hierarchy of evidence and its applicability for the Danish dairy herd health management context.

The purpose of the manuscript '**Diagnostic procedures and medical treatments for bovine genital disease in Denmark – a qualitative analysis of the potential for implementation of clinical field trials in herd health management**' is to exemplify the action patterns related to diagnosis and medical treatment as

Danish veterinarians apply them in their daily work. In the given context, the range of applied diagnosis and treatment of genital disease based on systematic clinical examinations is described. The results of the analysis are used to link the patterns of action to potential trial designs so that maximum adaptation to the individual herd problem and herd context can be achieved.

The purpose of the manuscript '**Veterinary decision making in relation to metritis – a qualitative approach to understand the background for variation and bias in veterinary medical records**' [published in *Acta Veterinaria Scandinavica* (2009), 51:36] is to exemplify the decision-making processes that take place in connection with the diagnosis and treatment of clinical conditions. The study illustrates that the veterinarian focus on clinical issues, counselling, law, or epidemiological issues affects motivation for systematic data collection.

The purpose of the manuscript '**Clinical field trials in a dairy herd health management program: treatment effectiveness on milk production in case of early postpartum vaginal discharge**' is to demonstrate and discuss the experience of the practical development and implementation of herd-specific randomized clinical field trials within one veterinary practice for dairy herd health management purposes. The study shows an example of how the data from randomized clinical field trials can be analysed and used to evaluate differences in treatment effect between two active treatments and the effect of a disease despite treatment.

Sammendrag (Danish summary)

Denne afhandling tager udgangspunkt i dyrlæger og mælkeproducenters samspil i rådgivningssituationen i malkekobesætninger. Den omhandler mulighederne for at udvikle viden (eller evidens) gennem systematisk evaluering, nærmere bestemt randomiserede kontrollerede forsøg, der er tilpassede til at skabe viden om specifikke besætningsproblemer. Projektet tager udgangspunkt i en evaluering af behandlingseffektivitet af medicinske børbehandlinger i et dansk rådgivningskoncept. Projektet viser, at den rådgivende dyrlæges 'værktøjs-kasse' kan udvides med praktisk anvendelige forsøgsdesign, der tilgodeser både menneskelige opfattelser og epidemiologisk metode. Således kan besætnings-specifik evidens kombineres med general evidens, erfaring og personlige præferencer under begrebet 'evidensbaseret dyrlægepraksis'.

Indledningsvis introducerer jeg læseren til projektets overordnede kontekst, samt en beskrivelse af projektets specifikke mål og de udførte studier (kapitel 1). Dernæst uddyber jeg kort projektets form, idet procedurer for indsamling af både kvalitative og kvantitative data samt analytiske metoder introduceres (kapitel 2). Efterfølgende fremlægger jeg hovedresultaterne for afhandlingens 5 manuskripter. Manuskripterne er inkluderet umiddelbart derefter, og de giver en detaljeret beskrivelse af studierne (kapitel 3). De fem manuskripter gennemgås desuden kortfattet nedenfor. I den afsluttende diskussion, konklusion og perspektivering samler jeg op på projektets resultater som helhed for at vurdere de fremtidige muligheder for implementering af kliniske besætningsbaserede forsøg i rådgivningen i malkekobesætninger.

Formålet med manuscript I '**Evidence-based veterinary practice for a dairy herd health management context – a tutorial**' er at introducere læserne til teoretiske og abstrakte begreber indenfor de videnskabelige discipliner, der ligger bag forsøg. Begrebet 'evidens-hierarki' er centralt. Der opstilles en syntese af projektet med forslag til udvikling af systematiske besætnings-specifikke effektevalueringer.

Formålet med manuscript II '**Review on effectiveness of medical treatment for early-postpartum bovine genital disease based on vaginal discharge**' er at demonstrere et eksempel på, hvor vanskeligt det kan være at finde viden (evidens), der kan anvendes direkte som beslutningsstøtte til løsning af kliniske og managementrelaterede besætningsproblemer. Den beskrevne litteratur vurderes i forhold til et evidenshierarki og dens anvendelighed i den danske sammenhæng.

Formålet med manuscript III '**Diagnostic procedures and medical treatments for bovine genital disease in Denmark - a qualitative analysis of the potential for implementing herd-specific randomized trials in a**

herd health management program' er at eksemplificere de handlemønstre indenfor diagnostik og medicinsk behandling, som danske dyrlæger foretager inden for i deres daglige gang i malkekobesætninger. I den givne kontekst vurderes diagnostik og behandling af børtilidelser ved systematiske kliniske undersøgelser. Resultaterne af undersøgelsen bruges til at koble handlemønstre og potentielle forsøgsdesign, således at størst mulig tilpasning til det enkelte besætningsproblem vil kunne opnås.

Formålet med manuskript IV '**Veterinary decision making in relation to metritis - a qualitative approach to understand the background for variation and bias in veterinary medical records'** (publiseret i Acta Veterinaria Scandinavica (2009), 51:36) er at eksemplificere de beslutningsprocesser, der foregår i forbindelse med diagnostik og behandling af kliniske lidelser. Undersøgelsen illustrerer, at dyrlægernes fokuspunkt på enten klinik, rådgivning, lovregler eller epidemiologi påvirker deres motivation for systematisk dataopsamling.

Formålet med manuskript V '**Clinical field trials in a dairy herd health management program: treatment effectiveness on milk production in case of early postpartum vaginal discharge'** er at demonstrere og diskutere erfaringer med den praktiske implementering af kliniske besætningsforsøg i et rådgivningsprogram. Der gives eksempel på, hvorledes data fra sådanne forsøg kan analyseres og anvendes til vurdering af forskelle i behandlingseffekt mellem aktive behandlinger og til vurdering af sygdomseffekt på trods af behandling.

Contents

PREFACE	I
SUMMARY	III
SAMMENDRAG (DANISH SUMMARY).....	V
CONTENTS	1
1 GENERAL INTRODUCTION.....	3
2 STUDY CONTEXT.....	9
3 RESULTS.....	17
3.1 Summary of results.....	17
3.2 Evidence-based veterinary practice in a dairy herd health management context – a tutorial (Manuscript I)	23
3.3 Review of effectiveness of medical treatment for early-postpartum bovine genital disease based on vaginal discharge (Manuscript II)	77
3.4 Diagnostic procedures and medical treatments for bovine genital disease in Denmark - a qualitative analysis of the potential for implementing herd-specific randomized trials in a herd health management program (Manuscript III).....	109
3.5 Veterinary decision making in relation to metritis - a qualitative approach to understand the background for variation and bias in veterinary medical records (Manuscript IV)	139
3.6 Clinical field trials in a dairy herd health management program: treatment effectiveness on milk production in case of early postpartum vaginal discharge (Manuscript V).....	151
4 DISCUSSION	181
5 CONCLUSIONS.....	189
6 PERSPECTIVES	193
7 REFERENCES – THESIS	195

8 APPENDIX	197
Appendix A	198
Appendix B	199
Appendix C	200

1 General introduction

Motivation

My Ph.D. project was motivated by my speculations while working in a Danish veterinary cattle practice from 2003 to 2007. I asked myself whether the medical treatments I allocated to dairy cows were beneficial for the animals themselves, the farmer, the dairy production, or society as a whole. Digging into the literature on **veterinary scientific evidence** of disease definitions and treatment effects, I often was unable to find general and convincing answers. With increasing insight into the epidemiological methods gained during my studies, especially concerning **causality and strength of scientific evidence**, I realized that no easy answers were readily available.

I was introduced to the general ideas of using cyclic **randomized field trials** for herd management procedures as proposed by Schwabe et al. (1977) [1] and elaborated on by the work of Enevoldsen (1993) [2]. Inspired by the ideas of the principles of **'local truth' in the herd context** proposed by Nir Markusfeld [3] and the increasing focus on **evidence-based veterinary medicine** [4,5], I saw randomized controlled trials within herds or within practices as a potential science-based improvement of my own (and potentially other veterinarians') pragmatic trade-offs in practice. Could decisions about treatment threshold and treatment protocols in the future be based on scientific criteria and estimates of effect of higher scientific validity as opposed to individual assumptions and perceptions based on undocumented experience?

However, at present, veterinary trials are primarily conducted by researchers and not by veterinarians in practice, so my challenges intensified. Can academic theories on trials be implemented in the often somewhat chaotic world of veterinary practice, in the tension-field among cows, disease, human decisions, and data recorded in the barn among the cows? I found that I had to include more humanistic aspects in my research. I also found that the aspects of **ontology (life-worlds) and individuality in human decision making and action patterns** [6-8] were a necessity for my study. **Cross-disciplinary research, mixed methods, grounded theory, and phenomenographic principles** [9-11] suddenly became highly relevant to study as these topics had not been part of my veterinary curriculum. The inclusion of such humanistic scientific methodologies again brought into question the validity of the purely positivistic epidemiological methods to identify evidence of effect [12]. Because of the time span of the study, the **legislation on an extended Herd Health Management Program (HHMP)** [13] changed from voluntary to compulsory data collection. These changes could potentially affect the **quality of data** on disease and medical treatments in the national Danish cattle database and influence the potentials of field trials in practice. The increasing legal requirements for recording, documentation, and effect evaluation in the dairy context led me to a

parallelization towards '**New Public Management**' (NPM) strategies and related consequences [12]. The implication of such NPM strategies in the HHMP context became an eye opener. Leblanc et al. [14] suggested turning dairy herd health management or production medicine into an integrated, holistic, proactive, data-based, and economically framed approach to prevention and enhancement of performance. As a specific component of this concept, I have explored the **conceptual idea of integrating randomized controlled trials (RCTs) into HHMPs**.

Context and concepts

Internationally and nationally, veterinary work has shifted from individual cow diagnosis and treatments toward herd medicine over the last decades [14]:

- from 'call on demand visits' producing treatment recordings primarily based on the farmer's definition of disease (in Denmark before 2006) and all medical treatment in adult cows performed by a veterinarian
- to **voluntary** advisory functions and systematic clinical examination and scoring of **all** cows at risk of predefined disease entities producing a continuum of scores and treatment data based on scoring charts and primarily the veterinarian's definition of disease (2006–2010) and the initial medical treatment of adult cows performed by a veterinarian
- towards **compulsory** advisory functions, audit functions, and systematic clinical examination and scoring of **selected** cows at risk of **selected** disease producing discontinuous recordings of scores and treatment data based on scoring charts and the veterinarian's definition of disease. Also, a liberalization of the antibiotic use towards farmer-initiated treatments has followed this legislation.

As a consequence of these changes, the veterinarians working with advisory functions in Danish HHMPs have also become increasingly dependent on analysis of quantitative measurements of production in the dairy herd to monitor and evaluate the development (health performance measurement). Such requirements for quantitative analysis are a demand from the veterinary public authorities focusing on welfare and antibiotic resistance, from the milk factory focusing on low somatic cell counts and no medical residues, and from the dairy industry itself.

In addition to the relatively systematic collection of data in the dairy herd context in the year 2012 (e.g., scores representing disease occurrence), all medical treatments, registrations of reproduction parameters, and milk production are recorded in the national cattle data base. However, the validity of the recording in the database most likely will change with the adjustments in legislation described above because farmers

resume responsibility for recording and the incentives for recording change. Despite such concerns, the data in the national cattle data base form a solid and useful basis for performance management.

Under Danish circumstances, methods and tools for herd health performance measurement (for instance, vaginal discharge scores and lactation curves) are developed and implemented in clinical practice via a database, the VPR platform [15]. For a detailed description of the historical development of the Danish data recording and herd-specific analysis for herd health management purposes, I refer to the work of Krogh (2012). However, the already established principles primarily work as monitoring systems with some possibilities of evaluation against defined limits, targets, and benchmarks. As such, the so-far-implemented principles in the Danish system cannot claim to show direct cause–effect relationships. The implementation of methods to establish and quantify cause–effect relationships in the dairy herd context could potentially support decision making in practice. From an epidemiological viewpoint, the strongest evidence for a cause–effect relationship is obtained through RCTs [16].

Implementing RCTs in the dairy context would seem straightforward if it were not for human beings! The circumstances for data collection in a Danish dairy context are complex. The health status of the cows is evaluated, determined, and acted upon by humans: the farm personnel and veterinarians. The resulting data on health status and treatment incidence are thus based on a complex web of the clinical signs of disease and individual human decisions, experience, and perceptions of why, when, and how to intervene. The human aspects that affect the validity of disease and treatment data seem to be somewhat neglected. Attempts to encounter the complexity with both quantitative and qualitative measures have been demonstrated within the context of the Danish dairy industry and HHMPs [7,17], and these mixed methods are adopted and adapted in the present work. The project includes qualitative research to elaborate on human influence on data obtained during diagnostic work and use of medical treatment in general, and in trial situations in particular. The project also includes quantitative research to provide estimates of intervention effects and to show the potential of integrating an advanced effect evaluation as clinical field trials in herd health management.

Ethical aspects of implementation of RCTs to study medical treatment of disease

Ethical concerns regarding animal welfare and use of medical treatments in the milk production industry play increasing roles in society. Considerations of the distinction between healthy and diseased, selective medication, preventive medication, and/or no medication in cases of disease or cows at risk of disease are important for decision making in practice. These topics probably were major reasons for the legislation

outlined above. All humans make their decisions based on an individual ethical or ideological standpoint related to their own ontology, being aware or unaware of it. In the field of trial theory, ethical considerations are important. First and foremost, to ensure animal welfare, trials should follow good veterinary practice and in most cases be registered by the legal authorities. Initially in the work with this thesis, the Danish authorities were contacted to seek advice in relation to the proposed RCT conducted by veterinarians in practice. No legal implications about special registration were needed as long as 'no seriously diseased animals were withheld treatment' and only procedures and treatment that veterinarians in practice would normally and legally use were tested. On this background, trials in the Ph.D. project and future similar trials can be conducted without any registrations and legal implications.

The scientific aim of the thesis

The overall aim of the Ph.D. study can be characterized as the initial phases of a concept development and is described as follows:

To develop, implement, and conceptually validate randomized controlled trial designs that can be used by practicing cattle veterinarians for continuous development and evaluation of current and new diagnostic criteria and medical interventions in a dairy herd health management context. The evaluation of the effectiveness of interventions should primarily be based on routinely collected production-related data.

This project is based on the specific veterinary case related to diagnosing and treating genital diseases (diagnosed as vaginal discharge 0–21 days postpartum) in the context of the HHMP described above. The 6 objectives listed below support the overall aim by together describing and exemplifying the context of the chosen case within the Danish HHMP. In this way, this entire project will be the development of a model or a concept to generate context-specific knowledge in commercial dairy herds. It follows that the methodologies applied are mainly of inductive character and thus hypothesis-creating.

Objective 1: Describe currently applied diagnostic treatment criteria and regimes for metritis treatment based on systematic clinical examination of cows by veterinarians 5-21 days postpartum.

Objective 2: Evaluate the theoretical justification of the current treatment criteria and regimes.

Objective 3: Describe the decision-making process concerning current diagnostic criteria and treatment regimes in the HHMP context.

Objective 4: Develop a practically applicable experimental design to estimate effectiveness of medical treatment of metritis on herd level and validate the conceptual idea of the proposed design.

Objective 5: Demonstrate an implementation of a trial to estimate the effectiveness of changing the regime for medical treatment of metritis on a production-related results measurement.

Objective 6: Identify possible interactions between metritis – and effectiveness of treatment - and relevant prognostic factors like parity, or selected clinical health indicators with respect to effects on production parameters.

The fulfilment of the overall aim and the 6 objectives was achieved by means of the work detailed in five separate manuscripts included in this thesis (note that the manuscripts do not follow the objectives chronologically). The contents of the manuscripts are described very briefly in the following:

Manuscript I (section 3.2) is dedicated to veterinarians in practice and is a synthesis of my suggestions for use of RCTs in the HHMP context. The manuscript is organized as a tutorial that also presents the theoretical background for the specific studies presented in manuscripts II–V and the contents of these studies. The topics are linked together by means of a concrete case: evaluation of the effectiveness of medical treatment of genital disease in dairy cows in the early postpartum period (0–21 days postpartum). The text is organized to promote common understanding for the veterinarian in dairy practice.

Manuscript II (section 3.3) contains a literature review of the effectiveness of medical treatments for early postpartum genital disease. Special emphasis is on the consequences of procedures in the Danish HHMP context.

Manuscript III (section 3.4) describes the currently applied diagnostic treatment criteria and treatment protocols and identifies the links between these observations and relevant trial designs and is based on a qualitative research approach.

Manuscript IV (section 3.5, published in *Acta Veterinaria Scandinavica* 2009, 51:36) contains a description of the decision-making processes in the HHMP related to genital diseases and the consequences for variation and bias in the recorded quantitative data and is based on a qualitative research approach.

Manuscript V (section 3.6) demonstrates the principles for implementation, conduct, and analysis of clinical field trials in 4 herds within one veterinary practice with a focus on treatment effect on milk production. The manuscript is based on quantitative research approach.

Outline of thesis

Chapter 2 describes the study context, including data collection and applied methods.

Chapter 3 contains a very brief summary of results in manuscripts I–V as an introduction to the manuscripts also included in chapter 3.

Chapter 4 contains a general discussion of studies presented in chapter 3 and the possible application of the trial concept in the HHMP.

Chapter 5 provides a summary conclusion of the entire project.

Chapter 6 outlines perspectives for implementation of herd-specific randomized trials in veterinary practice and outlines needs for future research and development.

2 Study context

Project description

This Ph.D. project was developed within the framework of the Research School for Animal Production and Health (RAPH), University of Copenhagen, Denmark, which promotes cross-disciplinary research. In addition to the traditional veterinary focus on the natural sciences, this project also was required to apply qualitative methodology from social science. This aspect made it very attractive to return to the university after working in veterinary dairy practice for 4 years, giving me a greater awareness of the importance of the sociological aspects of veterinary tasks. Understand and appreciating the different methodological methods has contributed much to my personal development and increased my awareness of barriers between science and practice.

The study context

The veterinary profession in the industrialized world in general has undergone a shift in paradigm from 'diagnosis and treatment of individual cows' towards 'herd diagnosis and management of production' [14]. In the late 1990s, a herd health management program (HHMP) involving systematic clinical examinations of all cows in periods of particular interest for disease diagnosis and preventive actions (postpartum, peak milk production/before start insemination, and before drying off) was introduced in Denmark. The concept was inspired by a similar system implemented in Israel in the late 1980s [3]. This HHMP is intended to collect uniformly diagnosed disease data for storage in central database(s) to facilitate subsequent qualitative and quantitative herd-level data analysis on health/disease, production, and reproduction [18]. A set of tools was developed for case studies of individual cows and evaluation of trends in disease occurrence to support decisions based on data obtained in the Danish HHMP [15]. From 2006 onwards, the HHMP was regulated (and therefore also changed) by the authorities as an integrated part of legislation on herd health and medicine.

The Ph.D. project presented here was developed within and around the framework of this HHMP. First, the background for data collection within the program was explored in depth using a qualitative research approach. Second, the trials were developed, implemented by veterinarians in herds working within the program for years, and analysed by a mixed approach (both qualitative and quantitative aspects). Because the main focus of the thesis was to develop the conceptual ideas concerning herd field trials, I outsourced the practical data collection to veterinarians in the field. This approach will reflect the challenges related to data quality and motivation for evaluating effects of medical treatment outside a laboratory or a university

herd. Moreover, this decision was taken to allow me as the Ph.D. student to function as a ‘trial manager’ (Farrell et al., 2010), to allow for inclusion of several participating herds with weekly clinical examinations over a long period of time (e.g., not possible for a single person), and to integrate and implement the project in veterinary practice from the beginning. It also allowed for better reflections on the potential problems related to implementation, such as agreement on disease definitions among veterinarians (data quality) and differences between herds.

Data sources and data collection

Qualitative research

Qualitative research methods (semi-structured interviews and observations) were used to explore the life-worlds of veterinarians in practice in relation to ‘diagnosis and treatment of genital diseases within the HHMP’ and ‘validation of the conceptual idea of a randomized controlled trial in a HHMP’¹ [19].

Between January 2008 and March 2008 I conducted all interviews (½–1¼ hours per interview) and made observations during visits in herds together with the veterinarians for 4–8 hours per visit with each of the 12 veterinarians. Interviews were guided by an interview guide (Table 1). A total of approximately 15 hours of interview recordings were fully transcribed and analysed. The applied analytical principles were inspired by both ‘grounded theory’ [inductive development of a theory from empirical data [20]] and ‘phenomenography’ [aimed at explicitly describing variation in ‘ways of experiencing a phenomena within a group of people’ and to relate these conceptions in a structural hierarchy or ‘outcome space’ [11,21]]. Phenomenographic research is based on the following assumptions according to the references given above:

- the existence of a finite number of ‘ways of experiencing a phenomenon’ within a group of people
- these ‘ways of experiencing a phenomenon’ being structurally related
- a perception of the world (ontology) that a person and the world are interlinked
- clear aim of exploring the range of meanings within a group, not the range of meaning for each individual

The practical method used in the analytical process in manuscript III was grouping of descriptions and perceptions and the use of a trial design tool (PRECIS) [22]. For manuscript IV, I applied ‘meaning

¹ ‘Conceptual validation’ refers to a process of validation of the concepts of the trial approach involving face validation by ‘experts’ (see Sargent, 1982)[19]. The validation process should investigate if theories and assumptions in the concept are ‘reasonable’. In our context, we asked the participating veterinarians and farmers and a number of other veterinarians (the end-user) if they found the concepts ‘reasonable and applicable in practice’.

condensation’ of transcribed interviews, grouping of ‘meanings’ across interviewees, and finally construction of the structural hierarchy or ‘the conceptual model of understanding’ [20]. For further details, I refer to manuscripts III and IV.

Table1. Interview guide

Interview themes
Veterinarian’s treatment of metritis
Clinical registration
Diagnostic criteria
Treatment strategies
Treatment effect in relation to production parameters
Control of clinical effect
Herd status
Farmer’s influence
Influence of strategy in veterinary practice
Ideology
Legislation

Mixed research methodology

Here follow descriptions of the participating veterinarians/herds, the process of developing the diagnosis and treatment protocol (qualitative approach), and the data collection and analysis (quantitative approach) in the HHMP trials.

The trial project was conducted in cooperation with one veterinary cattle practice (practice A) with 7–8 cattle veterinarians in 10 herds in southern Jutland, Denmark. Two of these veterinarians were authorized by the veterinary authorities and the farmer as ‘herd veterinarians’ and performed most of the systematic clinical examinations in the participating herds. A calibration of vaginal discharge scores (VDS) (part of the inclusion criteria; see below and Appendix A) was performed within practice A [23]. The agreement between veterinarians (weighted kappa=0.64 [0.62–0.67]) was considered sufficient (Appendix B) [23] and no further calibration was conducted. Practice A was selected because of my prior knowledge about the proactive attitudes to development of the HHMP and to clinical registrations in general. Additionally, one veterinarian in another practice (practice B) on the island of Fyn conducted an unrelated trial in one herd initiated by his own motivation to ‘try the concept of a randomized controlled trial in the HHMP’. A total of

11 trials were implemented and conducted (Table 2). Introduction, information, and follow-up meetings with veterinarians and farmers were held throughout the study period of one year to discuss practical issues and adherence to the protocol. Only data from four herds in practice A are used to demonstrate the quantitative analytical principles (manuscript V) because of considerations of sample size/power and the relevance of the tested protocols. But as the experiences of designing and implementing all 11 trials are relevant and have contributed to the process of conceptually validating of the trial-approach I have included a few details on all trials, such as overall description and protocols (see below). The results in the remainder of the herds were derived at herd level and presented to each herd owner. These results are not described in this thesis.

Table 2. Description of the herds participating in trials: identification number, affiliated herd veterinarians (1–4), study period, production type (conventional or organic), breed [Holstein/Danish Red (RD)], number of calvings in study period, and trial protocol (A–D).

Herd ID	Vet. practice	Herd Vet	Production type - Breed	No. calvings in trial period	Protocol (all protocols are described in Appendix C)
1	A	1	Conventional Holstein	467	A
2	A	1	Conventional Holstein/DR	196	A
3	A	1	Conventional Holstein/DR	160	A
4	A	2	Conventional Holstein	120	A
5	A	3	Conventional Holstein	288	B
6	A	3	Conventional Holstein	226	B
7	A	3	Conventional DR	150	B
8	A	3	Conventional Holstein	102	B
9	A	3	Organic Holstein	171	C
10	A	3	DROP-OUT*		
11	B	4	Conventional Holstein	111	D

*No description or analysis performed

The 11 private, free-stall dairy herds were purpose-sampled by the participating veterinarians based on their motivation and willingness to participate in long-term trials and not because of explicit problems related to genital diseases postpartum. Veterinarians and farmers were encouraged to be aware of non-

adherence to the protocol and note potential errors on a form available for that purpose in the barn office. I collected these forms twice during the trial period, and errors were coded and saved in files for later evaluation of adherence. All farmers gave their written informed consent to participate and for my access to all data regarding clinical registrations, medical treatments, insemination, pregnancy checks, and milk yield in the national cattle database. No compensation of any form was given.

The study period was predetermined to be approximately one year (herd #11 half a year). Data collection in herd #1–10 was initiated in May 2008 and in January 2009 in herd #11. One herd (#10) was withdrawn before the end of the trial because of problems with acceptance of allocation procedures. Trial designs were determined in all herds in June 2009. The final dataset was extracted from the national cattle database in August 2010, which allowed for a minimum of 1 year of follow-up.

Development of trial design in the practice context

The development of the actual study design and implementation of the protocol in each project herd were achieved in cooperation between the veterinarians in the respective practices and me. This ‘bottom-up’ approach was used to maximize ‘the feeling of ownership of the project’ among veterinarians and farmers.

The practical procedure was as follows:

- All farmers and veterinarians attended an information meeting to outline the perspectives of herd-specific randomized trials in HHMP.
- The affiliated herd veterinarian agreed on the aim of the trial with the respective farmers.
- A workshop was held for the veterinarians within the practices to discuss and describe their usual procedures under my supervision.
- I produced a ‘design and protocol draft’ based on the workshop.
- Another workshop was held for the veterinarians to discuss the draft and agree upon a final trial design and protocol.
- An information meeting was then held for all farmers and veterinarians on practical issues related to the trial design and protocol, including data registration and adherence to the protocol.
- Discussion groups were held during the trial period to improve understanding and remain focused on the principle of randomization and reduction of bias.

The final trial design in all 11 herds was a ‘*non-blinded, ear tag-allocated parallel group design with active control*’. Four treatment protocols (A–D) were developed following the choices of the veterinarians in practices A and B and the herd contexts. All treatment protocols are described in Appendix C. Figure 1 illustrates the design as it was presented to the veterinarians and farmers in the trial. Consult manuscript V for a detailed description of trial design, protocol A, and data management in herds 1–4.

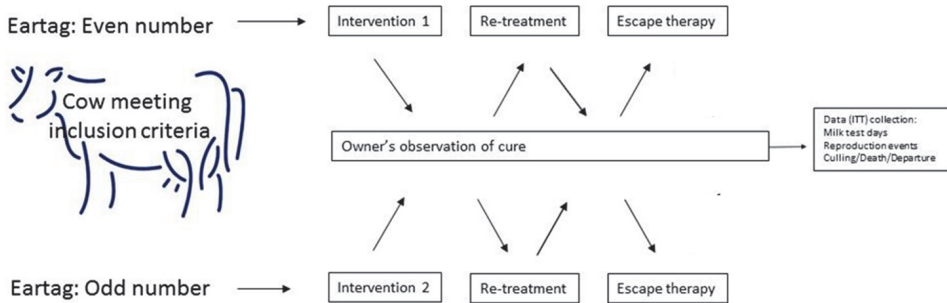
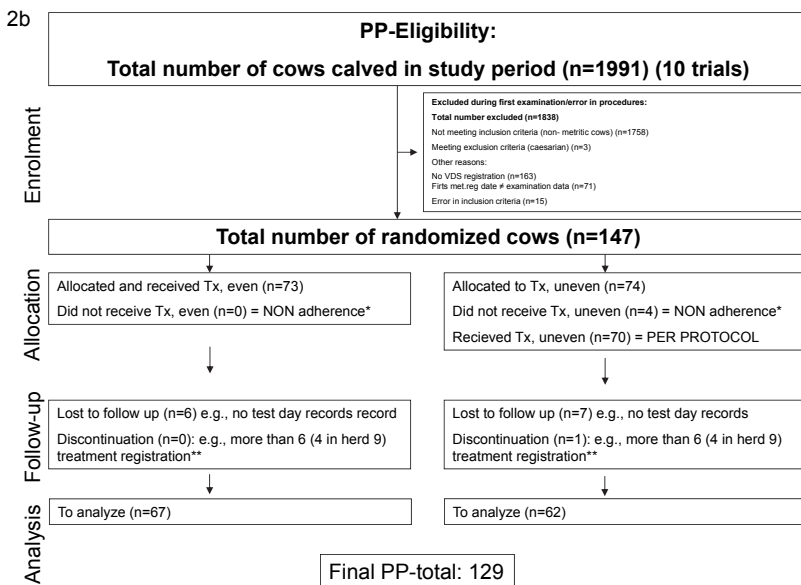
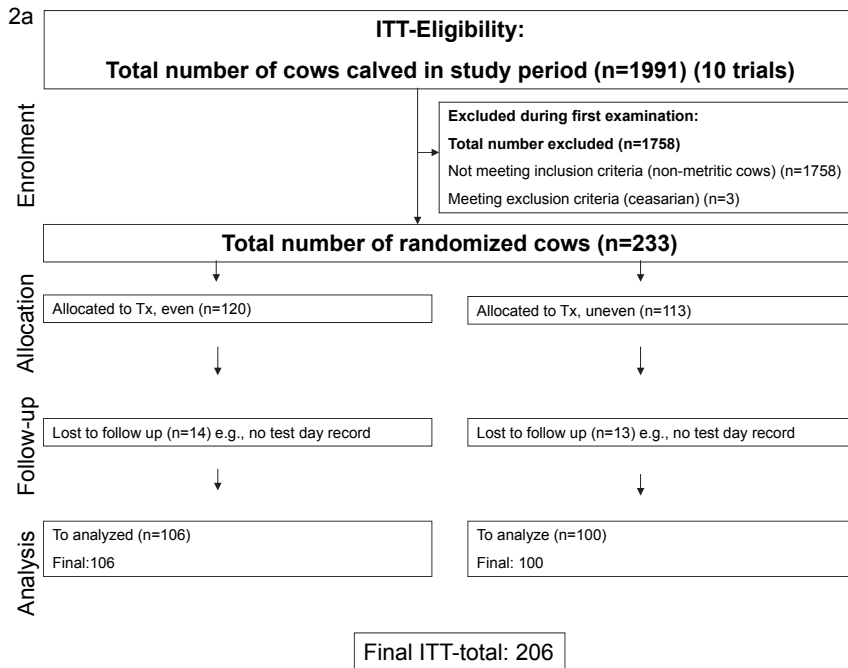


Figure 1. Diagram illustrating the trial design of 11 clinical field trials performed as an integrated part of a Danish herd health management programme in 2008–2009. ITT=intention-to-treat.

Data & Analysis

Both an ‘intention-to-treat’ (ITT) dataset and a ‘per protocol’ (PP) dataset were prepared to evaluate the overall level of adherence in the trials (figure 2a+2b). The ITT dataset included all cows with calvings in the study period with a metritis treatment 0–21 days postpartum (excluding cows with caesarean). The PP dataset was the result of the level of non-adherence to the protocol obtained in the field trial context. Consequently, all cows calved in the study period and examined and treated for uterine disease on the day of the herd visit 0–21 days postpartum were included if the cow was adherent (see definition below). If a cow calved twice within the study period, the first calving was included in the dataset (11 cases among 1,991 calvings were excluded).

Non-adherence was identified at both the barn and database levels. On barn level, veterinarians and farmers were asked to note any known deviation from the protocol and/or erroneous registrations on a form in the barn. The notes were collected twice times during the study period and categorized into two categories of non-adherence: 1) changed protocol (e.g., cow with even ear tag number receives treatment for uneven ear tag cows), or 2) a treatment other than the protocol used for retained placenta or metritis treatment (e.g., totally other treatment assigned). The non-adherent cases were excluded in the PP dataset. A third type of notification (e.g., cow supplemented with non-steroidal anti-inflammatory drugs in addition to correct protocol) was not considered as non-adherence. In addition, I detected a number of cows without VDS records (both metritis and non-metritic), cows treated for metritis before or after the systematic examination, and other errors in inclusion data. These cows were excluded from the PP dataset.



Figures 2a+ 2b These diagrams show the overall level of adherence in 10 clinical field trials integrated into the HHMP and subsequent reduction of the number of cows in the 'intention-to-treat' (ITT) dataset compared to the 'per protocol' (PP) dataset. The diagram is modified after 'consort diagrams' [24]. ** For the definition of discontinuation, refer to manuscript V.

At the database level, non-adherence can also be defined as cow data that does not follow the correct pattern of treatment registrations in the database. Two different errors of potential non-adherence were identified in the database when data were screened for irregularities during the study period: 1) missing registrations and 2) error in follow-up registrations. A pragmatic decision was made that database irregularities would not be regarded as non-adherence. These database-level cases were edited and cows included in both datasets. We made this judgement after asking the farmers for the reasons for these errors, and they described a 'pattern of convenience recordings' (e.g., easier to record one day instead of two). Thus, the cow would in most cases have received the medication according to the protocol. We also refer to manuscript V for a discussion of this decision of editing. Only the ITT dataset is used in manuscript V.

Statistical analysis was performed on data obtained in the randomized trials to illustrate possible methods of quantifying the difference in treatment effect and disease effect despite treatment for genital diseases on milk production. Random coefficient models were used to predict milk yield on day 60 postpartum and the total 305-day milk yield. Multivariable least-squares models were used to estimate effects of treatment and disease. I refer to manuscript V for a more detailed description.

3 Results

3.1 Summary of results

Here I summarize the results from the entire Ph.D. project entitled ‘Herd-specific Randomized Trials - an approach for Effect Evaluation in a Dairy Herd Health Management Program’. The tutorial in section 3.2 guides the reader through the concepts of effect evaluation of intervention, evidence-based medicine, and decision support in a dairy HHMP context. The tutorial gives examples from the other studies in this thesis. Subsequently, I present four studies. First, I present a review of the scientific evidence related to diagnosis and treatment of bovine genital diseases in the early postpartum period. The review focuses on metritis diagnosis based on vaginal discharge as it can be recorded in a Danish HHMP context. Second, I use the metritis case to describe the action patterns (diagnosis and treatment of metritis) used by veterinarians in the Danish HHMP and relate these action patterns to possible trial designs suitable for effect evaluation of interventions in practice. Third, I explore Danish veterinarians’ life-worlds and background for decisions related to metritis within the HHMP context and how these individualities can affect data quality. Fourth, I describe the implementation of a randomized clinical field trial in 4 dairy herds within a veterinary practice. This implementation allows me to elaborate on limitations and prospects of systematic effect evaluation based on randomized trials in HHMPs.

Manuscript I - section 3.2

Evidence-based veterinary practice for a dairy herd health management context – a tutorial

Veterinary medicine has developed by continuously implementing discoveries from human and veterinary medicine research into daily practice. Currently, as veterinarians organize into larger groups, consensus about best practices for decision making ought to be achieved in the practice unit. Therefore, the veterinary practice units need to get more involved in systematic evaluation of such new discoveries and the practices they currently apply. Clients and veterinary authorities also focus more on documentation for the applied interventions, including prudent use of antibiotics. Computerized data recording has facilitated the veterinarians’ analysis of data, which enables a professionally working group of veterinarians to set up their own systems for providing scientific evidence about the effects of current and new interventions in their larger dairy herds.

Structured around examples from dairy cattle practice, this tutorial provides step-by-step information that will allow a well-qualified and highly structured veterinary practice to create new science-based knowledge

and evaluation techniques of relevance to veterinary practice—the implementation of randomized trials in a HHMP.

We suggest an organizational framework for implementing the trial approach for effect evaluation in HHMP consisting of six phases in a systematic iterative cycle: 1) identification and reduction of the problem, 2) trial design, 3) starting phase, 4) trial conduct with data collection, 4) quantitative analysis, and finally, 6) qualitative effect evaluation and decision making. The trial could potentially lead to changes in action patterns, promoting an ‘evidence-based veterinary practice’.

To support such future development in the veterinary community, we propose the establishment of some unit for trial design and analytical support. This unit should support the development of competencies within the field of evidence-based veterinary medicine and practice among veterinarians in dairy practice. The supportive unit must have competencies in epidemiology, qualitative research, clinical science, education, and management of human resources to support the veterinarians in practice for on-going professional (and personal) development.

Manuscript II – section 3.3

Review of effectiveness of medical treatment for early postpartum bovine genital disease based on vaginal discharge

Evaluation of the disease effect and treatment effectiveness of genital disease (metritis) before 21 days postpartum in dairy cattle is important in contexts where the treatment criteria and the applied protocol have not been fully validated with a scientific approach. An initial step in evaluating applied protocols should be to compare the currently used clinical diagnostic criteria and treatment protocols in a Danish HHMP with the best available scientific evidence in the literature to identify practically relevant options for improvements. The review here is restricted to addressing occurrence of clinical metritis and evaluating medical treatment effects that are relevant in a real-world herd health management situation. We evaluated the scientific evidence from the literature and judged the relevance for a Danish HHMP setting where genital disease is diagnosed systematically by clinical examination of vaginal discharge in most cows during the 5–21-day postpartum period.

We found suggestions for uniform clinical definitions of bovine genital disease in the literature, but comparison between studies was complicated. Vaginal discharge was a major clinical sign of genital disease, but scoring scales varied. We found that the term ‘effect’ was used variably. For use in a HHMP, we suggest that effect should describe a disease effect, a disease effect despite treatment, a treatment effect, or a difference in treatment effect. This distinction will facilitate the organization of the evidence to be

used for practical decision making and recommendations. Often, the studies did not provide estimates of the relationships between the proposed clinical disease definitions and important key performance indicators, or the associations could not be estimated because the disease definitions contained information about the performance indicator. The most frequent key performance indicator evaluated was reproduction performance. We only found few randomized trials with negative and active control groups to evaluate treatment effectiveness and differences in effectiveness of medical treatment for early postpartum genital disease on milk yield. There was some evidence of milk loss and impaired reproduction (a general disease effect) caused by postpartum genital disease before 21 days postpartum (diagnosis based primarily on vaginal discharge). No general practical recommendation on treatment was available. The reviewed treatment trials and general recommendations in other reviews emphasized the following issues: choice of antibiotic including administration route and dosage, herd differences in treatment effect, effect of interaction between retained placenta and metritis, the importance of spontaneous recovery, and diagnosis and evaluation of ‘fatal cases’. For the Danish HHMP context with examination before 21 days postpartum, studies of spontaneous recovery, postponed treatment, and validity of the vaginal discharge score in relation to milk yield are particularly warranted.

Manuscript III – section 3.4

Diagnostic procedures and medical treatments for bovine genital disease in Denmark - a qualitative analysis of the potential for implementing herd-specific randomized trials in a herd health management program

Decision making in a HHMP should be supported by valid recommendations for diagnostic procedures, treatment thresholds, and treatment protocols. Genital diseases in a Danish HHMP are diagnosed with systematic clinical examinations of all or a majority of cows, 5–21 days postpartum, including a vaginal discharge score (VDS). This study addresses the potential of combining this systematic approach to diagnosis with experimental trials. We conducted semi-structured interviews with 12 veterinarians in the HHMP, and we observed their work in Danish dairy herds. With a tool designed to structure trial development (PRECIS), we linked the empirical descriptive data on how veterinarians work within the system to a pragmatic–explanatory continuum to identify conceptual trial designs that had potential for implementation in cattle practice. We also linked the identified action patterns to practical trial designs. We found a wide range of procedures implemented in the HHMP, despite the intended uniformity of the framework. Action patterns were linked to each veterinarian’s perception of focus point, treatment aim, and VDS scoring rationale. The results indicated the potential for implementing trials with pragmatic trial designs. That is, estimates of intervention effectiveness were more useful than evaluating treatment efficacy. Insights into veterinary procedures and attitudes in the HHMP context were used to discuss

practically relevant combinations of trial components. We concluded that because of literature discrepancies and shortages, limited scientific evidence exists to justify current Danish HHMP procedures. Clinical field trials within the pragmatic–explanatory continuum should be carefully adapted to individual veterinarians/practices and conducted by highly motivated participants to ensure success. We suggested three types of trial designs: (1) an explanatory within-herd trial with a clinical focus; (2) a pragmatic within-herd trial with a production focus; and (3) a pragmatic multi-herd/within-practice trial with a production focus. Furthermore, we proposed a practical approach, non-blinded treatment, and simple group allocations and trial designs that include active or negative-controlled parallel groups, modified cross-over treatments, or factorial groups, depending on the circumstances within the herd.

Manuscript IV – section 3.5

Veterinary decision making in relation to metritis – a qualitative approach to understand the background for variation and bias in veterinary medical records

Results of analyses based on veterinary records of animal disease may be prone to variation and bias because data collection for these registers relies on different observers in different settings as well as different treatment criteria. Understanding the human influence on data collection and the decisions related to this process may help veterinary and agricultural scientists motivate observers (veterinarians and farmers) to work more systematically, which may improve data quality. Based on observations and semi-structured research interviews of veterinarians working within a HHMP, we demonstrate how data quality can be affected during the diagnostic procedures, as interaction occurs between diagnosis and decisions about medical treatments. Important findings included vaginal discharge scores that lacked consistency within and between observers (variation) and scores that were adjusted to the treatment decision already made by the veterinarian (bias). The study further demonstrates that veterinarians made their decisions at three levels of focus (cow, farm, population). Data quality was influenced by the veterinarians' perceptions of collection procedures, decision making, and their different motivations for collecting data systematically. We conclude that both variation and bias were introduced into the data because of veterinarians' different perceptions of and motivations for decision making. Acknowledgement of these findings by researchers, educational institutions, and veterinarians in practice may stimulate an effort to improve the quality of field data, as well as raise awareness about the importance of including knowledge about human perceptions when interpreting studies based on field data. Both recognitions may increase the usefulness of both within-herd and between-herd epidemiological analyses.

Manuscript V – section 3.6

Clinical field trials in a dairy herd health management program: treatment effectiveness on milk production in case of early postpartum vaginal discharge

Effect evaluation of therapeutic intervention could be improved by adding a ‘local herd/practice trial approach’ to the tool box for veterinarians conducting a HHMP. We explored the practical potential and limitations of a trial approach that aimed at estimating the difference in treatment effect and the disease effect of metritis despite treatment on financially relevant performance measurements: predicted energy-corrected milk yield at 60 days postpartum and predicted 305-day total yield. We designed and implemented a pragmatic, ‘within practice’, multi-herd, ear-tag–allocated, non-blinded active controlled clinical field trial in four private Danish dairy herds where a highly structured HHMP was in place. We designed the study with the local veterinarians who conducted the practical work in the trial. We allocated 136 cows with vaginal signs of metritis before 21 days postpartum to one of two treatment protocols (penicillin or tetracycline) and included 744 non-metritic cows in the analysis. Superiority analysis of ‘intention-to-treat data’ was performed by means of a multivariable ANOVA model taking herd, parity, and retained placenta into account. The trials were integrated into the HHMP for a one-year period. We experienced some analytical problems related to small, unbalanced group size due to low disease incidence and the ‘pseudo-random’ allocation procedure. Also, variance heterogeneity was problematic for the model of total milk yield. We found no statistically significant difference in treatment effects of the two protocols on short- or long-term milk yield. The disease effect despite treatment was inconsistent and differed in both magnitude and direction depending on herd. Adjustment for parity and retained placenta was required but did not interact with treatment.

The study demonstrated that a trial approach can be used as a practical and feasible supplement to a highly structured dairy HHMP for improving evaluation of the effects of interventions as in the case of therapeutic treatment for metritis. Estimates of different effects can be obtained through a relatively pragmatic and simple data collection and corresponding statistical analysis. No evidence of a difference in treatment effect on milk yield between the antibiotic protocols was found. However, heterogeneity of disease effect despite treatment was evident across herds. Despite the motivation of veterinarians and farmers and professional supervision, the obtained data quality and non-adherence to protocol emphasize the importance of qualitative interaction with data.

3.2 Evidence-based veterinary practice in a dairy herd health management context – a tutorial

Manuscript I

D. B. Lastein & C. Enevoldsen
Department of Large Animal Sciences
Faculty of Health and Medical Sciences
University of Copenhagen
Grønnegårdsvej 2, DK-1870 Frederiksberg C
Denmark

Evidence-based veterinary practice for a dairy herd health management context – a tutorial

D. B. Lastein ^{a*} & C. Enevoldsen ^a

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

*Corresponding author: dorte.bay@gmail.com (Lastein, D.B.): phone 0045-20641151

About this Tutorial (summary)

Purpose: Veterinary medicine has developed by continuous implementation of discoveries from human and veterinary medicine research into daily practice. Currently, as veterinarians organize into larger groups, consensus about best practices ought to be achieved in the practice unit. Therefore, these units need to become more involved in systematic evaluation of discoveries in scientific research and the scientific evidence underlying the practices they currently apply. Clients and veterinary authorities also focus more on documentation for the applied interventions, including prudent use of medical treatments, especially antibiotics. Computerized data recording has facilitated the veterinarians' analysis of data. For these reasons, it becomes both relevant and possible for a professionally working group of veterinarians to set up their own systems for providing scientific evidence about the effects of current and new interventions in their herds. Because of the large population size and increased automation of data collection, this systematization is particularly relevant and feasible in dairy herds. Structured around examples from dairy cattle practice, this tutorial provides step-by-step information that will allow a well-qualified and highly structured veterinary practice to create new science-based knowledge and evaluation techniques of relevance to veterinary practice.

Is this paper for you? The tutorial is written for veterinarians who want to develop their practice unit to meet increasing demands from society and clients. Society requires more and more documentation concerning prescriptions and professional conduct. Owners of animals or herds become more and more informed and critical. Veterinary practice is part of an industry, which probably must self-develop its services with less and less support from public research. Because this tutorial is built around examples from the dairy industry, it will also be useful for non-veterinary research and development groups delivering products and services to the dairy industry.

The scope of this tutorial: First a table of contents is provided. The first section presents three examples of concrete attempts to provide evidence of effects of medical treatments of a typical disorder in dairy cattle. We use the case of genital disease shortly after calving. These examples give an introduction to the concepts and methods that are elaborated on in the subsequent sections. The context, the veterinarians' work and tasks in dairy herds as a clinical decision maker and a herd health management advisor, are presented in sufficient detail to help readers understand the examples. Built around the examples in section 1, section 2 gives an introduction to concepts, terms, and methods used for systematic effect evaluation. Section 3 provides details about the principles behind randomized clinical field trials ('trial theory'), which we also demonstrate in example three. Section 4 reviews the concept of evidence-based medicine with a focus on applications for dairy herd health management. Views on the hierarchy of evidence are presented. Section 5 addresses barriers to implementation of clinical field trials in veterinary practice, including a detailed discussion of the requirements for data and assessment of data quality. Section 6 gives suggestions for organization of veterinary cattle practice to meet the needs for dynamic generation of knowledge.

Table of contents

Evidence-based veterinary practice for a dairy herd health management context – a tutorial..... 25

 About this Tutorial (summary) 25

 1. Examples of tools to generate knowledge from a population 29

 The blessings and curses of definitions 29

 Example 1. A tool for health performance measurement in a dairy herd 30

 Example 2. How much milk is lost due to disease in a dairy herd?..... 31

 Example 3. A randomized controlled trial in a dairy herd 31

 2. An introduction to concepts, terms, and methods used for effect evaluation..... 32

 Causal effects, effect evaluation, comparability, randomization, and trials..... 33

 Evidence..... 34

 3. Randomized clinical field trials – practice and theory..... 34

 Fundamentals of trials 35

 Randomized clinical field trials in a dairy context – basic theory and technical details 39

 4. Wider perspectives on evidence 51

 Effect evaluation based on observational performance measurements 54

 Further about evidence 55

 Quantitative research methods and effect evaluation 56

 Qualitative research methods and effect evaluation 59

 Evidence-based decision making..... 62

 Evidence-based veterinary practice and policy in the herd health management context 64

 5. Barriers to implementation of evidence-based service and clinical field trials in veterinary practice ... 67

 Data quality 67

 Human involvement 68

 6. Suggestions for implementation of randomized controlled clinical field trials in herd health management 69

Evidence-based veterinary practice for a dairy herd health management context – a tutorial

The blessings and curses of definitions

Common understanding and learning of any issue requires common use and understanding of descriptive terms. Different scientific disciplines have their specific languages, which ideally make exchange of information very precise and condensed within each discipline. However, the applied terms or definitions may often be abstract to the novice reader. In addition, different disciplines may use the same term for different entities, which of course can be confusing. Such an inconsistency also complicates cross-disciplinary research and understanding.

Many key terms will deliberately be repeated and elaborated upon throughout the sections of this tutorial in the process of building up this common understanding. The definitions are in line with the vocabulary used by Krogh (2012). Details on statistical calculation, formulas, and the methodology of observational epidemiological studies are largely omitted. Relevant textbooks must be consulted to obtain such detailed knowledge (for instance: 'Veterinary Epidemiologic Research' by Dohoo et al., 2003). Key terms are marked in bold italics the first time they appear in the text and are subsequently in italics only if they are especially important for understanding. The listing of terms should not be considered exhaustive for the definitions related to scientific fields of herd health management and effect evaluation.

The contents of this tutorial are synthesised and adapted from several sources, in particular the European Medicine Agency, the Consort organization, and other individual references (Consort-statement.org, 2011; Dohoo et al., 2003; EMEA, 2001; EMEA, 2012; Habicht, 2011; Schulz et al., 2010; Thorpe et al., 2009; Zwarenstein et al., 2009). As terms and definitions are not used consistently in the existing literature, the choices in this tutorial are our interpretation of the sources reviewed. In specific cases, we give additional references. Further, we include all available knowledge obtained in a Ph.D. project on herd specific trial (Lastein, 2012). As mentioned, we use bovine metritis, treatment and prevention as an example of a herd problem in tutorial.

1. Examples of tools to generate knowledge from a population

This section presents three examples of concrete attempts to provide evidence of effects of interventions against a typical disorder in dairy cattle (e.g., treat or prevent genital diseases shortly after calving) as they are developed and implemented for use in Danish dairy herds. These examples give an introduction to the concepts and methods that are elaborated on in the subsequent sections. The context, the veterinarians' work and tasks in dairy herds, is presented in sufficient detail to help readers understand the examples.

Example 1. A tool for health performance measurement in a dairy herd

Figure 1 is a time-series plot that shows the development in occurrence of vaginal discharge among cows in third and later parities in a dairy herd. Each dot in the figure represents an individual cow that calved in the herd and subsequently was clinically examined by a veterinarian in the following 5–19 days. At the examination, the veterinarian assessed the vaginal discharge on a scale from 0 to 9 (left y-axis) and recorded the findings. A decision was made about whether to treat a cow as a result of the conditions found. If treatment was initiated, the date, diagnosis, and cow identification number (ID) were recorded in a database. As time (x-axis) goes by, the development in vaginal scores over a period of one year can be assessed. One purpose of the chart is to detect changes over time in the probability of occurrence of genital diseases (estimate of probability (e.g., a score ≥ 5) on the right y-axis). The position of the trend-line does not change unless there is sufficient statistical evidence (Thyssen and Enevoldsen, 1994). However, the cow ID (ear tag number) with high scores for vaginal disease is also added. These cow IDs allow the veterinarian and the farmer to evaluate the health history and the subsequent disease history of each individual cow from records or by memory. The design and use of this type of charts are described in detail by Krogh (2012) (download at www.vpr.kvl.dk). Such a series of case studies of individual cows may give useful qualitative information about identification of possible causes of genital disease and effects of medical treatments. This sequential qualitative evaluation of subsequent individual cases provides a typical and repeatable way for clinicians to generate experience-based knowledge.

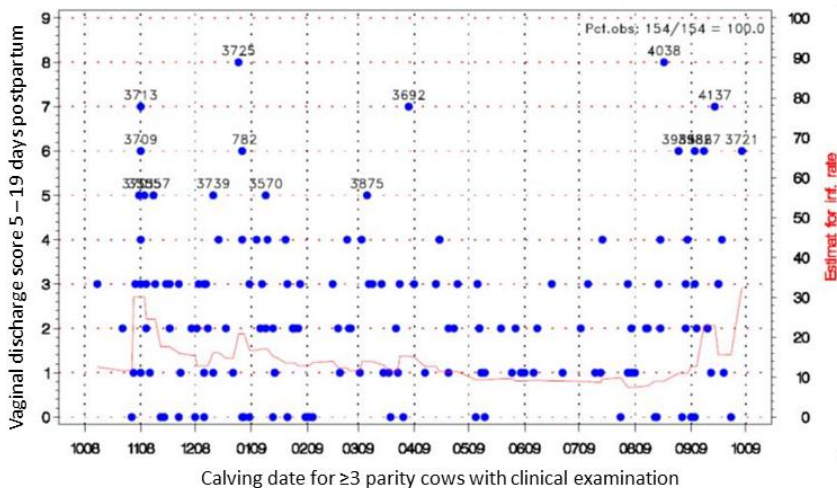


Figure 1. A time-series plot used to study occurrence of vaginal discharge of individual cows (dots) and trend-line of estimated risk (thin line) for third-parity and older cows at clinical veterinary examination 5–19 days postpartum in a Danish dairy herd. Causes of shift in the trend-line and evaluation of single cow cases can be used to generate experience-based knowledge in veterinary practice.

Example 2. How much milk is lost due to disease in a dairy herd?

Table 1 shows an extract from a report issued by a commercial Danish service to support herd health management (www.dhd-vpa.dk). There is a line of numbers for each of the parities 1, 2, and ≥ 3 . Each line shows the number of calvings in the study period, the average kilograms of energy-corrected milk per day 9–92 days postpartum, the percentage of calvings with one or more medical treatments for a genital disease, the estimated loss of milk in case of a genital disease, and the so-called p-value (statistical significance).

Table 1. An extract from a within-herd multivariable analysis (www.dhd-vpa.dk) estimating the disease effect (despite treatment) of genital diseases on average daily milk production 9 to 92 days postpartum (pp). For first parity, the 88 cows produced 30 kg energy-corrected milk (ECM) on average, 19% had experienced one or more genital diseases, and each cow with a genital disease produced 4 kg ECM less per day. The probability was less than 0.01 that this 4 kg difference (less) could be due to random fluctuations (the p-value). The results in the table are adjusted for a number of other factors that may have affected milk yield together with genital disease (for example, season of the year and other diseases).

Parity	No. of calvings	Average ECM/day 9–92 days pp, kg	% of cows with genital diseases	Less ECM/day if a disease record, kg	p
First	88	30	19	4	<0.01
Second	56	39	14	2	=0.38
Older	85	38	21	3	=0.06

The purpose of the analysis shown in Table 1 is to point out groups of cows that produced substantially less milk than their herd mates. With this information, the veterinarian can start a focussed search for causes of this exceptional variation in the herd. The applied medical treatment might have been inefficient or the genital disease might have been atypical. The hope is that the veterinarian can find the fundamental causes and solve the problem. If successful, the veterinarian has generated more experience-based and some analysis-based knowledge to supplement the knowledge obtained with the concept demonstrated in example 1. Numerous disease recordings can be analysed in a similar way, which allows for a screening of several potential causes of impaired performance in the herd. Krogh (2012) describes the use of this type of analysis and the interpretational pitfalls, which also are addressed below.

Example 3. A randomized controlled trial in a dairy herd

Table 2 shows the results of a randomized controlled trial in four dairy herds in a veterinary cattle practice. Ear-tags were used to allocate animals to either the previous medical treatment for metritis or a new treatment. The number of cows that received the new/old treatment, the additional milk production caused by the new treatment, and the p-values are shown for each herd.

Table 2. A subset of results from a clinical field trial performed in 4 Danish dairy herds as an integrated part of an ongoing herd health management program involving systematic clinical examinations. The trial was set up to demonstrate differences in treatment effect on milk yield (305-day yield) between two antibiotic treatment protocols. The trial did not provide sufficient evidence to claim that there was a substantial difference in treatment effect (Lastein , 2012).

Herd	Calvings with metritis	# new/old treatment	Difference between milk yield [kg ECM]	P=
A	70	36/70	482	0.21
B	18	7/18	-216	0.74
C	20	10/10	-228	0.70
D	28	21/7	-86	0.88

The analysis in Table 2 appears very similar to the analysis in Table 1. There is a fundamental difference, however. In Table 1, we categorized the cows according to the disease status that we **observed**. In Table 2, we **decided actively and at random** which cows should receive the new medical treatment and which cows should receive the previously used medical treatment. The categorization of cows in Table 1 might have been influenced by the herd manager’s preferences. For example, cows with high genetic merit might have received more intensive treatment. Such preferential treatment could cause systematic differences in milk yield that were **not caused** by disease per se. The randomization procedure used to create the data in Table 2 is an efficient tool to minimize systematic errors such as those arising from preferential treatment. Consequently, the scientific evidence of causal effects is stronger when randomization has been applied. We have demonstrated that it is feasible to conduct this type of randomization and subsequent analysis as part of a herd health management program (HHMP). The trial approach can thus be used to guide practical decision making in specific herd contexts.

2. An introduction to concepts, terms, and methods used for effect evaluation

In this section, we introduce definitions, key concepts, terms, and methods used to understand and create new science-based evidence of relevance to veterinary practice. References follow the principles introduced in section 1. That is, only in special case a specific reference is provided. The introduction is structured around the examples presented in section 1. The introduction in this section is followed by an elaboration in sections 3 and 4.

Causal effects, effect evaluation, comparability, randomization, and trials

When we use a plot like the one shown in Figure 1 (example 1) and see a marked change in the trend-line of occurrence of genital disease, we should start a search for the causes of this exceptional variation. We might discover that a simultaneous change in management of parturitions has occurred. If there is sufficient theoretical justification for a causal link, it can be justified to claim that the change in parturition management was the cause of the change in occurrence in genital diseases. First of all, we need to consider other simultaneous changes, such as changes in age distribution. That is, the cows calving before and after the management change should be ***comparable***, except for the management change. We can estimate the magnitude of a possible causal link with some metric. In example 1, we can calculate the difference in the risk of genital disease before and after the management change and call this estimate a ***causal effect***. The consequence of a truly causal link would be that elimination of the change in parturition management would prevent the change in the risk of genital disease. The ***time series analysis*** and the qualitative interpretation together make up one approach to ***effect evaluation*** for risk factors for disease, but also to evaluation of treatment effect on single cow cases given systematic clinical control after treatment. We elaborate on the justification issue in section 4.

When we use an analysis as described in Table 1 (example 2) and find a lower milk yield in the group with disease treatment compared to the group without disease treatment, it may also be justifiable to claim that the difference is *caused by* the disease (process) combined with a more or less sufficient medical treatment. This metric difference is then a ***causal effect of disease*** (despite treatment). To justify this claim, we at least need to consider whether there were reasons for disease treatment other than signs of disease. For example, the farmer might have used medical treatment to prevent disease in cows with high genetic merit for milk yield. The statistical analysis and the more or less qualitative interpretation of the assumptions and the results together form another approach to ***effect evaluation***. We will also elaborate on the required justification issue for this example in section 4.

In both examples 1 and 2, the claim of causality requires numerous assumptions, which may be hard to justify because known or unknown influential factors cannot be accounted for in all analyses. Many of these assumptions are avoided in example 3, which demonstrates the use of ***randomization***, a very strong tool for ***quantitative effect evaluation***. By randomly allocating the intervention (e.g., medical treatments) in question to the cows, we maximize the chance that ‘everything else is equal’; that the cows are ***comparable*** except for the medical treatments, the preventive management procedures, or other relevant procedures to be tested in the HHMP. Under practical circumstances, ‘random’ allocation can be conducted by systematically allocating the new treatment to every second cow on a list with calving dates. In this third

example, we used active interventions; we decided which interventions we would implement during the study. This is in contrast to examples 1 and 2, where we simply observed what had happened. Active intervention of some kind (ideally allocated by use of randomization) is usually called a **trial**. However, the benefits of using a trial do have a price. Numerous problems cannot be studied with trials for ethical or economic reasons (e.g., we cannot inflict severe disease at random). Practical issues may also hinder implementation of ideally designed trials in the field. We will elaborate on these issues in section 3.

Evidence

Knowledge-based, experience-based, analysis-based, trial-based, or a combination, is the scientific base from which veterinarians should take decisions in the HHMP. Ideally, this knowledge should represent the **evidence** for the justification of a given decision, so that the best advice is given. In this tutorial, we define *evidence* as one result of an **effect evaluation**. Basically, the strength of the evidence expresses the individual veterinarian's belief in the causal mechanisms that are being studied. We can use qualifiers like convincing, strong, or weak to describe (qualify) our belief in the causal mechanism and the estimated magnitude of the *causal effect*. We elaborate on these issues on evidence in section 4.

The terms related to population differs even between the worlds of observational and experimental studies. The terms presented here relate to the trial approach. Figure 2 illustrates some essential issues related to the use of evidence obtained from a study (in our case a trial). The trial is conducted in the **study population**. The study population can be a subset of cows from a herd (the **reference population**). If this subset is a random sample, we have maximized the chance that the trial effects are applicable to the *reference population*. If we have conducted the *trial* in the same way in multiple herds selected at random from all Danish herds, we have maximized the chance that the *trial effects* are applicable to the **general population**. That is, we have obtained **general** (universal) evidence of interest to the Danish veterinarians. Non-random exclusion of cows from the *study population* can cause **systematic errors** in the estimated trial effects (**selection bias**).

3. Randomized clinical field trials – practice and theory

In section 2 above, we argued that a randomized trial is the strongest tool to identify and quantify causal effects, in line with Habicht (2011). Therefore, we present first and mainly detailed information about trials. We expect that knowledge about trials makes it easier to understand other methods for *effect evaluation*. The following section is divided into two parts: 1) the fundamentals of trial conduct in practice (including a vocabulary); and 2) a rather detailed description of theory, definitions, and concepts related to trials that

have relevance for veterinary practice, particularly for implementation in veterinary dairy cattle practice. The aim is to provide sufficient technical detail to allow the reader to understand, plan, and conduct a simple trial in veterinary practice. Professional assistance will be required for the analytical phase depending on the statistical skills of the veterinarian in question. More advanced trial issues are briefly mentioned, but they most probably require professional assistance (in particular concerning statistical issues) and may be of limited relevance to the contexts of veterinary practice.

Fundamentals of trials

The following description puts words on the actions described in example 3 and adds definitions of fundamental importance for the design of simple field trials suitable for the HHMP context. We present a vocabulary to facilitate a common understanding. Again we refer to our general sources of information in particular the European Medicine Agency, the Consort organization, and other individual references (Consort-statement.org, 2011; Dohoo et al., 2003; EMEA, 2001; EMEA, 2012; Habicht, 2011; Schulz et al., 2010; Thorpe et al., 2009; Zwarenstein et al., 2009). As terms are not used consistently in all literature, the terms defined below and used in this tutorial is our interpretation of the sources reviewed. To exemplify the theory in the following text, we use bovine metritis as the disease entity and medical treatment of metritis as the experimental intervention.

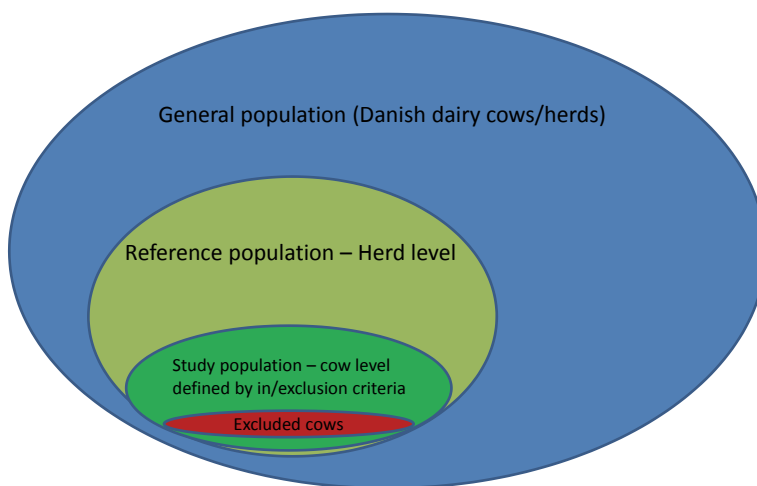


Figure 2. The definitions of general, reference, and study populations in the dairy herd context.

A vocabulary

In the following sections, we will primarily use the terms that are in line with the vocabulary suggested by Krogh (2012) and inspired by Krogstrup (2011). The overall principle of trials is to compare the **responses of**

interest (e.g., milk yield) to two or more different interventions like medical therapies, treatment strategies, screening, or management procedures. Numerous terms are used to describe or define the responses of interest and the measurements of these responses. **Manifestations** are frequently used in medicine to describe the responses to pathological processes in, for example, cells or organs. **Output** is also used to describe what comes from a process. The term **process measurement** is also used to describe the responses to processes. For example, vaginal discharge is one manifestation of the inflammation process, and recordings of vaginal discharge scores can be called *process measurements*. The term **results measurement** we use to describe the end-product of some chain of events or the consequences of the preceding procedures. For example, because cows are kept for milk production, recordings of milk production can be called *results*. The term **performance indicators** is widely used for business management. An **indicator** is a variable that measures the state of the trait (condition) of interest. **Measurements** can be either raw data or indicators, not to be confused with ‘a measure to handle something’. For use in statistical analysis, common general terms are *outcome*, *dependent variable*, *response variable*, or simply *Y-variable*. In contrast to *results measurements*, **effect** always relates to an incident (e.g., a randomly occurring disease) or an *intervention* (e.g., a treatment or preventive procedure) that is initiated to affect the results measurements (Krogstrup, 2011). More specifically, *effect* relates to the differences in *results measurements* caused by a given incident or intervention (**cause–effect relation**). Examples 2 and 3 above demonstrate tools to **estimate** the effects of disease (despite treatment) and medical treatment, respectively. We use the term ‘*estimate*’ to indicate the inherent uncertainty associated with the assumptions, the study design, the measurements, statistical methods, and the inferences. The **effect estimate** is the result of an **evaluation** (see further below).

A **clinical trial** is characterized by involving patients (in our examples, cows or herds) in any context. However, **clinical field trials** are explicitly conducted in ‘real world’ conditions (in our examples, cows in private commercial dairy herds) (Kastelic, 2006). If two (or more) interventions are compared concurrently, the trial is called a **controlled trial**. If no control intervention is applied, but an intervention is implemented, the trial can be regarded as a case study. Onwards in this text, we will only describe controlled trial even if the term is not used explicitly. The intervention in question is termed an **experimental intervention** and the comparison is a **control intervention**. If the control intervention is not an intervention per se (that is, we did nothing) or is a placebo intervention (e.g., injecting water), the trial is classified as a **negative or placebo-controlled trial**. Alternatively, if the control intervention is another treatment or procedure, the trial is an **active controlled trial**. If the interventions are allocated into different time intervals, the term **historical controlled** can be used. The before-and-after evaluation in example 1 can be considered a ‘*historical controlled*’ observational (not experimental) study. The choice of historical, active, or negative/placebo

control groups is a matter of clinical, ethical, and financial considerations. The interventions (two or more) are each assigned to a group of cows. These groups can be called **intervention groups**, **treatment groups**, or **treatment arms**.

Specification of the comparison group is essential and determines which *effect is estimated*. We propose the use of the terms: **disease effect**, **disease effect despite treatment**, **treatment effect**, or **difference in treatment effect** to distinguish these from each other. In the case of metritis in dairy cows, the definitions are exemplified as follows and in Figure 3. We define the disease effect of metritis as the difference in *results measurements* (e.g., milk yield, see elaboration below) between non-metritic cows (healthy) and metritic cows (diseased) without medical treatment. We define disease effect despite treatment as the difference in *results measurements* (e.g., milk yield, see elaboration below) between non-metritic cows (healthy) and metritic cows (diseased) with medical treatment. We define *treatment effect* as the difference in *results measurements* between non-treated metritic cows and metritic cows with medical treatment. Furthermore, we define *difference in treatment effect* as the difference in *results measurements* between metritic cows treated according to different protocols (e.g., two different drugs that presumably are active). If a trial is designed to evaluate *disease effects* or *treatment effects*, one of the intervention groups must be a placebo or a diseased group without treatment. A *difference in treatment effect* can be evaluated in active controlled trials where all cows will receive active treatments.

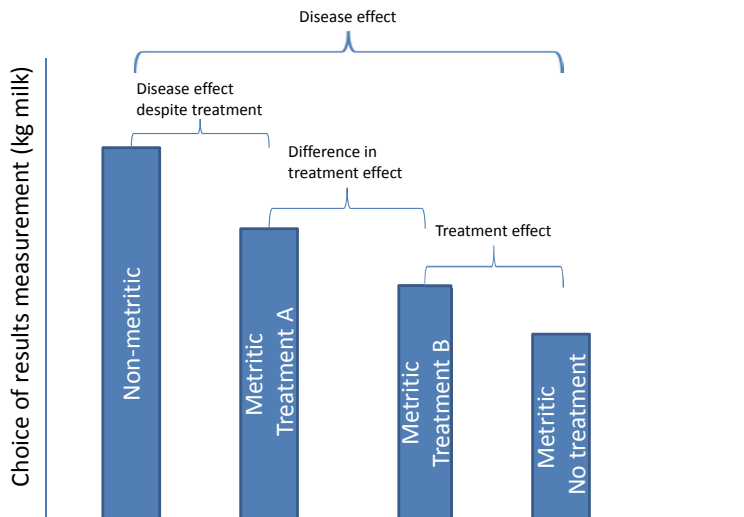


Figure 3. The distinctions among disease effect, disease effect despite treatment, treatment effect, and difference in treatment are illustrated with a constructed example of ‘effects’ of metritis and medical treatment on a given results measurement (kg milk).

Randomization, as mentioned earlier, is one of the most important issues related to trial theory because this process maximizes the probability that cows enrolled in a trial have equal chances to be assigned or allocated to either intervention group (Habicht, 2011). The issue will be described in practical detail below.

Blinding refers to unawareness of group allocation among one or more levels of the participants in the trial (in our case: the farmer, the veterinarian, and/or the analyst). In example 3, we could have masked the drugs so that farmers and veterinarians could not distinguish them, in which case the trial would have been double blinded. In human medicine, single blinding refers to the 'patient allocated being unaware of which groups he/she are allocated to'. In our case, this could mean that the patient is the cow and thus single blinding is irrelevant (e.g., no placebo effect is assumed). However, we adhere to the principles of considering the 'patient' to be the farmer (Kastelic, 2006). In our context, blinding is a very important tool if the different participants in a trial might have an interest in receiving or using one of the interventions or in promoting one of them (e.g., for a trial conducted by a drug producer, this conflict of interest might be suspected). Also, blinding is important if there are assumed risks of preferential care or treatment of some groups (e.g., farmer wants to protect high yielding cows). Also, this topic will be addressed in more detail later.

The most commonly used and accepted trial design in human and veterinary pharmacological research seem to be **parallel group design** that compares two or more intervention groups simultaneously. This design requires a relatively large sample size compared to designs that use the same animal as its own control (**cross-over trial or before-and-after**) or designs that adapt to the results measurements collected during the trial (e.g., **adaptive allocation** and **sequential design**) (Whitehead, 1992).

Superiority trials aim at showing that the average of the results measurements in one intervention group is statistically different by a predefined clinically relevant difference compared to the average in the other intervention group (can be designed to test both superiority and/or inferiority). **Equivalence trials** aim at showing that interventions have equal results or at confirming absence of a meaningful predefined difference (**equivalence margins**). **Non-inferiority trials** aim at showing that an experimental intervention is no worse (non-inferior) than a control intervention by a predefined margin (**non-inferiority margin**) (Christensen, 2007).

In the ideal trial, all cows that are randomized and allocated to an intervention group will 1) comply with the protocol (**compliance or adherence**) and 2) stay within the herd and be identifiable until the results measurements have been recorded (follow-up and no drop-out). Reality is another issue. Cows are **lost to follow-up, drop out**, or are discontinued because of culling, sale, death for competing reasons, or

insufficient human rigor in trial procedures. Maximal adherence and minimal drop-out are the goals in all trials. However, non-random non-adherence and/or drop-out can be considered worse because these factors will not only dilute any ‘true effects’ but also bias the results.

The responses to the interventions in a trial (the results measurements) can be divided into primary and secondary results measurements (outcomes). Ideally, a trial should have one **primary results** measurement (e.g., clinical cure or milk yield). In addition, **secondary results** measurements (also called exploratory outcomes resulting from subgroup analysis) can be evaluated in the same trial, but these results must be considered less valid. An example could be to evaluate across different prognostic factors (e.g., the effect of metritis on milk yield with or without retained placenta). Such analysis can demonstrate **heterogeneity in effect**; that is, the effect of treatment or disease varies or depends on the level of another factor (e.g., another good example are herd differences in treatment effect). However, trials can be designed to have the objective of estimating **multiple results** measurements or outcomes, sometimes referred to as multiple endpoints.

In addition, **heterogeneity** and **homogeneity** in the trial context are terms to describe the degree of biological or environmental variation. However, if the trial population is very heterogeneous (and the sample size large) and the analytical methods appropriate, it is more likely that the effect estimates from the trial are applicable to other herds (**generalizable** or **universal**).

Randomized clinical field trials in a dairy context – basic theory and technical details

Enrollment, randomization, and allocation

All or some (potentially selected at random) cows in one or multiple herds (so-called multi-centre trials) can be enrolled in a trial. When a cow is enrolled, it will be examined for eligibility based on a predefined set of inclusion criteria. The cow can subsequently be either excluded based on predefined exclusion criteria or be allocated to an intervention group. This allocation procedure can and should preferably be a random method (randomization). When a cow is allocated and has undergone the allocated intervention, results measurements are collected during a follow-up period. Analysis can begin when the follow-up period is finished and the results measurement of the last allocated cow is recorded.

Some additional details on the randomization procedures are warranted in this practical HHMP context. A complete random allocation process maximizes the probability that cows have equal chances to be enrolled in either intervention arm (e.g., tossing a coin or random number charts generated in a computer). Consequently, the chance is maximized that the cows in each intervention group are comparable with respect to potentially prognostic (disturbing) factors (e.g., parity or prior milk yield) that could also

influence the results measurements. Furthermore, randomization maximizes the chance that the intervention groups are of equal size at each given size of trial population (so-called balanced). As a consequence of randomization, assuming that randomization and subsequent allocation to intervention groups are performed correctly, the analytical methods applicable can be simplified (e.g., uni-variable instead of multivariable analysis). However, some caution must be considered concerning some methods for 'random allocation' that are more correctly termed as 'systematic assignments', non-random allocation (Dohoo et al., 2003; Kastelic, 2006) or perhaps pseudo-randomization. Examples are: identification number (even/uneven to either group) or sequential allocation (first animal to experimental group, next animal to the control group). Such alternative methods could interfere with blinding procedures (e.g., introducing risk of bias) and/or facilitate preferential allocation, in which case, the distribution of prognostic factors can be disturbed (Kastelic, 2006). An example of systematic assignments in a dairy context and related problems could be using ear-tag identification in herds where the numbers are NOT applied chronologically at calving but according to a (un)identified pattern decided upon by the herd-manager (e.g., equal number for heifers, uneven for bulls would be a problem in trial evaluating the effect of disease on growth). In human trials where the 'recipient of the intervention' is aware of allocation (e.g., risk of placebo effect and deliberate non-adherence and drop-out) and self-reporting on the outcome is more frequent (e.g., asking the patient instead of measurements), true randomization (and blinding) could seem of greater importance than in this dairy context. In all cases, the distribution of the most important prognostic factors in the intervention groups must be described in detail to evaluate the randomization procedures (also called baseline comparison). Statistical testing of difference is not recommended (Dohoo, 2003). In a **multi-herd trial**, the randomization and allocation can be done within each herd (can also referred to as block design; more similarity within herds than between herds).

Another important feature of an ideal trial is that the cows in different intervention groups are randomized, allocated, and treated within the same period of time (in trials with a non-historical comparison group). This design eliminates the time factor from the analytical perspective. For example, in the dairy herd context, feeding can vary considerably during the year, and feeding often can influence results measurements. However, if recruitment lasts longer than one production cycle (e.g., calving interval) or the effects of the intervention depend on the season, then season should be considered during design and analysis (Dohoo et al., 2003) (e.g., because of heat stress, metritis occurring in summer may be more severe and refractory to treatment).

The problem of preferential allocation of high-yielding cows or cows with severe clinical signs to one intervention group by the farmer or veterinarian can be prevented by *blinding*. For trials conducted in the

context of a single herd for the purpose of herd management, blinding probably is of minimal relevance because the manager should have no interests in cheating himself/herself and the results of the trial are not intended for marketing purposes. Trials can be either blinded (e.g., cows have no detectable awareness of their intervention group), double-blinded (e.g., the veterinarian or farmer that enrolls, allocates, and treats the cows does not know the intervention group), or triple-blinded (the cow, the person treating, and the person analysing do not know the intervention group).

Choice of primary results measurements

The choice of **primary results measurements** depends on the overall purpose of the trial and can range from a short-term clinical 'surrogate endpoint' (e.g., rectal temperature) or a long-term cow/farmer relevant outcome (e.g., survival relative to calving). We will demonstrate use of milk yield as a *results measurement* as an example of a pragmatic and relevant variable in the HHMP context. It could also have been a measurement of reproduction performance, an animal welfare indicator, or other relevant health performance indicators. Milk yield is also used in the subsequent text because it is considered the most important financial key performance indicator in the dairy context (Kristensen et al., 2008a). Careful considerations are needed when we select the results measurements for a study of the disease or treatment effects on milk yield. Are we interested in demonstrating short- or long-term fluctuations, daily yield, full lactation yield, peak milk yield, estimated 305-day energy-corrected milk yield, or persistency (slopes from peak to 305 days postpartum)? Krogh (2012) describes the components of the so-called lactation curve in detail.

Figure 4 shows hypothetical average lactation curves for groups of non-treated and diseased cows, treated and diseased cows, and non-diseased cows to illustrate some considerations that are necessary concerning potential early drying off in case of disease, short-term fluctuation during disease, etc. Fourichon (1999) discusses this issue in great detail. The illustration also implies that some *results measurements* appear more suitable for disease effect evaluation than for treatment effect evaluation from a statistical point of view (e.g., need to show a possible statistically significant difference). For instance, using estimated or predicted 305-day yield to detect a disease effect (blue (top) versus red (bottom) line) is 'statistically easier' than using the same results measurement to demonstrate an effect of treatment of an early postpartum-disease (red (bottom) versus green (middle) line), given that the difference is true. This is because the numerical difference between the compared results measurements is larger in the former (better power of the test). Numerous other statistical issues influence the 'ease or difficulties' of establishing of statistical evidence (discussed below). However, our point is that the statistical point of view might not coincide with a pragmatic point of view in a trial context: The pragmatic veterinarian and farmer could argue that if no

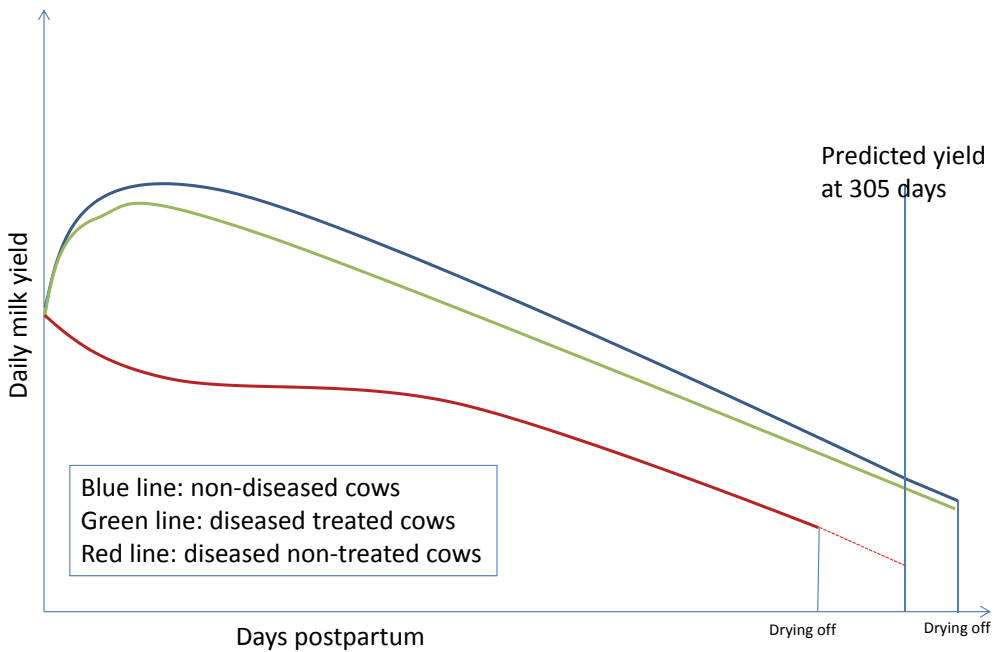


Figure 4. Three average lactation curves illustrating the need for clear definitions of results measurements in a clinical trial on disease or treatment effects related to bovine genital disease. Modified after Fourichon et al. (1999).

large treatment difference between diseased treated and diseased non-treated cows in 305-day yield exists, then the treatment is not worth the time and money. This is the same as saying that short term results measurements are irrelevant from their perspectives. The choice of milk yield as a results measurement requires knowledge and assumptions about the ‘behaviour’ of lactation curves for healthy and diseased animals (the pathogenesis) and knowledge of the preferences of the farmers and veterinarians in the HHMP to pick an appropriate disease or treatment results measurement. A recent analysis illustrates a way to estimate the impact of postpartum genital diseases on parameter estimates from the entire lactation curve (in that case, a disease effect despite treatment) including correlations between the different parts of the curve (Hostens et al., 2012). Similar principles of correlations can be used to evaluate treatment effect on lactation curve parameter estimates in a trial situation if an untreated control group is included and interaction terms between treatment and disease status are tested. Our description here concerns milk yield as a choice of results measurement to illustrate the complexity of

deciding which definition of outcome to choose. A similar complexity would arise for other disease entities, for instance, mastitis or ketosis.

Sample size

To allow for statistically valid inferences from trial data, the required **sample size** should preferably be estimated prior to the trial when the *primary results measurement* is chosen. The *sample size* required for a study can be affected by both statistical (e.g., prior knowledge of variation) and non-statistical (e.g., clinical, practical, or financial) considerations. Factors with influence on the sample are listed below. Several decision criteria are also presented. They are chosen by the research communities, perhaps arbitrarily, or by convention. The *sample size* usually must be higher if the trial context is very heterogeneous (e.g., several breeds in the trial population) compared to a homogeneous population (e.g., the same breed and/or parity).

The significance level [in technical terms, the type I (α) error] determines the probability that a given observed difference in the averages of *results measurements* (clinically relevant, e.g., 3 kg milk at peak yield) between intervention groups is due to chance (in a superiority trial; see below). A very commonly used α -value is 0.05, which means that there is less than 5% probability that the observed difference was due to chance (given the null hypothesis that there truly was no difference). The required sample size can be reduced by choosing higher α -values, but higher values simultaneously cause more *false-positive trial results*. However, the intended use of the trial results should determine the choice (Christensen, 2007). The α -value is directly related to the confidence interval, which gives lower and upper limits of the observed difference. If α is 0.05, there is a 95% chance that the true difference is within the confidence interval of the difference observed in the trial. For example, a 3-kg difference in peak milk yield observed in a trial could have a 95% confidence interval between 1 and 5 kg.

The power of a trial describes the probability of detecting a difference (larger than the clinically relevant difference) between intervention groups **when a difference truly exists**. In technical terms, power is 1 minus the type II (β) error. The required sample size can be reduced by choosing lower values for power (that is, increasing type II error), but lower values simultaneously cause more *false-negative trial results*. Often the selected (based on tradition or specific purpose) power is 90% in trials compared to 80% in observational (non-experimental) studies (Dohoo et al., 2003).

Clinically relevant difference: This criterion should be determined by clinical—**not** statistical—considerations. The value is used in trials that aim at detecting a certain difference in results measurements between intervention groups (*superiority trials*). When the statistical significance value is fixed (e.g., at

0.05), a small value for clinically relevant differences requires a larger sample size than a high value. The clinically relevant difference can be evaluated from the perspectives of cows, farmers, veterinarians in practice, or researchers. For instance, a 5-kg loss in peak milk yield due to disease may be very relevant for a farmer, but if the cow is not systemically affected, the welfare of the cow may not be compromised. Furthermore, researchers might find a difference of 1 kg milk relevant for a national breeding program, but the veterinarians in practice could argue that this difference is too small to consider for herd management purposes. In studies that aim at detecting **equivalence** (similar to) or **non-inferiority** (no worse than) of the effect estimates, the magnitudes that you search for are called equivalence or non-inferiority margins, respectively. These margins should be planned to be smaller than the clinically relevant difference (Christensen, 2008). Therefore, these types of studies require sample sizes up to four times higher than superiority trials (Christensen, 2007).

The statistical properties of results measurements: The scale of the chosen primary results measurements will usually be continuous/interval (e.g., milk yield), binary (e.g., treated/not treated), ordinal (e.g., vaginal discharge score at clinical examination), a count (e.g., number of re-treatments), or time to an event (e.g., time to pregnancy). The type of measurements (the scale) affects the choice of analysis method and sample size. For instance, a continuous *results measurement* requires a smaller sample size than a binary measurement.

Level of adherence (or compliance) and loss to follow-up (or drop-out): Non-adherence to protocol and loss of patients before measurement of results must be encountered. The sample size should be adjusted for the expected loss of study units (e.g., cows dying in the first three months of lactation if peak yield is to be estimated) (Dohoo et al., 2003) or analytical methods to account for this drop-out (selection bias) can be used (e.g., for instance, extrapolating short lactation curves to 305-day curves – predicted milk yield). We find that the terms ‘adherence’ and ‘compliance’; can be used interchangeably. We use adherence in the following. The degrees of non-adherence and loss to follow-up in data and the analysis determine which inferences that can be made on the results of the trial (details below) given that randomization procedures and sample size are adequate. The analysis of a trial can be very simple if we have conducted a trial with several hundred cows in a homogeneous environment (see below), full adherence, perfect randomization, and no loss to follow-up.

Heterogeneity > homogeneity in the reference population: The terms refer to the large or small (biological) variation, respectively, in the characteristics of the population you sample from (the *reference population*, e.g., a herd). For instance, the larger the variation [standard deviation (SD)] in milk yield, the larger the sample size that is required to detect a statistically significant difference between intervention

groups, if a difference truly exists. In a multi-centre trial (e.g., the same trial conducted in several dairy herds), a cow is usually more similar to her herd-mates (within-herd variation) than to cows in another herd (between-herd variation) because the cows within a herd are exposed to the same environment or are more alike genetically. A larger sample size is required to account for this environmental heterogeneity, which is also called *clustering*. Very strict criteria for inclusion and exclusion of cows in the trial (eligibility criteria or sampling frame) increase the homogeneity and decrease the required sample size. The disadvantage of homogeneity is that the cows in a *trial population* differ much from the general population, where the results of the trial should be used (not generalizable or applicable results). Consequently, the application of the intervention in the real world may give results that differ substantially from the results of a trial with a narrow sampling frame. If that is the case, the *external validity* of the trial results is poor. Logically, heterogeneity in the trial population increases the probability that the trial results are applicable to a wider range of animals and environments, and the ideal situation is that the trial is conducted in the context where the results are to be used. Note the difference to the term '*heterogeneity of effect*' that we use when the analysis has demonstrated the degree of variability of effects of treatment in different sub-populations.

Multiplicity problems (error inflation) must be considered (Bender and Lange, 2001; Proschan and Waclawiw, 2000) if the trial design contains the following components: multiple (>2) comparison groups (e.g., two treatments and one negative control), multiple results measurements (e.g., both milk yield and pregnancy rate), repeated measurements (e.g., milk test days records), subgroup analysis (e.g., testing difference between parities), or interim analysis (e.g., analysis before the trial has ended) (Dohoo, 2003). The multiplicity problem may increase the risk that a truly non-significant (statistically) difference is declared significant. Techniques for adjusting sample size to account for this problem in the analysis exist but may require professional expertise. Results from subgroup analysis should be considered when developing hypotheses for new trials.

The issues addressed above show that sample size estimation can be a complicated task that probably requires professional statistical assistance, especially in cases of multi-herd trials or *multiplicity problems*. Calculations are straightforward in cases of a relatively simple trial design in a single herd with one primary outcome. Suitable software programs for sample size calculations are available free of charge on the internet (e.g., 'Winpepi' at Brixton University). However, in practice, the 'effective sample size in a HHMP trial' will be determined by circumstances such as disease incidence and study period. For instance, a low disease incidence will require a long study period to obtain the 'estimated sample size', which could decrease the applicability of the results as herd-specific circumstances might have changed over time.

Conceptual trial design - the 'pragmatic–explanatory continuum'

The overall conceptual purpose of a trial should be considered in an initial trial design planning phase. You need to ask and get answers to the following questions:

- Is the trial intended to demonstrate and explain biological associations (e.g., does a medical treatment of puerperal metritis reduce the risk of subsequent subclinical endometritis)?
- Is the trial intended to provide pragmatic decision/management support in practice (e.g., does medical treatment of puerperal metritis reduce milk loss in a specific herd)?
- Is the trial intended to provide pragmatic decision support in politics (public management) at the industry or national level (e.g., should postpartum examination and treatment of clinical metritis be implemented nationally by means of legislation)?

When you have answers to these questions, you can position your trial within the '***pragmatic–explanatory continuum***', a framework that defines how a trial design fits into either pragmatic or explanatory design principles (Thorpe et al., 2009). Pragmatic relates to the term *effectiveness* and how interventions work if used in practice. ***Explanatory*** relates to the term *efficacy* and how interventions work under ideal settings. The first question above illustrates an explanatory trial seeking universal or general evidence concerning biological relations. The second question illustrates a pragmatic trial seeking local evidence for herd management purposes. The third question illustrates a pragmatic trial seeking general evidence for public management purposes. The consideration of 'pragmatism versus explanation' or 'efficacy versus effectiveness' and the importance of being consistent within the continuum of trial designs between the extremes are described in detail by others (Gartlehner et al., 2006; Thorpe et al., 2009; Zwarenstein et al., 2009). Another type of effect evaluation, *efficiency*, is addressed with cost–benefit analysis (Haynes, 1999) and will not be discussed further.

Our emphasis in the following will be on the perspective that the justification of trials in the whole of the pragmatic–explanatory continuum is that they can serve as evidence for herd management decisions. The important issues are that different trials answer different research questions and that they are designed and conducted consistently. That is, the specified question should guide all other choices taken. [Table 3](#) demonstrates the extremes of the 'pragmatic–explanatory continuum' by means of examples from the context of studying genital disease in a dairy herd (this context is described in the examples in section 1 and in detail in related work (Lastein, 2012)).

Table 3. The key differences between trials with extreme explanatory versus extreme pragmatic attitudes are described. The table is adapted according to tables and text by Zwarenstein et al. (2008) and Thorpe et al. (2009), and subsequently modified to handle trials related to genital diseases in a dairy herd health management context. Consequently, some very extreme explanatory elements, such as laboratory trials, are omitted.

	Explanatory extreme: Efficacy trial – ‘clinical trial’	Pragmatic extreme: Effectiveness trial – ‘clinical field trial’
Question or Overall aim	Efficacy – can the intervention work under ‘ideal settings’?	Effectiveness – does the intervention work when used in practice?
Settings	Well-resourced ‘ideal’ setting (e.g., university herd and trained personnel/researchers)	Normal practice (e.g., private herds and veterinarians in practice)
Recipient of intervention	Highly selected cows. Cows are excluded if they are non-adherent or otherwise atypical so that they may dilute or distort the effect	Little or no selection (e.g., few in/exclusion criteria)
Intervention	Strictly enforced protocol and adherence monitored closely	A flexible protocol applied, as in normal veterinary practice
Outcome	Often short term and frequent surrogate measurements (e.g., daily temperature measurements and blood parameters)	Measures that are directly practically or financially relevant to farmer and veterinarians in practice (e.g., clinical treatment success/failure, milk yield, or pregnancy chance)
Relevance to practice	Indirect relevance: little or no effort is made to match design of trial to practical decision making among veterinarians or farmers	Direct relevance: the trial is designed to meet the needs of those making the decisions about treatment options
Adherence of participants	Close continuous monitoring of adherence to protocol (cows) and by trial personnel (farmers/veterinarians). Strategies for improvement predefined.	Unobtrusive or no monitoring of adherence to protocol (cows, farmers, and veterinarians)
Analysis	Analysis is performed on both reduced ‘correct’ and ‘real-world’ data [per protocol (PP) and intention to treat (ITT), respectively]. See definitions and description elsewhere in this section.	Only ‘real-world’ data are analysed (ITT data)

Analytical trial design

We use the term ***analytical trial design*** when we refer to the strategies concerning principles of data management and analysis. To ensure coherence with the choice of conceptual design (see Table 3, last row), essential decisions are described below regarding how to organize data before analysis and how to analyse the data. Implications of the choices of analysis and validity of results are also described below.

Data management principles

Non-adherence to protocols (e.g., a cow receives only 1 out of the 3 required doses of antibiotic) and loss to follow-up (e.g., a cow dies before peak milk yield) results in a difference between ‘real-world’ data (raw

data) and reduced 'correct' data. The real-world data obtained in a trial situation represent data on **the intention to treat** or to follow a trial protocol (called *ITT data*). They include all the noise of practice, just as the real world does not follow a strict protocol. Corrected data are obtained through a data management process so that data include only the cows that followed the protocol to which they were randomized (e.g., cows were excluded from the dataset if they received the opposite treatment than the one they were randomized to get or they were misclassified as diseased based on incorrect inclusion criteria). Correct data are called **per protocol (PP) data**. A third category is designated 'As treated' (*AT*). Here, cows are analysed according to the protocol they were exposed to no matter the randomization and allocation. In the following, we elaborate on these three distinctly different principles of data management for ITT, PP, and AT.

- 1) **Management of ITT data** requires that all cows that are enrolled and have a results measurement (or a value estimated from raw data) are analysed as they were randomized. That is, all cows are included in the analysis despite failure to start intervention, non-adherence, and erroneous inclusions. Methods for handling missing results measurements are available (Hollis and Campbell, 1999). The ITT procedure maximizes the positive effects of randomization. That is, the influence of disturbing prognostic factors (e.g., parity) in the analytical phase is minimized.
- 2) **Management of PP data** excludes cows that did not receive the randomly assigned intervention in full compliance with the protocol or that did not have record of a results measurement. This exclusion implies that all non-adherent cows and cows lost to follow-up are excluded from analysis. Such a procedure might cause systematic errors (bias) in the trial's effect estimates if non-adherence or loss to follow-up were non-random: for example, if the experimental treatment was 'forgotten' more often and replaced by the active control treatment (which could be standard treatment before the trial started) or if culling before results measurements were available occurred more frequently after one of the treatments (loss to follow-up). The consequence could be a skewed (uneven) distribution of possible influential prognostic factors among the intervention groups, which also could bias the trial's effect estimates. In addition, there are fewer cows in the PP data than in the ITT data, which may reduce the chance of detecting a (true) statistically significant effect of treatment.
- 3) **Management of AT data** is controversial because it interferes with the randomization procedure. With this principle, cows are re-allocated if they received the opposite intervention than the one they were randomized to. This approach is not recommended because of changes it causes in balance in number and base-line comparability caused by re-allocation (Lee et al., 1991). If, for instance, cows are treated with the opposite intervention in an active controlled trial for non-random reasons (e.g., in cases of severely acute clinical signs of metritis that the veterinarian forgets to allocate correctly), then the distribution of cows with acute and less acute disease will no longer be comparable in the two groups. The results of the effect will be biased (e.g., the group with more cows with severe signs might produce less milk compared to the other group, a result that is not due to 'true difference in treatment effect' but to more sick cows in one group).

Analytical principles

The choices of conceptual design, comparison group, and data management principle influence the choice of ‘analytical principle’. We describe three analytical principles: superiority, equivalence, and non-inferiority. The analytical principles determine which possible and plausible statistical inferences can be made concerning the groups we compare.

Superiority testing implies that a statistical test is applied to show whether the intervention groups differ with respect to the chosen results measurement and in which direction. The superiority principle fits well into situations where we want to evaluate the effect of new interventions against placebo or negative control groups (e.g., to evaluate a ‘true’ treatment effect). We must be able to ethically justify use of negative comparison groups. In case of a trial with two intervention groups, a two-sided statistical test provides statistical evidence of a positive or a negative difference. For example, the average of results measurement Y was better after treatment A compared to treatment B, but the opposite outcome could have occurred (Habicht, 2011). The test is based on the assumption that there was no difference (the null hypothesis). The statistical test and the selected significance level dictate whether or not the null hypothesis of no difference can be rejected statistically (falsification). However, a superiority trial test with a statistically non-significant result does not show that the results measurements in the intervention groups are equal. High p-values (statistically non-significant) can be a result of things other than equality (e.g., low sample size and bias). In a superiority trial conducted according to the ITT principle, allocation errors will ‘dilute the difference’. That is, the estimate of treatment effect will be conservative e.g., a reduced risk of a false positive trial result. One important notion always to recall when interpreting a superiority trial is stated by Altman and Bland (1995): “Absence of evidence is not evidence of absence”. This means that if a study gives no evidence of difference in effect between groups, this does not indicate that there is equality in effect.

Equivalence trials and non-inferiority trials aim at showing that the results measurements in different intervention groups are equal or ‘no worse’, respectively. Non-inferiority trials are one-sided alternatives to equivalence trials. Equivalence and non-inferiority trials are suitable for situations where an effective intervention is already established and a new intervention can be tested against this active and effective control (e.g., active controlled trials). They are also useful or even needed where placebo intervention cannot be ethically justified. The use of these principles is also appropriate in situations where new interventions are expected to be less harmful (‘risk–benefit situations’) when comparing different doses of the same medical drug, and if the experimental intervention is less expensive or is not expected to have an

improved therapeutic effect compared to the control intervention (Christensen, 2007; Gøtzsche, 2012). Statistical analyses of equivalence and non-inferiority trials are based on the predefined margins (equivalence and non-inferiority margins) and the confidence intervals of the difference between groups with respect to the outcome measurements. That is, p-values are not used directly. If the confidence interval of the difference in results measurements between groups is entirely contained in the equivalence margin, equality or non-inferiority can be claimed if both the ITT and the PP analyses show the same result. The reason for including both ITT and PP principles in case of these trial types is as follows: If the ITT principle is exclusively used, the non-adherence and effect dilution result in overestimation of equality and increase the likelihood of a false-positive result (being equality or non-inferiority). Therefore, both PP analysis and ITT analysis of data are needed in equivalence and non-inferiority trials to give a valid evaluation of the effect (Habicht, 2011). The margins must be of clinical relevance (often set too large!) and smaller than the clinically relevant difference for superiority testing. In opposition to a superiority trial where the statistically non-significant differences in effect can never be used to state equality, the results of an equivalence (or a non-inferiority trial) that indicate differences in effect between interventions can be validly used to state that there are differences in effect (Habicht, 2011). The design, analysis, and interpretation of equivalence and non-inferiority trials might require professional statistical assistance.

The decisions about conceptual design (pragmatic–explanatory continuum), the data management principle (ITT, PP, AT), and the analytical design (superiority, equivalence, or non-inferiority trial) are complicated and context dependent. Within the practical trial dairy context, some interactions between different decisions will result in different consequences for trial conduct, the possible inferences (local or general effects, biological associations, or decision support), and the possible actions based on results of the trial. These interactions will be discussed here.

First, there are ethical considerations in case of a ‘disease or treatment effect’ study. Is it ethically acceptable to include a placebo or negative-control group? Legislation probably gives some constraints. The ethical perspective of risk of side effects of medical treatments should also be considered. Second, if ‘the active comparison approach’ is chosen, we need to know whether there is valid scientific evidence to document the effect of any comparison treatment (by means of other studies including a negative-control group). Such documentation is recommended in case of active controlled studies in all analytical designs but is not crucial in superiority trials (EMA, 2001). Third, the level of adherence is very context dependent. We speculate that the closer to ‘real-world’ circumstances, the lower the adherence to protocols. This association implies that equivalence and non-inferiority studies, which require PP data management and PP analysis, will benefit from high levels of adherence and thus be less relevant to ‘real-world situations’.

Consequently, PP principles adhere better with explanatory trials. In addition, non-adherence resulting from inadequate trial design (e.g., inadequate randomization or blinding procedures) can cause non-random error that could introduce systematic errors into the results (bias).

As a final remark on the analytical design, we find that superiority trials are most widely used in scientific research and that therefore most veterinarians in practice probably understand and accept them better than equivalence and non-inferiority trials. Also, in the veterinary community, we find reflections scarce on the distinction of the pragmatic–explanatory continuum and the difference in ITT/PP approaches (e.g., local/general policy decision versus biological inferences). In essence, the discussion also relate to the distinction between internal and external validity of a study and the *local* versus *general* estimates of effect. Lack of awareness about these issues could result in inconsistency in trial design and inferences with subsequent reduced validity and usefulness of the results. Elaboration on the issues is given below.

Practical trial design

We use the term ‘practical design’ to designate the practical setup for allocating and comparing cows in different systems. All practical trial designs have their advantages and disadvantages. Table 4 describes some relevant designs with the characteristics that are relevant to the dairy HHMP context. Further details on practical design are readily available in textbooks on trial theory and from reliable internet sources (e.g., www.ema.europa.eu, www.consort-statement.org).

4. Wider perspectives on evidence

The description of *trials* above demonstrates how to use *randomized controlled trials* to generate new knowledge about individual cows. Analysis can also be performed on higher level e.g., allocating herds to different interventions (Tempelman, 2009). Although the techniques may appear complicated and logistically challenging, we have demonstrated that they can be implemented in practice (Lastein, 2012), and apparently there is consensus about the major theories, practical techniques, and interpretation of results. However, the dairy cattle veterinarian faces much more complicated problems in which generation of new knowledge also is required. Figure 5 is a modified schematic representation of a ***farm*** ***system*** where the complex pathways involving animal, human, and additional farming resources are illustrated (Andersen, 2004). Raw data measured on cows, feed, housing, and production are collected and

transformed via analysis into information or results and used to measure performance. Dialogue among stakeholders combines quantitative information with qualitative data (e.g., farmer, veterinarian, and other stakeholder preferences, perceptions, and experiences). Decisions (at either a higher or lower level of the personal structures) are taken by the stakeholders, actions are adjusted or interventions are considered

Table 4. Advantages and disadvantages of practical trial designs of relevance for dairy herd health management are listed. The designs are mainly used to compare two intervention groups (A and B). If relevant, the designs are illustrated schematically and referenced. Continued on next pages.

Design/short description	Advantages	Disadvantages	Illustration									
<p>Parallel group (2, >2)</p> <p>Cows are randomized to A or B and outcome measurements are compared</p>	<p>Simple to implement and understand</p> <p>Widely accepted in scientific community</p> <p>Negative or active control</p>	<p>Large sample size</p> <p>Multiplicity problem if multiple outcome or interim analysis ('un-authorized pre-look')</p>	<p>A >< B</p>									
<p>Stratified group</p> <p>Stratified on prognostic factor (e.g., parity) and comparing A and B within strata</p>	<p>Relatively simple</p> <p>Takes prognostic factors into account (e.g., parity) to reduce heterogeneity in effect</p>		<p>Parity 1: A >< B</p> <p>Parity >1: A >< B</p>									
<p>Full cross-over: Each cow is randomized to A or B, is treated, and the outcome is measured. A washout period with no treatment follows before the same cow is randomized again to either A or B.</p>	<p>Cows are their own controls, which gives reduction of heterogeneity</p> <p>Gives opportunity for assessment of the sequences of treatments</p>	<p>Only suitable for persistent chronic incurable diseases</p> <p>Washout period between treatments</p> <p>Multiple randomization procedures</p>	<p>A>washout + randomization>A or B</p> <p>B>washout+ randomization >B or A</p>									
<p>Semi-cross-over</p>	<p>Possible to test different sequences of treatments (or re-treatments)</p>		<p>A>±washout>B</p> <p>B>±washout>A</p>									
<p>Factorial</p> <p>All combinations of 4 (or more) interventions</p> <p>A/B x +C/-C</p>	<p>Possible to test concurrently applied treatments of different kinds (for instance, antibiotics and hormones)</p>	<p>Complicated allocation procedures due to many intervention groups</p>	<table border="1"> <tr> <td></td> <td>+C</td> <td>-C</td> </tr> <tr> <td>A</td> <td>X cows</td> <td>X cows</td> </tr> <tr> <td>B</td> <td>X cows</td> <td>X cows</td> </tr> </table>		+C	-C	A	X cows	X cows	B	X cows	X cows
	+C	-C										
A	X cows	X cows										
B	X cows	X cows										
<p>Adaptive</p> <p>Characteristics of the study itself change during the trial in response to data being collected. E.g.,</p>	<p>Follows the clinical mind-set</p> <p>Ethical due to reduced number of cows getting inferior treatment</p>	<p>Complicated allocation procedures</p> <p>Not widely used in human or veterinary medicine</p>	<p>Example of 'play the winner' design. Figure adapted from (Bjerkset et al., 1997).</p>									

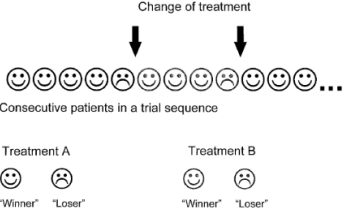
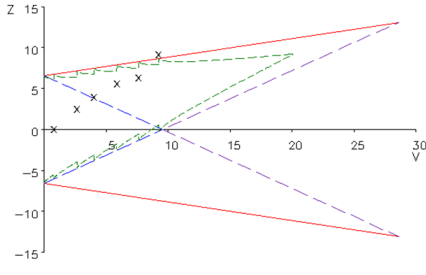
<p>the proportion assigned to active intervention versus control changes in response to results measurements</p>	<p>Well suited for critical disease</p>		
<p>Sequential Implementation of statistical procedures that allow for valid interim analysis for each cow (or group of cows) by predefined clinically relevant stopping parameters</p>	<p>Ethical due to stopping rules Can be superimposed on other designs (e.g., parallel group) Potential reduction of sample size Ensure control of multiplicity problems</p>	<p>Complicated theoretical background Workload due to interim analysis Statistical expertise needed at all phases; sample size estimation and analysis Sample size adjustment due to multiplicity problems (error inflation)</p>	<p>Example: Triangular design with Christmas tree correction (Whitehead, 1992)</p> 

Table 4 continued

and may be implemented. The subsequent results of the production can be measured and compared to the pre-intervention results to provide some level of effect evaluation.

Until now, this tutorial has primarily addressed the relationships in the little ‘cow’ box. However, we suggest that the farm manager also need support in herd level decisions. An example e.g., it is worthwhile to apply the HHMP as a whole? Should the farm manager choose another milking management system instead of the present one? Because the HHMP and other herd-level systems may affect the entire *farming system*, we clearly cannot provide an answer by using randomization of individual cows within the herd or similar cow-level studies to answer these questions.

In theory, we could do a large-scale herd-level study where we randomly allocated herds to different types of HHMPs to estimate a ‘*difference in effect of HHMP*’ at herd-level, as mentioned above. For financial and practical reasons, such an approach probably would be untenable. A major difference between cow and herd level studies is that in the latter we deal an even more complex structure of humans, values, complicated feedback mechanisms, and long- and short-term effects as each herd has their own structure as the one indicated schematically in Figure 5. Within herd, we farmers and veterinarians still need to

evaluate the effects of our interventions to develop ‘best practices’ with a resulting improvement of effectiveness, but we need increased knowledge of the structures in the specific farm. The system described in Figure 5 is analogous to other manufacturing or service industries and organizations in the public sector (public management) where systematic effect evaluation also is needed (Krogstrup, 2011).

In the following, we present approaches to estimation of effects (to provide evidence) from a sociological perspective primarily based on the work of Krogstrup (2011) and evaluate these approaches in relation to possible applications in the dairy herd context. The randomized controlled trial principle has been applied for evaluating the effectiveness of ‘public programs’ (e.g., teaching or health promotion) to support public management. Numerical principles similar to those of examples 1 and 2 are also widely used [observational data in contrast to randomized (experimental) data from trials]. Finally, qualitative research tools are used for effect evaluation. We give a brief overview of the options for evaluation based on observational numerical methods (mainly time-series analyses) and qualitative research principles.

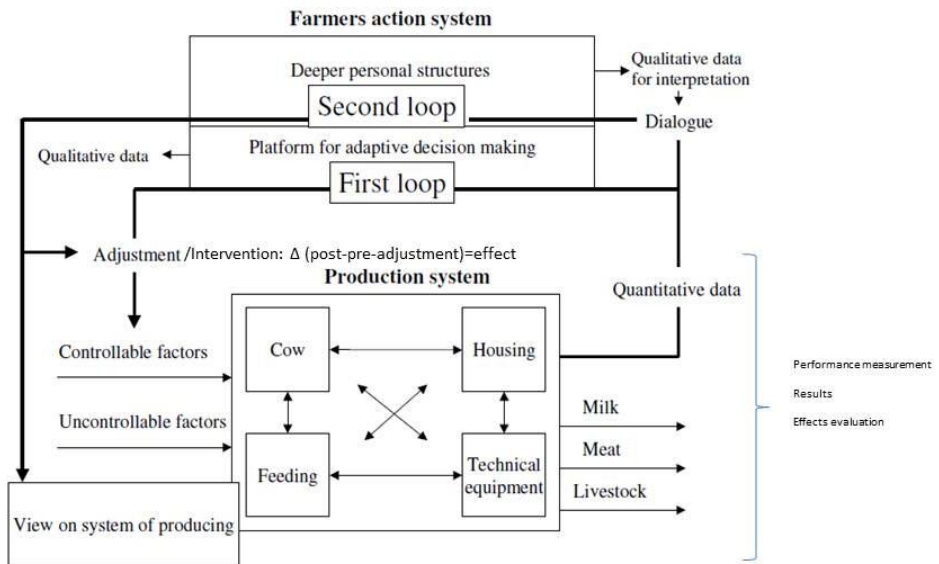


Figure 5. The complicated framework of herd health management requires disciplines from both natural (quantitative) and social (qualitative) science to evaluate performance and effects of interventions (Andersen, 2004). Modified to consider HHMP terms used in the present paper.

Effect evaluation based on observational performance measurements

A series of terms is used for description of tools to evaluate the performance of a system (e.g., a dairy herd or any other business). To clearly define effect evaluation, we must differentiate between effect evaluation and the following herd health management terms: measurement, monitoring, surveillance, control, and

evaluation. The distinctions are elaborated upon by Krogh (2012) with special emphasis on health performance measurement in dairy herds and by Krogstrup (2011) with emphasis on public management. Although originally used in a public management context, we decided to use an analogous definition and rank the different management tools used for analysis-based decision support:

- *Measurement* describes the process of recording raw or minimally reduced data on herd performance; for instance, milk production or health data. In the HHMP context, performance measurement can be divided into process measurement (for instance, proportion of calved cows that are examined between 5–21 days postpartum) and results measurement (for instance, average peak milk production).
- *Monitoring* describes a more or less systematic approach for collecting and using production or health information (information being indicators or results measurements produced from reduced or aggregated raw data). An example is the qualitative interpretation of a time-series plot of vaginal discharge and trend-line development (example 1, section 1).
- *Surveillance* describes a more systematic and active process to follow the development in production and health performance. An example could be to do regular multivariable analysis (example 2, section 1) to allow for decisions on potential intervention. Often, the distinction between monitoring and surveillance is unclear. Krogh (2012) suggests that *performance measurement* is a sufficient and informative term that covers both monitoring and surveillance.
- *Control* describes the process of evaluating the importance of a deviation between an obtained result (performance) and a target. In broader terms, ‘to control’ means to keep performance within certain limits (e.g., budgets).
- *Evaluation* describes a systematic, transparent, and retrospective (or prospective) process of distinguishing between valuable and non-valuable or acceptable and non-acceptable (most often) causal relations in production. That is, an evaluation seeks to verify if production, health, or the intervention is meeting predefined goals (targets, limits, or clinically relevant differences). An evaluation is aimed at yielding practical actions (after Krogstrup, 2011). In this definition, there is no distinction between quantitative and qualitative methods; both methodologies can be used to evaluate effect.

For use in the HHMP, we suggest the following elaboration of Krogstrup’s definition of effect evaluation: ***Effect evaluation** is the detection of causal relationships between an intervention and a performance measurement (results measurement related to production, fertility, health, or welfare). The evaluation includes a subsequent judgment of the practical usefulness of the effect estimates for implementation of corrective interventions.*

Further about evidence

The word *evidence* seemingly is not used nor perceived neither uniformly in everyday language nor in different scientific communities. Krogstrup (2011) indicates the divergence of opinions on scientific evidential support of evaluation of public policies merely by the title of her book: *The Fight about Evidence*

– *Performance Measurement, Effect Evaluation, and Evidence*. Elsass (1993) describes a similar controversy in the tension-field between psychology and human medicine (Elsass, 1993). Andersen (2004) and Kristensen (2008) also explore the paradigmatic relation between social science and the field of veterinary medicine. Historically, *the fight about evidence* may relate back to the following contrasts, as outlined by Krogstrup (2011):

- Ontological conflict between constructivism (humans ‘construct’ facts) and positivism (universal facts exist)
- Inductive (hypothesis creating) versus deductive (hypothesis testing) research
- Bottom-up (starting from working with real-world problems) versus top-down approach (starting from theory and experimental settings)
- No ranking of evidential strength versus an established hierarchy of evidence
- The aim of context-specific versus universal knowledge
- Divergence between the methodological approaches to research: qualitative versus quantitative

Essentially, this list of topics shows that a discussion about evidence is also a discussion of theories of science. The contrasts between qualitative and quantitative science are described briefly below in addition to some methodological approaches to uniting the forces and potentially providing more contextual, useful, and meaningful scientific evidence. The main emphasis will, however, be on quantitative analysis and experimental trials because these methods are especially suitable for demonstrating cause-and-effect relations in a herd management setting where there are numerous options for health performance measurements based on numerical clinical data, as described in human medicine by Habicht (2011).

Quantitative research methods and effect evaluation

The following text covers the essential parts of the relationships between study design and evidence of effect based on statistical methods.

In quantitative science, ‘scientific evidence’ is not the product of a single uniform methodology but is derived from the sum of results of a range of accepted methods that make associations between causes and effects plausible within each scientific field (you can compare evidence with a tower of bricks – each brick being one study using one method). This range of methods can be arranged in a ‘hierarchy of evidence’ (Figure 6) (Habicht, 2011). The hierarchy illustrates that the choice of methodology and study design dictates how well a statistically established association between two entities implies a ‘cause-and-effect’ relationship between the two. The hierarchy is not exhaustive. The scientifically best founded general evidence for a cause-and-effect relation between two entities requires the use of the methods in the top of the hierarchy. By ***general (or universal) evidence***, we mean evidence that is valid for extrapolation to the entire population due to representative data. In opposition to *general evidence*, we

define **local evidence** as evidence that is valid in highly specific contexts (e.g., herd-specific) and where the issue of representativity to populations beyond herd level is subordinate.

The validity of the evidence is a central issue at all levels of the hierarchy and corresponds to some extent to the *homogeneity* and *heterogeneity* issues addressed above. **Internal validity** refers to the extent to which the study designs and results of the effect evaluation are valid within the study population (e.g., cows included) and can be extrapolated to the *reference population* [e.g., all cows in the studied herd(s)] (Figure 2). **External validity** of a study refers to the extent to which the results can be generalized to the *general population* (e.g., to all Danish dairy herds). *Internal validity* is a prerequisite for *external validity* (Dohoo et al., 2003). Depending on the goals of a trial (universal or local evidence), the *study population* is defined through different eligibility criteria at cow or herd level. These eligibility criteria include both *inclusion criteria* that define which cows can be enrolled in the trial and *exclusion criteria* that exclude potential outliers or cows that would blur or bias the results. For instance, if you wished to obtain universal evidence of effects of treatment of genital diseases, a trial could be performed in a random sample of herds and cows in Denmark in a random sample of affected cows within a herd. This trial would be called a ‘multi-centre (multi-herd) trial’. To obtain local evidence at herd level, a trial could be performed in the concrete decision-taker’s herd(s) including all cows with genital diseases. In this latter case, the study population will be the same as the reference population and, consequently, would be as representative for the reference population (the herd) as possible.

In general, the representativeness of trial-based effects estimates is strongly influenced by the eligibility criteria. A narrow set of criteria could result in highly comparable intervention groups and thus increase the statistical power of the design and the statistical tests with a given sample size, but generalizability to the general population is reduced (Dohoo et al., 2003). Random selection of cows within herd and herds within country maximizes the generalizability of the results, thus increasing the likelihood of universal evidence.

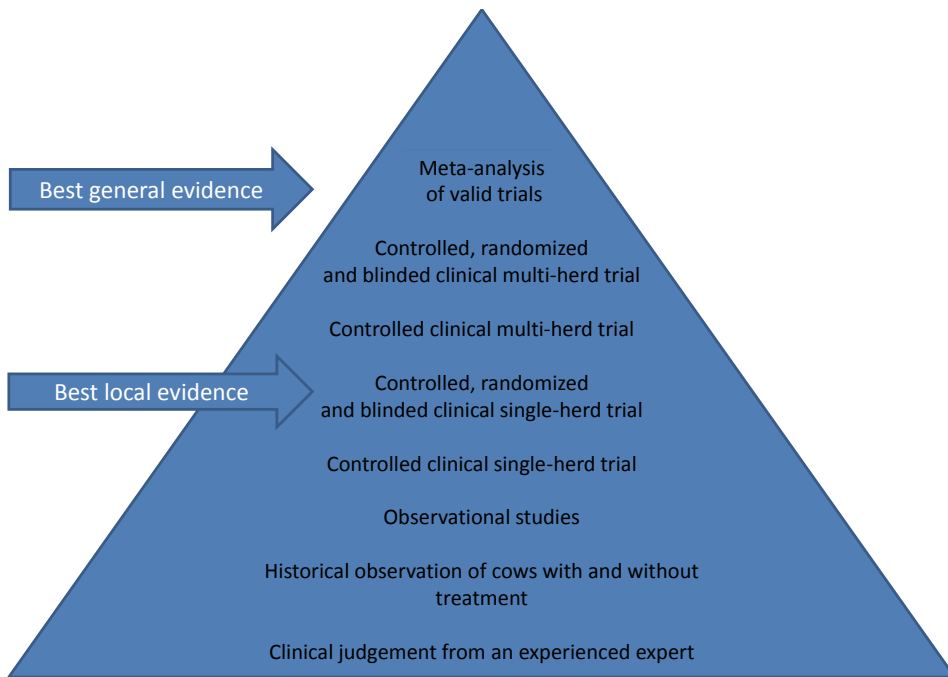


Figure 6. A suggested ‘*hierarchy of evidence*’. Methodologies at the top are superior to establish general evidence for cause-and-effect relationships between interventions compared to the methodologies at the bottom. For instance, a comparison of the effects of two medical treatments of genital disease on a chosen outcome could be evaluated with the entire range of methods. Inspired by and modified from Habicht (2011).

The sequential clinical judgements described in example 1 illustrate the lowest level of evidence shown in Figure 3. An issue related to expert knowledge is that it may often be ‘tacit knowledge’ (you know but you don’t know why). That is, the tacit knowledge can be extremely useful for problem-solving in the specific context but, by definition, it is difficult for the expert to share the knowledge with others (Heiberg Engel, 2008). The best *general evidence* for *cause-and-effect* can be found in meta-analyses with the homogeneous results of multiple externally valid multi-herd *controlled, randomized, and blinded* trials. This high level of evidence is difficult, expensive, or perhaps impossible to obtain in most fields of veterinary medicine at present (Kastelic, 2006). Also, in the HHMP context, general evidence of effect could be irrelevant because such studies seek to estimate average general effects. If a given cow or herd is not ‘average’ with regard to multiple prognostic factors relevant for disease and treatment effect of genital diseases, then decisions and recommendations based on general average estimates could have serious negative effects. Therefore, we claim local evidence is warranted to support ‘evidence-based decision making’ in specific herd contexts and that the trial approach could contribute with such valuable local

evidence concerning for instance difference in treatment effectiveness of metritis, as shown by Lastein (2012). Within these herd-specific contexts, the identification of the aim of the ‘support’ is still central (choice of the conceptual design): Do we want exploratory evidence (biological effect in case of treatment) or pragmatic evidence (effect in case of ‘intention to treat’)?

Qualitative research methods and effect evaluation

In the following section, we give an ultra-short description of a few fundamental ideas in qualitative research, often used in humanities and in social science. Special emphasis in this section will be on conduct, analysis, and interpretation of so-called qualitative research interviews and is based on the descriptions given in an article and two textbooks on qualitative research (Aagaard-Hansen, 2007; Flick, 2002; Kvale, 1994). Where appropriate, we have associated the qualitative description with some quantitative terms used above to enhance the possibility of the reader (assumed to be primarily educated within natural sciences) to acknowledge the similarities and differences between the two methodologies. We acknowledge that there is much variation in aims and methodology that we do not cover in the following short description.

The aim of qualitative research can be to promote understanding of a given phenomenon; **to obtain an in-depth and contextual understanding of the associations between human’s life-worlds, and human actions around this phenomenon (the research question)**. Emphasis on context is of utmost importance. Often, an inductive approach of analysis is applied. That is, a theory is induced from empirical data [e.g., you explore the complexity in human perceptions (data) to create an understanding of a coherent structure around the phenomenon, often called a ‘grounded’ approach or theory]. Data in qualitative research are empirical material like interviews, dialogue, discussions, observations, and pictures. This inductive approach should be seen in contrast to the often deductive approach often used in quantitative research where you know which variables to specify in a quantitative model to get the best estimates of effect before you collect and analyse your data. However, induction and deduction can be used in all scientific fields, and can be used in the same analysis at different stages of the process. In veterinary practice, a practical example of induction and deduction is the traditional ‘clinical decision process’ (Kristensen, 2008). The clinical examination is induction. That is, you observe a clinical manifestation and build a theory (the diagnosis). The diagnosis is followed by a deductive process, because you formulate a prognosis based on the diagnosis (a theory) and the potential interventions (or treatments) given by the diagnosis.

The diagram (Figure 7) below illustrates the differences between different approaches to the scientific study of a phenomenon (deductive and linear versus inductive and circular or ‘grounded’) (Flick, 2002).

Qualitative research can follow both linear and cyclic processes. We will focus on the inductive and circular methodology, which contrasts with the linear methodology often used in quantitative research.

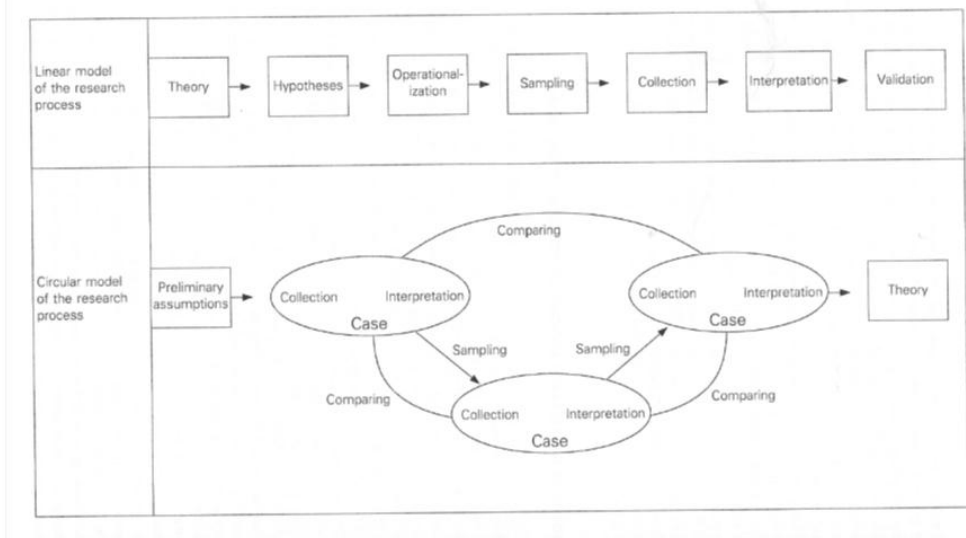


Figure 7. A linear research approach beginning with a theory that starts with deduction (often used in quantitative research) and a circular research process ending with a theory that starts with induction (often used in qualitative research). From Flick (2002).

The sampling process of study units (e.g., humans) in qualitative research is based on some of the same principles as in quantitative research (e.g., purpose sampling, convenience sampling, random sampling, etc.) (Onwuegbuzie and Leech, 2007). In qualitative research, iterative sampling is fully accepted (e.g., more interviews are conducted until consistency and coherence of the phenomenon ('saturation') is accomplished). The choice of sampling frame will influence the qualitative results and their generalizability and validity similar to the experience in quantitative research. However, in the case of qualitative research, these issues are somewhat differently evaluated, as described below. For instance, the number of participants in an interview study will typically be smaller than in any quantitative study. This is because the validity of an interview study also relies on the interaction time (depth and duration of interview) and not only on number and sampling procedure. We will elaborate further on the differences of 'tradition of sampling and inferences' within the disciplines. It is important to note that qualitative sampling (often *not* random and *not* representative) does not allow for any valid numeric inferences such as proportion or averages and that the analytical circular process involving the interaction with the researcher destroys the independence of the observations required for numeric inferences. Accordingly, we find it important to notice that random and representative sampling does not strive to include 'extreme observations'

(outliers); thus, the inferences represent only the average and related measurements and *not* thorough observations of the entire spectrum within a phenomenon.

In the case of interviews, these can be more or less structured (potentially following a written interview guide). Optimally, the interviewer will ‘influence’ the interviewed (or the observed) and vice-versa to a minimal degree. Often, interviews are recorded and subsequently transcribed (typed) full length to facilitate analysis. The practical analytical methods used can be described as meaning condensation, meaning categorization, narrative structuring, interpretation, and combined ‘ad hoc’ methodology to create meaning of the empirical material (Kvale, 1994). The analytic process is often cyclic as described above. Also, the sampling can be cyclic, that is if there are areas of the research context that require additional support during data collection then additional interviewees can be included. The results (the theory) could be presented as a diagram or figure describing relations in the observed data, and the theory are supported by quotes and descriptions of meaning.

Issues as generalizability, reliability, and validity are also relevant in qualitative research (Kvale, 1994). However, the terms are used in a different way because qualitative research often builds on the perception of ‘knowledge and evidence’ as a social construct (e.g., knowledge and evidence are formed by the analysis and discussions of the observed phenomena) and not as a manifestation of ‘an objective and universal truth’ because of a rejection of the positivistic approach that is dominant in quantitative research. **Generalizability** can be judged either as naturalistic (experience-based), statistical (random selection and quantification), or analytical (thorough judgement of the extent of the value of the results in other contexts, related to external validity in quantitative science). **Reliability** refers to the extent of the researcher’s subjective influences in the interview and transcription phases (e.g., the fewer leading questions and predefined conceptions, the higher the reliability; related to quantitative bias). In qualitative research, **validity** can be defined as: “to which extent do our observations reflect the phenomenon or variable that we are interested in” (related to internal validity in quantitative science). According to Kvale (1994), validity depends on the “quality of the craftsmanship” and should be ensured and evaluated simultaneously through the multiple stages of a qualitative research project using interviews (Kvale, 1994). Qualitative studies are relatively rare in the field of veterinary science, but there are HHMP-related examples of applications within the field of mastitis treatment and management, metritis treatment, and calf mortality (Jansen et al., 2010; Lastein et al., 2009; Nielsen et al., 2008; Vaarst and Sørensen, 2009; Vaarst et al., 2002; Vaarst et al., 2003). Recently, the Scandinavian veterinary research community was encouraged to use these methods to supplement their national database studies (Hansen et al., 2011).

The research community appears somewhat divided in its views on quantitative and qualitative research methods. However, a ***mixed methods research (MMR) approach*** is suggested to combine the qualitative approaches with quantitative methods in either the same or parallel studies of the same phenomenon. This MMR approach corresponds to the principles of supplementary validation ('ask again'), triangulation ('measure the cows and both observe and ask the farmer and the veterinarian'), and knowledge generation (an iterative change between hypothesis creating and hypothesis testing in a HHMP) (Kristensen, 2008; Kristensen et al., 2008b). The strength of the MMR approach is that the same problem is addressed from different angles, elaborating on both human involvement and statistical inferences. Several areas related to HHMP have been studied using MMR principles, e.g., calf mortality, data collection on metritis, and valuation of the HHMP (Andersen, 2004; Kristensen, 2008; Kristensen and Enevoldsen, 2008; Kristensen et al., 2008b).

Effect evaluation by means of qualitative research methods

Qualitative research methods can be used to evaluate the perception of effect of a given phenomenon among people who experience it. For instance, the value (or effect) of the Danish HHMP has been evaluated by these methods (Kristensen and Enevoldsen, 2008). Also, smaller parts of a larger context (e.g., the effect of examination procedures on the quality of data) have been evaluated this way. By means of examples derived from the metritis context, Lastein et al. (2009) showed that some veterinarians used clinical scores very differently and inconsistently and that scores sometimes are adjusted according to the 'situation' and not according to the clinical sign on the cow. When such uses of scores have been identified, it is clear that the intended application of the scores as 'objective' welfare measurements by means of frequencies and associated variance will fail. That is, a qualitative piece of evidence like this does provide evaluation about the effectiveness of at least one component in a trial process.

Evidence-based decision making

Above we have sketched how different data sources become information by means of various approaches to an effect evaluation. The final step in the evaluation process is the evaluation of the usefulness for concrete actions in a specific herd. Even if a trial was conducted in a herd, circumstances might have changed after completion of the study so that the quantitative or qualitative evidence was no longer valid in the new context. Consequently, effect evaluation will be an on-going (dynamic) process in the dairy herd context. It should basically be an integrated part of planning. This notion leads to the concept of ***supported decision making*** in an iterative pattern. By *supported decision making*, we mean that data derived from the production and organized into information in different ways can help farmers and herd consultants (including veterinarians) make informed (perhaps improved) decisions about why, when, and whether to

take corrective actions. As such, the information plays the role of *evidence* that the decisions will affect the production in the needed direction (if the evidence is valid!).

Evidence-based decisions can be divided into decisions related to either practice or policy (Krogstrup, 2011) and to either local or general decisions. If you recall the continuum of the different conceptual trial designs from the explanatory to the pragmatic trial, the differences between efficacy and effectiveness and the difference in testing a biological causal relation and testing ‘an intention to follow a decision’, you will also find these differences integrated. In our dairy herd management context, we see also the division as multi-dimensional. *Evidence-based (veterinary) practice* is the situation in which decisions are concerned with the individual cow (e.g., to treat or not to treat puerperal metritis and how to interpret a given diagnostic test), and *local or general evidence-based policy* is the situations in which the decisions are concerned with either the herd (e.g., if and which standard treatment protocol to use) or the society/dairy industry (e.g., to implement or not implement the HHMP as a compulsory national program). Each individual decision being reasoned by short or long term effect measurement must be placed in the continuum between practice and policy. In addition, for each decision, the evidence underlying the decisions must be based on appropriate scientific support of the methodology that is best suited to cover the decisions in question.

Krogstrup (2011) elaborates on the use of evidence as background for decisions by the distinction between 1) *deterministic evidence*, 2) *probabilistic evidence*, 3) *day-to-day evidence*, and 4) *other factors than evidence* (such as personal preference or opinions, and legal implications). **Deterministic evidence** is characterized by meeting the counterfactual principles, which means relying on an ability to control all influential factors (e.g., ‘all other things equal’ as in a ‘true’ pure causal–effect relation). Thus, deterministic evidence is the most extreme of general evidence. Krogstrup (2011) states that no convincing example of deterministic evidence has been found in situations where ‘the subjective human behaviour’ could influence the relation (Krogstrup, 2011) (as we saw in qualitative research: ‘evidence is regarded as a social construct’). In our context, this means that only decisions of purely biological and/or technical character (so-called ‘tame problems’) can be based on deterministic evidence. It also means that no decisions involving any aspects of human influence (veterinarian, farmer, banker, spouse, etc., on decisions related to, for example, treatment threshold and expenses for HHMP) in any procedures in the herd rightfully can be based on deterministic evidence. Instead, they can be considered ‘wild problems’. **Probabilistic evidence** is based on the contextual relations between entities, and interactions between all variables in the ‘equation’ are possible. Uncertainty of the evidential basis is a fundamental issue when using probabilistic evidence. Examples of decisions based on probabilistic evidence could be quantitative evidence of average effect of metritis treatment or decisions based on qualitative evidence supporting the

value of implementing postpartum examinations. **‘Day-to-day’ evidence** is based on continued observations and ‘non-scientific’ evaluation performed by the HHMP stakeholders (e.g., decisions based on ‘perception’ or experience of the success of a given treatment protocol [similar to local (herd) expert knowledge that may be tacit or non-transferable to others]).

Evidence-based veterinary practice and policy in the herd health management context

The descriptions above concerning evidence and decision making lead us to explore the description of evidence-based (human) medicine as described by Sackett (1996):

“Evidence-based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence-based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research. By individual clinical expertise, we mean the proficiency and judgment that individual clinicians acquire through clinical experience and clinical practice. Increased expertise is reflected in many ways, but especially in more effective and efficient diagnosis and in the more thoughtful identification and compassionate use of individual patients’ predicaments, rights, and preferences in making clinical decisions about their care. By best available external clinical evidence, we mean clinically relevant research, often from the basic sciences of medicine, but especially from patient-centred clinical research into the accuracy and precision of diagnostic tests (including the clinical examination), the power of prognostic markers, and the efficacy and safety of therapeutic, rehabilitative, and preventive regimens. External clinical evidence both invalidates previously accepted diagnostic tests and treatments and replaces them with new ones that are more powerful, more accurate, more efficacious, and safer. Good doctors use both individual clinical expertise and the best available external evidence, and neither alone is enough. Without clinical expertise, practice risks becoming tyrannized by evidence, for even excellent external evidence may be inapplicable to or inappropriate for an individual patient. Without current best evidence, practice risks becoming rapidly out of date, to the detriment of patients.”(Sackett et al., 1996).

In the text above we have underlined, for us, central areas of the quote to emphasize the importance of combining context with scientific evidence to obtain the goal of making the best decisions for the cow or the farmer. Accepting this definition and translating to our dairy and veterinary context (replacing doctor with veterinarian or herd consultant, and patient with cow or herd), this definition implies that evidence-based decisions and actions within the HHMP should be based on a combination of available best scientific evidence of different methodologies (qualitative and quantitative), clinical expertise and experience (including updated education), and a combination of veterinarian and farmer personal preferences based on personal ideology and meaning (‘deeper personal structures’; Figure 5). We use the term ‘practice’ instead of ‘medicine’ to emphasize the importance of the actions that follow a decision based on evidence. In our context, decisions must also adhere to social and cultural norms and legal implications regarding, for

instance, animal welfare and drug usage. We have found little discussion of the use of local herd-specific retrospective data analysis in literature on ‘evidence based veterinary medicine or practice’, although clearly exemplified in the dairy context several times (Enevoldsen, 2006; Nir, 2008). Inspired by these sources and by Schmidt (2007), we propose the following framework for evidence-based work in the HHMP context (Figure 8).

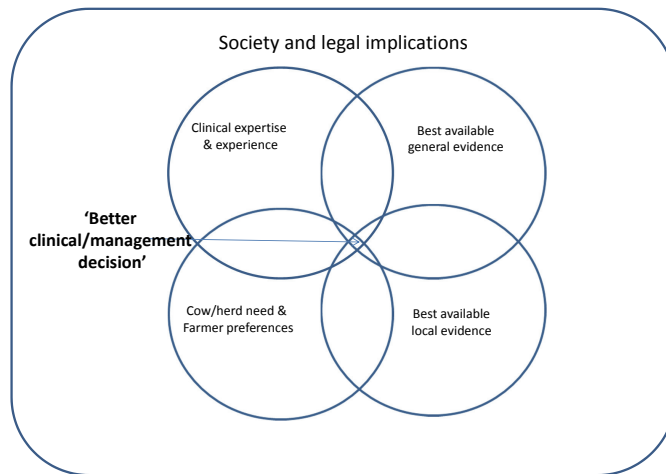


Figure 8. A framework for decision making by the veterinarians in the HHMP context: Use the combination of both general and local scientific evidence (qualitative and quantitative methodologies), personal qualifications, and cow and herd contextual needs and preferences to reach the best decision for solving clinical or management-related problems. Modified after Schmidt (2007).

Problem reduction – can wild problems be tamed?

We (the farmer and the veterinarian) may ask this question: “How do we most effectively treat bovine genital disease”? This question implies that there is a universal truth concerning a biological causal relation between a treatment and a given outcome. The description above concerning effect evaluation also indicates that it might be what Krogstrup (2011) calls a ‘wild problem’ to provide an answer to the question. A wild problem is characterized by being difficult to define and with no objective best solution (Krogstrup, 2011). Our question above is a wild problem because it involves multiple whole farming systems with humans, cows, and data in different contexts with feedback mechanisms between at least veterinarians and authorities. If we decide to continue a search for an answer, we need to reduce the *wild problem* to a *tame problem*, that is, to something tractable. We could start by reducing ambition and asking this question instead: “How do we most effectively treat bovine genital disease in this particular herd?” We now request *local evidence* and have reduced the contextual influence to a less complicated problem involving only a few humans concerned with cows and data in a more controllable context. If humans in the

herd context are organized, systematic, and able to formulate their own preferences and procedures, the problem could potentially be further reduced to a 'tame' problem (e.g., easy to define, of technical character). By a process of problem reduction and a 'bottom-up approach', effect evaluation of therapeutic interventions in a herd context can meaningfully be evaluated in trial settings in 'the real world'. This approach is explained as exemplified in detail in relation to bovine genital disease in Denmark (Lastein, 2012). Schwabe et al. (1977) suggest implementing the Evolutionary Operation principle (EvOp) (Rieman and Aalund, 1975). EvOp is an iterative process in which trials are conducted to improve the production process while the production continues. That is, the risk of major detrimental effects of the interventions should be minimal. The experimental interventions might be defined from experiences in the actual production system or they might be inspired from external evidence. This iterative study design should gradually improve best practice. Such an EvOp design could potentially be used to optimize the use of medical interventions or management interventions in HHMPs.

In Denmark, a HHMP with routinely systematic clinical examinations of cows in a predefined risk period and associated tools for monitoring and evaluation of farm functions (e.g., measuring of disease incidence, lactation curves) was developed and implemented from the late 1990s [read a thorough description and history by Krogh (2012)]. The legislation concerning the mandatory part of this program includes documentation of requirement for 'effect evaluation of initiated medical treatments' (<https://www.retsinformation.dk/Forms/R0710.aspx?id=132648&exp=1>) [in Danish]. Additionally, increased documentation of the effect of procedures and interventions on the farm are central for following the development with fever, but larger herds, fewer people to tend them, decreasing milk prices, etc. Also, the surrounding society is concerned about microbial antibiotic resistance due to the use of therapeutic antibiotics in the farming industries, one reason evaluation of effect can become an issue of public debate in the future.

However, the monitoring and evaluation principles developed so far mainly relate to process evaluation (e.g., number of inseminations per pregnancy), results evaluation (e.g., litres of milk per cow per day) over time, identification of outliers (e.g., identification of single disease cases), trends (e.g., changes of fat/protein ratio) (Krogh, 2012), and simulation of the economic effects of changes under Danish conditions (www.simherd.com). So far, the evaluation of the effects of the preventive or therapeutic interventions initiated by the herd manager and his herd health advisors, such as veterinarians or feed consultants, has not been addressed with a systematic approach. If problem reduction could be successful (tame the people!), then a quantitative trial approach to effect evaluation combined with a qualitative approach to design and results evaluation could be promising. The herd manager and the veterinarian in charge of an

EvOp process certainly could claim that local evidence-based decisions to optimize, for example, the medical interventions, were documented convincingly. Such a claim could be valuable in the future if debates concerning antibiotic treatment in agricultural production and bacterial resistance are frequent. If EvOp-type studies of similar issues (e.g., medical treatments with a certain drug) were conducted in multiple herds, general evidence might emerge based on meta-analytical principles.

5. Barriers to implementation of evidence-based service and clinical field trials in veterinary practice

In the following section, we address potential barriers to a successful implementation of clinical field trials in veterinary practice. Several issues will be dealt with to discuss the potential of the trial approach. First, issues on the requirements for data and assessment of data quality are addressed. Second, issues related to human impact on data quality and success of implementation are discussed.

Data quality

In a clinical trial context related to metritis, data collection on the signs of metritis depends on human clinical skill. Recordings of clinical manifestations can be difficult to calibrate (Baadsgaard and Jorgensen, 2003; Krogh et al., 2011). However, studies have shown that efforts to define and calibrate clinical assessments in areas such as udder health, body condition scoring, and vaginal discharge scoring (Klaas et al., 2004; Kristensen et al., 2006; Lastein and Enevoldsen, 2010) can result in clinical data of a quality that most certainly can be used within herd (local validity), but use between herds is problematic (general validity low). This issue is elaborated upon and discussed intensively in a HHMP context (Lastein et al., 2009).

Again, we recall the continuum of the different conceptual designs, from the explanatory to the pragmatic (Thorpe et al, 2009). If results of explanatory trials on effect are to be used to evaluate effect, then results from both PP and ITT analysis trials should be conducted and be consistent. For such similarity of results to be found, a high level of adherence is required (because non-adherence will dilute any effect in ITT results, as described above). If the results from the two analyses differ (e.g., PP analysis shows an effect and ITT analysis does not), then the results of explanatory trials used to evaluate effect are considered inconclusive. In a pragmatic trial, only ITT analysis is performed, and the level of adherence will influence the result in the same way as real-world circumstances influence the ITT-data: Some cows receive the intended treatment and others do not. However, if the differences in effect between the 'intention compared' is large enough, the results will demonstrate which one to prefer. A high level of adherence might not be achievable in the 'trials in real world HHMP' despite intensive instruction, calibration, and motivation among the participants (Lastein, 2012). This 'relatively low level of data quality' will imply that only relatively pragmatic trials within

each small unit (herd or veterinary practice) and subsequent coherent ITT analysis and interpretation of the results—evaluation of the ITT—are valid under HHMP conditions. If the end-users accept these premises of the results, i.e., ‘what is the effect given the intention of treating according to a protocol?’, then implementation of trial results from local effect evaluation can be successful.

Human involvement

We addressed acceptance of the ITT principles in the section above. Such acceptance can be obtained only through understanding. The scientific veterinary literature seems to almost neglect these aspects on evidential issues. Thus, for the trial approach to be implemented in a broad spectrum of dairy herds, educational efforts are warranted to ensure a common understanding of the advantages of local pragmatic estimates of effects.

Given that trials are implemented, will the evaluation of effect estimates change the management decisions? This is not a given consequence. Maybe the participants do not believe in or accept the results despite their own involvement, or circumstances have changed during the trial process (e.g., change to organic farming). That is, the qualitative prerequisites for the trial had changed fundamentally and the (internally valid) inferences from the effect evaluation are no longer meaningful in the actual context.

Another issue is that some veterinarians (and farmers) from the start might not want to invest personal effort in obtaining understanding and education and updating their professionalism or in investing business resources in developing ‘tailor-made trial solutions to suit their own context’. We found large variation among veterinarians working within the Danish HHMP in motivation to evaluate effect, and evidence of reluctance towards issues such as systematic data collection and statistical analysis (Lastein et al., 2009). Therefore, we find it likely that a group of veterinarians would reject participation in the proposed ‘systematic effect evaluation scheme’ to support decision making. They probably would prefer to accept a ‘day-to-day evaluation’ and the ‘general evidence’ as it seems to be taught in the usually positivist-oriented veterinary curriculum. That is, they assume that their recommendations and actions are based on (perhaps tacit) expert knowledge that is applicable in general. Therefore, they may not want or be able to appreciate that you could challenge whether general evidence in ‘complex farming systems involving humans’ exists. A similar situation is described among surgeons in the human clinic, and the following quote summarizes very well the veterinary situation as we see it (replace surgeons with veterinarians):

“Surgeons need to make changes to the health care they provide all the time as new clinical evidence emerges. If the small change results in worse outcomes in the clinical setting, it can be simply reversed; however, this requires the surgical unit to evaluate outcomes related to that change. Small, frequent

changes are more likely to avoid failure of the system than large dramatic changes. Continuous information delivery to surgeons is required to enable them to keep up to date and to improve knowledge transfer. There is no progress with no change and no change with those practitioners who cannot change.” (Daves and Marko, 2007).

Even for veterinarians who are willing and motivated to follow the principles of ‘evidence-based practice’, barriers are evident. The journey from ‘craftsmanship’ to ‘evidence-based practice’ is long and requires personal will, motivation, and perhaps some change in attitude (Hansen and Mikkelsen, 2012). Furthermore, veterinarians in practice (or in general) have difficulties in deciphering or ‘dissecting’ the available scientific literature, which seems to be more and more technical. That is, they lack the scientific competencies to fully understand and acknowledge the methodological differences and the importance of these when it comes to critically judging the internal and external validity of the quantitative studies. For qualitative studies, this problem probably is even bigger because these methods are not standard in the veterinary curricula we are acquainted with.

The dairy industry in the year 2012 calls for ‘leadership’ or managing as other similar-sized industries or organizations. In addition to managing multiple employees and/or industrial (mechanical) processes, managing of a dairy herd also relates to health and production of living organisms. Managing can be described as the act of getting people (e.g., herd owner, employed herdsman, veterinarians, etc.) together to accomplish desired goals and objectives with available resources (e.g., the cows, facilities, land, etc.) efficiently and effectively. The term ‘herd health management’ thus describes methods and actions taken to measure, monitor, evaluate, and control the on-going farm functions and performance over time and intervene if goals related to animal health and production are not reached. Herd health has also been described as a ‘self-generating and self-regulating’ complex organizational ecosystem that requires context-specific feedback mechanisms and potential interventions (management) performed by humans (herd manager and advisors) to meet predefined herd-specific targets and limits (Krogh, 2012). This description implies that the HHMP is a highly complex and very contextual work field involving humans, cows, and data, of which humans most likely are the hardest component to control.

6. Suggestions for implementation of randomized controlled clinical field trials in herd health management

We propose that establishment of some unit for a trial design and analytical support is needed for trials to be implemented in the HHMP. This unit should support the development of competences within the field of evidence-based veterinary medicine and practice among veterinarians in dairy practice. Because we now

have Danish cattle practice units with up to around 90 veterinarians serving 1200 herds (www.dyr-laegerogko.dk, accessed 10.09.2012), some practice units have sufficient volume for such a subunit. The supportive unit must have competences in epidemiology, qualitative research, clinical science, education, and management of human resources to support the veterinarians in practice for on-going professional (and personal) development. Incentives for the establishment of such a supportive unit could be governmental regulations or motivation within the management of private veterinary practices. The cooperative Israeli veterinary service has established such a unit (<http://www.hachaklait.org.il/english/haklait-english.pdf>). Inspired by initiatives within the field of human nursing (Hansen and Mikkelsen, 2012), we propose the following areas of general support given by the unit:

1. Introduction to evidence-based medicine and practice to understand the definitions of experience-based and evidence-based clinical decision making
2. Definition of research question and search of the scientific literature
3. Qualitative research: methodology and critical judgement
4. Basic statistical and epidemiological knowledge
5. Quantitative research: methodology and critical judgement of evidence
6. Evidence-based decision support: protocols and implication of the applied practice
7. Development and implementation of clinical field trials in the HHMP

We suggest an organizational diagram for a dynamic generation of best available local evidence of effect. We propose a framework for implementing the trial approach for effect evaluation in the HHMP (Figure 9). The diagram shows six phases in a systematic iterative cycle for conducting clinical field trials. The phases are 1) identification and reduction of the problem; 2) trial design; 3) starting phase; 4) trial conduct with data collection; 4) quantitative analysis; and finally, 6) qualitative effect evaluation and decision making.

We have used the term herd health management and the related *herd health management program* (HHMP) if a specific set of predefined activities (centred on regularly planned herd visits and raw data recording) are implemented with the aim of monitoring, evaluating, and controlling dairy health and production. The Danish approach as briefly outlined in the introduction is an example of a HHMP. The Danish HHMP is one approach to such systematic activities. Alongside the Danish HHMP, a set of management tools for performance measurement and monitoring are developed (Krogh, 2012), which can be helpful in the problem-reduction phase.

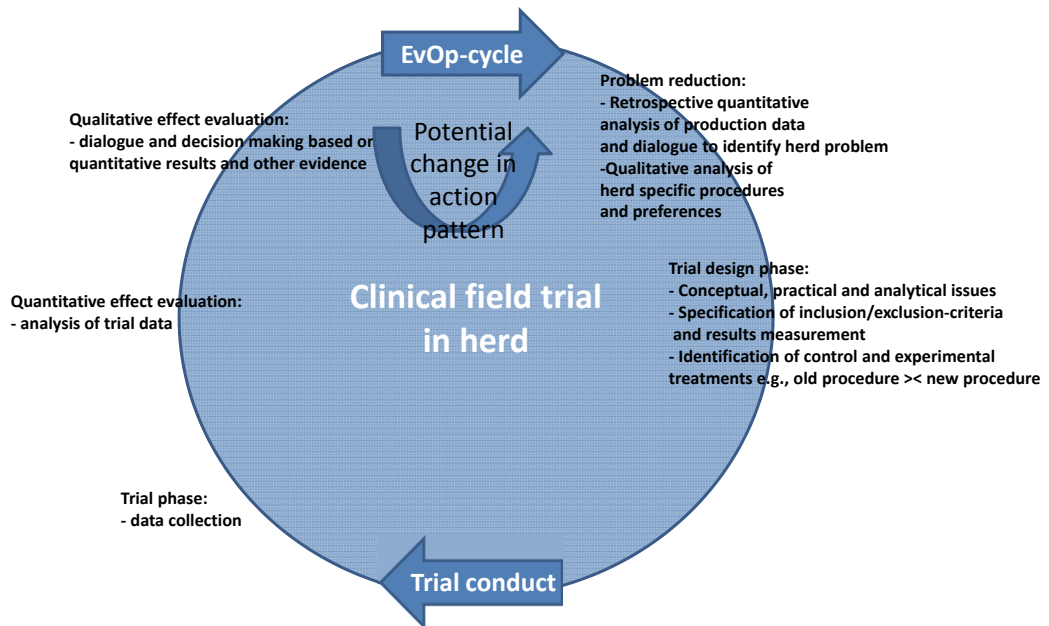


Figure 9. The proposed mixed approach for systematic effect evaluation in a dairy herd health management program (HHMP) involves both herd-specific qualitative and quantitative elements. A systematic iterative cycle of a randomized controlled clinical field trial conducted in six phases including problem reduction, trial design, and starting phase, trial conduct with data collection, and quantitative analysis. The final evaluation and decision-making phases are based on the trial results and any other relevant evidence available. Subsequent changes in action patterns in the herd-specific procedures could arguably be a result of evidence-based effect evaluation in the HHMP.

In this tutorial, we have described and explored the theoretical possibilities for integrating systematic effect evaluation into the decision-support toolbox of HHMPs. To do so, we have defined ‘effect evaluation and evidence of effect in the HHMP context’ and described an approach that can be used to evaluate relevant effects. We have suggested solutions to reduce the complex problems of the herd context to simpler, explicitly described problems within the herd context.

In summary, we have:

1. Described and defined effect evaluation, the concepts of evidence, and decision making in the HHMP context.
2. Described trial theory in the context of the HHMP.
3. Proposed a coherent approach to systematic iterative effect evaluation in the HHMP.

References

Consort-statement.org, 2011. CONSORT principles. www.consort-statement.org. Assessed 18-09-2012

Aagaard-Hansen, J., 2007. The challenges of cross-disciplinary research. *Soc Epi* 21(4), 425-438.

Andersen, H.J., 2004. Rådgivning, bevægelse mellem data og dialog. Ph.D.Thesis. Mejeriforeningen, Århus.

Baadsgaard, N.P., Jorgensen, E., 2003. A Bayesian approach to the accuracy of clinical observations. *Prev Vet Med* 59(4), 189-206.

Bender, R., Lange, S., 2001. Adjusting for multiple testing--when and how? *J Clin Epidemiol* 54, 343-349.

Bjerkset, O., Larsen, S., Reiertsen, O., 1997. Evaluation of enoxaparin given before and after operation to prevent venous tromboembolism during digestive surgery: Play the winner designed study. *W J Sur* 21, 584-589.

Christensen, E., 2007. Methodology of superiority vs. equivalence trials and non-inferiority trials. *Journal of Hepatology* 46(5), 947-954.

Christensen, E., 2008. Metoder ved ækvivalens-og non-inferioritetsundersøgelser (in Danish). *Ugeskrift for læger* 170(38), 2977-2979.

Daves, L., Marko, L., 2007. Knowledge transfer in surgery; skills, process and evaluation. *Ann R Coll Surg Eng* 89, 749-753.

Dohoo, I.R., Martin, W., Stryhn, H., 2003. *Veterinary epidemiologic research*. AVC Inc., Charlottetown, Prince Edward Island, Canada. ISBN 0 919013 41 4

Elsass, P., 1993. *Sundhedspsykologi [in Danish]*. Gyldendal, Copenhagen, Denmark. ISBN 87 00 15468 7

EMA, 2001. Choice of control group in clinical trials. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002925.pdf . Assessed 18-6-2012.

EMA, 2012. Guideline on statistical principles for clinical trials for veterinary medical products (pharmaceuticals). http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/01/WC500120834.pdf . Assessed 18-6-2012.

Enevoldsen, C., 2006. Epidemiological tools for herd diagnosis. *Proc. WBC XXIV, Nice*. 376-383.

Flick, U., 2002. *An introduction to qualitative research*. SAGE Publications Ltd, London, Great Britain, London. ISBN 0 7619 7435 0

Fourichon, C., Seegers, H., Bareille, N., Beaudeau, F., 1999. Effects of disease on milk production in the dairy cow: a review. *Preventive Veterinary Medicine* 41(1), 1-35.

Gartlehner, G., Hansen, R.A., Nissman, D., Lohr, K.N., Carey, T.S., 2006. A simple and valid tool distinguished efficacy from effectiveness studies. *Journal of Clinical Epidemiology* 59, 1040-1048.

Gøtzsche, P, 2012. Noninferioritets- og ækvivalensforsøg: erfaringer og forbehold - sekundærpublikation. *Ugeskrift for læger* [in Danish] 168(37).

Habicht, A., 2011. *Vurder selv evidens* (in Danish). Munksgaard, Denmark. ISBN: 978 87 628 1111 9

Hansen, B., Schei, V., Greve, A., 2011. When counting cattle is not enough: multiple perspectives in agricultural and veterinary research. *Acta Veterinaria Scandinavica* 53 (1).

Hansen, S.R. and Mikkelsen, M.R., 2012. Videnskabelige kompetancer til udvikling af evidensbaseret sygepleje [in Danish]. *Sygeplejesken* 7, 69-74.

Haynes, B., 1999. Can it work? Does it work? Is it worth it? *BMJ* 319, 652-653.

Heiberg Engel, P.J., 2008. Tacit knowledge and visual expertise in medical diagnostic reasoning: Implications for medical education. *Med Teach* 30(7), 184-188.

Hollis, S., Campbell, F., 1999. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ* 319, 670-674.

Hostens, M., Ehrlich, J., Van Ranst, B., Opsomer, G., 2012. On-farm evaluation of the effect of metabolic diseases on the shape of the lactation curve in dairy cows through the MilkBot lactation model. *J. Dairy Sci.* 95, 2988-3007.

Jansen, J., Steuten, C.D.M., Renes, R.J., Aarts, N., Lam, T.J.G.M., 2010. Debunking the myth of the hard-to-reach farmer: Effective communication on udder health. *J. Dairy Sci.* 93(3), 1296-1306.

Kastelic, J.P., 2006. Critical evaluation of scientific articles and other sources of information: An introduction to evidence-based veterinary medicine. *Theriogenology* 66(3), 534-542.

Klaas, I.C., Enevoldsen, C., Vaarst, M., Houe, H., 2004. Systematic Clinical Examinations for Identification of Latent Udder Health Types in Danish Dairy Herds. *J. Dairy Sci.* 87(5), 1217-1228.

Kristensen, E., Dueholm, L., Vink, D., Andersen, J.E., Jakobsen, E.B., Illum-Nielsen, S., Petersen, F.A., Enevoldsen, C., 2006. Within- and across-person uniformity of body condition scoring in Danish Holstein cattle. *J. Dairy Sci.* 89(9), 3721-3728.

Kristensen, E., Ostergaard, S., Krogh, M.A., Enevoldsen, C., 2008a. Technical indicators of financial performance in the dairy herd. *J. Dairy Sci.* 91(2), 620-631.

Kristensen, E.L., 2008. Valuation of dairy herd health management. Ph.D. Thesis. Faculty of Life Sciences, University of Copenhagen, Denmark.

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

Kristensen, E., Nielsen, D., Jensen, L., Vaarst, M., Enevoldsen, C., 2008b. A mixed methods inquiry into the validity of data. *Acta Veterinaria Scandinavica* 50(1), 30.

Krogh, M.A., 2012. Management of data for herd health performance measurements in the dairy herd. Ph.D. Thesis. Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.

- Krogh, M.A., Toft, N., 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.
- Krogstrup, H., 2011. Kampen om evidens; Resultatmåling, effektevaluering og evidens [in Danish]. Hans Reitzels Forlag, Copenhagen, Denmark. ISBN 9788741255163
- Kvale, S., 1994. Interview - en introduktion til det kvalitative forskningsinterview. Hans Reitzels Forlag, [in Danish]. Copenhagen, Denmark. ISBN 87 412 2816 2
- Lastein, D.B., Enevoldsen, C., 2010. Visual assessment of within and between observers' agreement on vaginal discharge scores in a cattle practice context. *Proc. WBC XXVI*, Santiago, Chile, 2010.
- Lastein, D.B., 2012. Herd-specific randomized trial - an approach for effect evaluation in a dairy herd health management program. Ph.D. Thesis. Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.
- Lastein, D.B., Vaarst, M., 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 Veterinaria Scandinavica* 51(1), 36.
- Lee, Y.J., Ellenberg, J.H., Deborah, Hirtz, G., Nelson, K.H., 1991. Analysis of clinical trials by treatment actually received: Is it really an option? *Statist. Med.* 10, 1595-1605.
- Nielsen, D., Jensen, L., Jespersen, R., Enevoldsen, C., 2008. Use and abuse of a uterine scoring system. *Hungarian veterinary journal. WBC XXV. Hungarian veterinary journal. Suppl. 2.* 130
- Nir, O., 2008. The multifactorial approach to fertility problems in dairy herds. *WBC XXV. Hungarian veterinary journal. Suppl 1,* 77-81.
- Onwuegbuzie, A.J., Leech, N., 2007. A call for qualitative power analysis. *Qual & Quan* 41, 105-121.
- Proschan, M.A., Waclawiw, M.A., 2000. Practical guidelines for multiplicity adjustment in clinical trials. *Control Clin Trials* 21, 527-539.
- Rieman, H., Aalund, O., 1975. Anvendelse af Evolutionary Operation (EvOp) in veterinær sygdomskontrol. In *Proceedings: Cost-benefit analyse i forbindelse med husdyrbrugets sundhedsstyring* [in Danish], KVL. 40-48.
- Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M., Haynes, R.B., Richardson, W.S., 1996. Evidence based medicine: what it is and what it isn't. *BMJ* 312(7023), 71-72.
- Schmidt, P.L., 2007. Evidence-Based Veterinary Medicine: Evolution, Revolution, or Repackaging of Veterinary Practice? *Veterinary Clinics of North America: Small Animal Practice* 37(3), 409-417.
- Schulz, K., Altman, D., Moher, D., 2010. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *Trials*, 11:32
- Schwabe, C., Riemann, H., Franti, C., 1977. Herd health programs in: *Epidemiology in veterinary practice.* Lea & Febiger, Philadelphia, USA, 246-248. ISBN: 0-8121-0573-7
- Tempelman, R.J., 2009. Invited review: Assessing experimental designs for research conducted on commercial dairies. *J. Dairy Sci.* 92(1), 1-15.

Thorpe, K.E., Zwarenstein, M., Oxman, A.D., Treweek, S., Furberg, C.D., Altman, D.G., Tunis, S., Bergel, E., Harvey, I., Magid, D.J., Chalkidou, K., 2009. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *Journal of Clinical Epidemiology* 62(5), 464-475.

Thyssen, I., Enevoldsen, C., 1994. Visual monitoring of reproduction in dairy herds. *Preventive Veterinary Medicine* 19, 189-202.

Vaerst, M., Sørensen, J., 2009. Danish dairy farmers' perceptions and attitudes related to calf-management in situations of high versus no calf mortality. *Preventive Veterinary Medicine* 89 (1-2), 128-133.

Vaerst, M., Paarup-Laursen, B., Houe, H., Fossing, C., Andersen, H.J., 2002. Farmers' choice of medical treatment of mastitis in Danish dairy herds based on qualitative research interviews. *J. Dairy Sci.* 85(4), 992-1001.

Vaerst, M., Thamsborg, S.M., Bennedsgaard, T.W., Houe, H., Enevoldsen, C., Aarestrup, F.M., de Snoo, A., 2003. Organic dairy farmers' decision making in the first 2 years after conversion in relation to mastitis treatment. *Livestock Production Science* 80, 109-120.

Whitehead, J., 1992. *The design and analysis of sequential clinical trials*. Wiley, West Sussex, England. ISBN 0 741 97550 8

Zwarenstein, M., Treweek, S., Gagnier, J.J., Altman, D.G., Tunis, S., Haynes, B., Oxman, A.D., Moher, D., 2009. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *BMJ* 2008; 337: 2390

3.3 Review of effectiveness of medical treatment for early-postpartum bovine genital disease based on vaginal discharge

Manuscript II

D. B. Lastein & C. Enevoldsen
Department of Large Animal Sciences
Faculty of Health and Medical Sciences
University of Copenhagen
Grønnegårdsvej 2, DK-1870 Frederiksberg C
Denmark

Review of effectiveness of medical treatment for early-postpartum bovine genital disease based on vaginal discharge

D. B. Lastein ^{a*} & C. Enevoldsen ^a

^a Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen

Grønnegårdsvej 2, DK-1870 Frederiksberg C, Denmark

*Corresponding author: dorte.bay@gmail.com (Lastein, D.B.): phone 0045-20641151

Abstract

Background

Evaluation of the disease effect and treatment effectiveness of bovine genital disease before 21 days postpartum is important in both specific herd and veterinary practice contexts. An initial step in evaluating applied protocols would be to identify the best available scientific evidence in the literature to validate the applied clinical diagnostic criteria and the protocols.

Results

The review is restricted to encounter only studies of clinical disease occurrence and evaluating medical treatment effectiveness that are relevant in a herd health management situation. We discuss scientific evidence in the field and compare the findings of relevance to a Danish dairy herd health management program setting where genital disease is diagnosed systematically by clinical vaginal examination of most cows during the 5 to 21 days postpartum period. We found that attempts to obtain uniform clinical definitions of bovine genital diseases are proposed in the literature. Vaginal discharge is a major clinical sign of genital disease. However, the term 'effect' is not used systematically to describe either a disease effect, a disease effect despite treatment, a treatment effect or a difference in treatment effect which complicates systematic reviewing. We found that the proposed clinical disease definitions are not in all cases validated against important key performance indicators. Only few randomized trials with negative and active control groups are performed to evaluate treatment effectiveness and difference in treatment effectiveness of early postpartum genital disease, respectively.

Conclusion

Evidence of a general disease effect of postpartum genital disease before 21 days postpartum based primarily on vaginal discharge indicate milk loss and impaired reproduction. No general practical recommendation on treatment and effect hereof can be given. The following issues are important for practical decision making: choice of antibiotic including administration route and dosage, herd differences in treatment effect, the interaction of retained placenta on the effect of treatment and the importance of both spontaneous recovery and diagnosis and evaluation of 'fatal cases'. Within a Danish herd health management program issues as spontaneous recovery, postponed treatment in relation to effectiveness on reproduction performance and further validation of a vaginal discharge score are warranted.

Keywords

Effectiveness, treatment, metritis, review, vaginal discharge, bovine

Introduction

Bovine genital diseases are expected to cause reduced cattle welfare, loss of milk production, and impaired reproduction performance in the dairy industry. In addition, some genital diseases require the use of antibiotics, which may increase the risk of antibiotic resistance. At present, it has been proposed that specific bovine genital diseases (excluding retained placenta [RP] and pyometra) should be defined according to their occurrence relative to calving (before or after 21 days postpartum [pp]) and the presence of clinical signs (e.g., rectal temperature [RT] or vaginal discharge [VD]) or subclinical signs (e.g., pathological findings in uterine cytology or bacteriology) (Sheldon et al., 2006; Sheldon et al., 2009).

In a Danish herd health management program (HHMP), genital disease is diagnosed systematically by clinical vaginal examination of most cows during the 5 to 21 days pp period (Anonymous, 2010; lr.dk, 2012). VD evacuated during the examination is scored on a ordinal scale (1 to 9) to quantify the severity of the disease at the cow level (Lastein et al., 2009). The examinations and VD scores (VDS) are intended (1) to specify and record defined criteria (e.g., severity of disease) to use for determining whether to give medical treatment to individual cows and (2) to identify trends in the occurrence of genital disease at the herd level, based on disease definitions that are uniform within herds (and potentially, consistently used by veterinarians within and across herds). Unacceptable trends should trigger preventive actions to reduce the occurrence of genital disease. The VDS could be useful for assessing herd-level effects of preventive actions. Once treatment is initiated, a follow-up examination with the VDS might be useful for assessing the clinical effects of treatment.

In a HHMP context, the use of medical treatment is an input factor, like feed or labour. Consequently, evaluating the effect of this input factor should be based on causal relationships between the treatments and some disease indicators. These indicators must have financial or ethical consequences. Examples of key financial performance indicators in the HHMP context are reduced milk production and impaired reproduction performance (Kristensen et al., 2008). Examples of ethical (welfare) indicators can be clinical manifestations of pain or premature culling.

The treatment criteria for genital disease and treatment protocols applied in the Danish HHMP have not been fully validated with a scientific approach (Lastein et al., 2009). Only recently, studies have demonstrated that VDS ≥ 4 was associated with delayed involution (Gorzecka et al., 2011) and impaired reproduction (Elkjær, 2012). An initial step in evaluating these protocols applied in practice would be to identify in the literature the best available scientific evidence of disease effect and treatment effectiveness.

The objective of this review was to evaluate the uniformity of treatment criteria and the effectiveness of therapeutic interventions for early postpartum genital disease (called 'metritis' in this paper), diagnosed by clinical examination of VD. We have restricted the review to studies relevant to the context of recording disease occurrence and evaluating medical treatment effects in a real-world HHMP situation. Consequently the reviews is not performed as a strictly systematic (Liberati et al., 2009), but rather as a 'narrative' (Collins and Fauser, 2005) approach. The paper is organized as follows: (1) A brief description of bovine genital disease with a focus on definitions of metritis in the early postpartum period; (2) a summary of concepts and methods for evaluating effects of disease and effects of medical treatment; (3) a literature review on the effects of bovine genital diseases with a focus on metritis; (4) a literature review of effects of medical treatment on bovine genital diseases with a focus on metritis; and (5) a summary and discussion of the needs for future studies.

Bovine genital disease in the early postpartum period

Here, we will review the definitions of genital diseases in the context of the Danish HHMP. The terms we define will appear in italics throughout this paper. Disease manifestations can be defined from multiple different perspectives (e.g., pathological, bacteriological, or clinical). We will focus primarily on the clinical manifestations, where the term 'clinical' refers to procedures performed as part of a routine examination of cows in veterinary practice with simple cow-side devices. That is, without the use of complex, time-consuming, laboratory equipment. The term 'subclinical' refers to procedures that require rather complex technical equipment or procedures (impractical for cow-side use). In keeping with the HHMP context, we will focus on the literature on genital diseases and treatment effect indicators that are most relevant to the farmer or health consultant. That is, we will include mainly key financial performance indicators and few

animal welfare indicators. As the reader might notice, we use the general term 'genital disease', because the definitions used in epidemiological studies, clinical trials, and literature reviews make it difficult to compare results from different studies. When possible, we will distinguish between relevant clinical entities, as defined below. When we use the term 'metritis', it refers to disease entities before 21 days pp, which is in line with Sheldon et al. (2006).

Lastein et al. studied the attitudes and recording methods used in Danish veterinary cattle practice related to metritis (Lastein et al., 2009). They showed considerable ambiguity in the records and associated definitions, even within the relatively homogeneous Danish context, where substantial effort was made to standardize procedures and recording methods to meet the HHMP requirements defined by veterinary authorities. With this much ambiguity in a homogeneous population, it should be no surprise to find ambiguity in the records and associated definitions among the wide range of contexts presented in the literature. Consequently, a major issue in this review was the evaluation of the comparability of records and criteria in the reported studies.

Uterine involution

The uterine involution process is dynamic. It starts before parturition, as the immune system gradually degrades the attachment of the placenta (LeBlanc, 2008). At parturition, the calf is expelled from the uterus into a more or less contaminated environment - with or without human interference. Postpartum uterine disease is most often caused by an ascending bacterial infection of the genital tract (Thompson, 2011). In 80-100% of cows, a variety of bacteria is present in the uterus within the first 2 weeks postpartum. However, the presence of bacteria in the genital tract does not necessarily cause pathological manifestations (Sheldon et al., 2006; Sheldon et al., 2009). The immune system reacts to a bacterial load by initiating clearance with several defence mechanisms. One is the expulsion of lochia, pus, etc. through the vagina, also called VD in this article. The development of pathological manifestations in the uterus, cervix, and vagina depends on the balance between the amount of tissue damage, the immune system, and the bacterial load. The on-going involution of the uterus in the first couple of weeks postpartum coincides with the processes of follicular activity that initiate the next gestation. Due to this delicate interaction, potential disturbances within the involution process can affect both the re-organization of the uterus and the likelihood of normal follicular development (Sheldon et al., 2008).

The outcome of uterine clearance over the first months postpartum can result in a variety of clinical scenarios, ranging from no effect to toxæmia and death. Within this spectrum, a threshold for medical treatment should be determined by practicing herd veterinarians, who, under Danish regulations, are responsible for the reasonable administration of drugs. The HHMP context calls for practical, inexpensive,

cow-side diagnostic methods. The natural variation in expelled VD complicates the clinical diagnosis of a genital disease within the postpartum period.

Definitions and occurrence of postpartum genital diseases

Table 1 presents definitions of genital diseases that appeared to be consistent among recent scientific papers. The time intervals (e.g., before or after 21 days pp) described in the reviewed articles and in the Danish HHMP are not always consistent and unique. Here, before and after 21 days pp are defined as ≤ 21 and > 21 days pp, respectively. However, we acknowledge that biological variation in clinical signs makes the distinction somewhat arbitrary.

We found some practical inconsistencies in the definitions presented in Table 1. For instance, the definitions do take into account that some cows can have both purulent and fetid VD. Such cows are left without a definition in the definitions proposed by Sheldon and co-workers. Also, a validation problem arises in some of these definitions for the HHMP setting, because, when reduced milk yield is considered a diagnostic indicator of metritis (grades 2 and 3), then it becomes circular reasoning to evaluate milk loss due to disease.

The occurrence of a (genital) disease in a herd can be described in two fundamentally different ways: from records of a disease indicator (e.g., VDS) (*disease records*) or from records of treated cows (*treatment records*).

The incidence risk/rate derived from *treatment records* appears to be both relevant and practical from a monitoring perspective in a HHMP, because the main purpose is to detect changes in the health status **within** the herd over time. Clearly, for a valid comparison of herds, the treatment threshold and diagnostic methods must be uniform within and between herds during the observation period. However, incidence measurements are often used to compare health states between herds without validation of uniformity. As, factors related to personnel, costs, legal constraints, etc. often introduce variability into the definitions and criteria for diagnosis and medical treatment between herds such comparison between herds are problematic. Krogh (2012) and Vaarst et al. (2002) demonstrated and discussed this issue previously (Krogh, 2012; Vaarst et al., 2002).

Examples of reasons for non-comparability between *treatment records* from different herds include the following: different interpretations of disease definitions, different practices of routine examinations, which facilitate detection and treatment of pathological manifestations (Bruun et al., 2002); and different judgments of the severity of manifestations (treatment threshold). If *treatment records* are used to compare the true occurrence of genital disease between herds, it is necessary to evaluate the importance

of bias sources, like those listed above; this may be a daunting task. Therefore, we propose that, when the purpose is to estimate the difference in the true occurrence of genital disease between herds, it is meaningless to use data based on *treatment records*; that comparison would only estimate the variability of definitions and difference in treatment threshold between the herds. To estimate the difference in the true occurrence of genital disease between herds, the disease indicators must be based on classifications of well-defined diagnostic signs. If, a score, like VDS is used concurrently **and** is recorded uniformly in all herds the relationship between the treatment-based diagnosis and the score-based diagnosis can be estimated.

In summary, the occurrence of clinical disease and the subsequent occurrence of medical treatment records depend on the method used to diagnose the disease. In the VDS context, several studies have investigated interrelationships between different clinical methods for evacuating vaginal discharge and different methods for evaluating uterine discharge (McDougall et al., 2007; Pleticha et al., 2009; Runciman et al., 2009; Runciman et al., 2008b). Those studies showed that the gloved hand, a speculum/vaginoscopy, and a device for retrieving discharge (“metri-check”) could be used interchangeably, and these methods were superior to rectal examinations alone for identifying and classifying diseased cows. However, these very practical procedures are expected to be less sensitive than cytology for detecting disease after 21 days pp (Barlund et al., 2008). The method chosen in clinical practice depends on the time and/or equipment requirements for vaginoscopy, ultrasonography, or cytological examinations; the use of these tools to diagnose subclinical conditions is probably limited in every-day veterinary cattle practice, due to logistic (time and hygiene) and economic constraints.

VDS-based disease records, collected under Danish regulations from the HHMP, were summarized for 132 herds in the year 2007. The following estimates of the *period prevalence* were found for genital disease in the period of 5 to 21 days pp: VDS 0-4 = 87%, VDS 5-6 = 8.6%, VDS 7-9 = 4.4% (Sloth et al., 2008). Unfortunately, only cows with at least one breeding were included in this analysis; this could lead to an underestimation of the incidence risk of high VDS values, because fatal or very severe cases were likely to be culled without breeding. We are not aware of studies that estimated the occurrence of fatal cases of genital disease. Indisputably, fatal cases cause pain and discomfort, and thus, they constitute a welfare problem. In a study of metritis treatments conducted by veterinarians in Danish dairy herds in 1993-1994 (call-on-demand treatment records), at least one treatment was performed in 391 of 2144 herds, and the across-herd incidence risk was 0.7% (Bruun et al., 2002). Due to the considerable costs associated with calling a veterinarian, those cases might have been severe or fatal. Consequently, the 0.7% estimate might be a reasonable estimate for the occurrence of indisputable welfare problems due to metritis in a large cattle population.

Table 1. Definitions and abbreviations of bovine genital diseases used in the scientific literature

Definition [Abbreviation in article]	Time postpartum	(Sub)Clinical* signs	General condition	Reference
Retained placenta [RP]	>24 hours	Fetal membranes not expelled		(LeBlanc, 2008)
Puerperal metritis [PM]	0-21 days	Abnormally** enlarged uterus and a fetid, watery, red-brown, vaginal discharge	RT*** >39.5°C, reduced milk yield and dullness (grade 2), and toxemia (grade 3)	(Sheldon et al., 2009)
Clinical metritis [CM]	0-21 days	Abnormally** enlarged uterus and a purulent vaginal discharge	No systemic effect (grade 1) RT≤39.5°C	(Sheldon et al., 2009)
Clinical endometritis [CE]	>21 days	> 50% purulent content of the vaginal discharge at >21 days pp or mucopurulent vaginal discharge at >26 days pp	No systemic effect RTs 39.5°C	(Sheldon et al., 2006) ¹ (Dubuc et al., 2010a) ² (Runciman et al., 2009)
Subclinical endometritis [SE]	>21 days	Absence of vaginal discharge, but >18% neutrophils in uterine cytology (UC) at 21-33 days pp or >10% neutrophils in UC at 34-47 days pp	No systemic effect RT≤39.5°C	(Sheldon et al., 2006)
Pyometra		Accumulation of purulent material within the uterine lumen in the presence of a persistent corpus luteum and a closed cervix	No systemic effect RT ~normal range	(Sheldon et al., 2006)
Cervicitis	<35 days	>5% neutrophil in endo-cervical smear and cytology; can occur independently of endometrial changes	No systemic effect RT ~normal range	(Deguillaume et al., 2012)
Urovagina	>15 days	Urine covering more area than solely the vaginal floor; diagnosed by vaginoscopy	No systemic effect RT ~normal range	(Gautam and Nakao, 2009)
Vaginitis (necrotic) [^]	4-6 days	Inspection and vaginal examination: superficial or worse injuries of the vaginal mucosa; necrotic tissue in the vaginal circumference; and slightly swollen vulva		(Goshen et al., 2012)

* Clinical: procedures that can be performed as part of a routine clinical examination of cows in veterinary practice without the use of laboratory equipment. Subclinical: procedures that require complex, time consuming technical equipment (not practical for cow-side use).**The definition of abnormal versus normal size of the uterus is based on a vaguely-defined classification. Other authors rate relative ability to retract uterus into the pelvis. *** RT = rectal temperature; ¹Purulent vaginal discharge (PVD) can occur independently of cytological endometrial changes. ² The term 'bovine reproductive tract inflammatory disease (BRTID)' is proposed for cases of reproductive tract disease diagnosed using methods that only sample vaginal contents. [^]only examined in first parity cows

Risk factors for metritis

A causal diagram of the relations and timing between different disease entities and other risk factors can clarify a complex disease for practicing veterinarians and researchers. The complexity is immense in a

complete casual diagram for bovine genital diseases. Because we restricted the diagram to the data and the context of a HHMP, we developed a simplified version of cow-level relationships between the records of some important calving events and different postpartum disease entities (Figure 1). However, we expect that the entire causal web in a HHMP context is most likely influenced by herd factors, including hygiene and milk production level. Identification of risk factors for early and late genital diseases are important in analytical studies of treatment and disease effects as they should be examined in light of confounding or interactions (Dohoo et al., 2003). In the following, we will focus on experimental studies or trials to evaluate effect of treatment and disease.

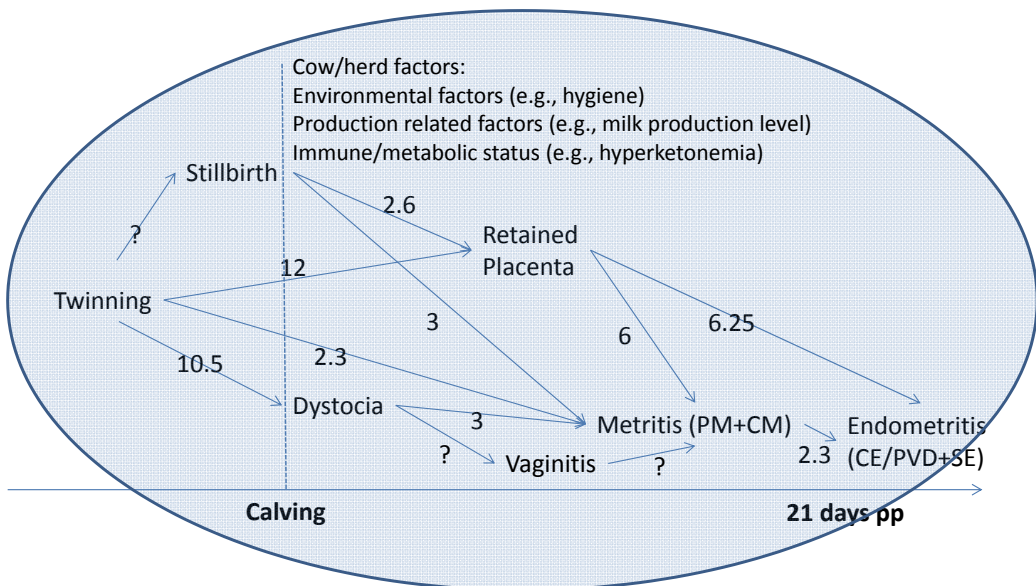


Figure 1. A simplified diagram of the cow-level associations between records of some important calving events and different postpartum genital disease entities. This causal web is adapted and modified from Smith and Risco (Smith and Risco, 2002b) and Dubuc (Dubuc et al., 2010b). The values represent odds ratios; thus, the entity at the left end of an arrow gives increased odds of developing the condition at the right end of the arrow. As indicated, some specific relationships, marked with (?), remain unclear. PM - puerperal metritis, CM - clinical metritis, CE - clinical endometritis, PVD – purulent vaginal discharge, SE – subclinical endometritis; disease classifications are based on definitions used by Sheldon et al. (2006) and Dubuc et al. (2010b).

Concepts and methods for evaluating effects of disease and medical treatment

The literature is inconsistent concerning the terms “disease effects” and “treatment effects”. We find these terms essential for understanding and comparing scientific evidence of effects. We define the *disease effect*

of a genital disease as the manifestation of disease mechanisms (also called outcome, indicator, or *effects measurement*) in untreated diseased cows compared to non-diseased cows. Unless the medical treatment is harmful, the *disease effect* includes the *disease effect despite treatment*, defined as the difference in effect measurements between non-diseased cows and treated diseased cows. We define the *treatment effect* as the difference in effect measurements between untreated diseased cows and treated diseased cows. The *difference in treatment effect* is the difference in *effects measurements* between treated diseased cows that received different treatment protocols (e.g., a comparison between cows treated with drug A or protocol 1 and cows treated with drug B or protocol 2).

In experimental settings, diseased, untreated cows are often called 'negative controls' (or placebo, when a placebo-treatment is used). Alternatively, because all treatments are expected to influence the outcome when a trial aims at estimating a *difference in treatment effect*, the term '*active control*' are used to describe a control group that consists of treated diseased animals.

The definitions above imply explicit treatment criteria for classifying cows into two categories: 'non-diseased' (not to be treated) and 'diseased' (to be treated). The difference in *effects measurements* between these two categories should be relevant to the end-user. Thus, the applied treatment protocol should benefit a sufficient proportion of cows, and cause a sufficient *treatment effect* (often referred to as a clinically relevant difference, in a trial context). The determination of 'sufficient' is influenced by contextual circumstances, including ideology, financial and animal welfare considerations.

Trials should be designed to detect superiority, equivalence, or non-inferiority (not worse) of one intervention compared to another. In equivalence or non-inferiority studies, a clinically relevant difference (margin) must be defined (Habicht, 2011). Additionally, a trial evaluation requires an awareness of the overall purpose of the trial. When the *effects measurement* is chosen to evaluate the usefulness of the intervention for the end-user (in this case the farmer), we use the term '*evaluation of effectiveness*' (Thorpe et al., 2009). A prerequisite for an *evaluation of effectiveness* is the establishment of a causal relation between the clinical manifestations and this effects measurement, often being a financially relevant performance indicator. Under the assumption that we are able to record clinical manifestations uniformly, several additional obstacles must be overcome to estimate causal relations (*disease effects* and *treatment effects*):

- The clinical measurement scale(s) must be valid for the purpose. For example, when puerperal metritis (PM) is defined by two or more clinical signs (e.g., dullness and fetid odour) that are recorded on a dichotomous scale, they can be combined into four categories. With some sort of

weighting, a linear or ordinal scale may be constructed based on general knowledge; ideally that scale will require empirical evidence and subsequent validation. Substantial work is needed to define appropriate measurement scales. A straightforward approach would be to relate individual clinical signs to a performance indicator. The observations could then be used to construct the clinical scale. This type of analysis might also show that some clinical signs are irrelevant for herd management purposes. The index created by Gorzecka et al. (2011) provided correlations between records, but individual clinical signs were not correlated with performance. In the Danish HHMP, the VDS appears to be based on general knowledge; the ordinal scale is linked to the severity of disease, by some veterinarians also measured in terms of systemic effects (Lastein, 2012). This makes sense when the aim is to relate milk yield loss or poor appetite to subsequent loss of body condition. However, the VDS scale also contains multiple clinical signs that can be combined in multiple ways. Consequently, the VDS could be improved with empirical validation.

- The trial design must be effective. Knowledge about the true *disease effect* (on some performance indicator) is essential for rational decision making. In a complex computer model of an entire dairy herd (www.simherd.com), the user (e.g., a practicing veterinarian) can specify *disease effects* and *treatment effects* for the major disease complexes observed in dairy cows. Estimates of causal effects are required. The default estimates in the SIMHERD-model are based on information from the literature, which probably include some of the sources used for the current review. To obtain unbiased *disease effects* and *treatment effects* from a trial, it is essential to include untreated diseased animals. To estimate unbiased *treatment effects*, randomization (and blinding) is also essential. In case of systemic signs of disease, randomization and negative controls may be impossible to include in a trial, due to ethical constraints. These constraints may explain why the discussion below contains few studies of effects related to PM based on trials with negative controls. Those trials are rarely conducted in the early postpartum period. The so-called evolutionary operation (EvOp) design may be able to overcome these constraints (Lastein, 2012).

A final issue is uncertainty. The proportions of cows that are allocated to treatment in veterinary practice are governed by diagnostic precision and accuracy (Lastein et al., 2009). Even with consistent, uniform measurement scales and valid (unbiased) estimations of causal relationships between measurement scales and performance indicators, the predictions are associated with uncertainty (random error). Krogh et al. (2011) demonstrated an approach for estimating this uncertainty of clinical diagnosis in entities without access to a perfect test.

The characteristics of the trial design determine what type of scientific evidence of *treatment effects* can be obtained from the trial. Habicht (2011) suggested that a hierarchy of evidence could be used to classify scientific literature on clinical trials in human medicine. Applied to our context, best **general** evidence of treatment effect should be based on multi-herd, randomized trials, with a negative control and sufficient sample size, or meta-analysis of multiple of such trials.

Disease effects related to genital disease in dairy cows

Unfortunately, the literature seldom explicitly specifies *disease and treatment effects* as defined above. This deficiency reduces the value of those reports. In this section, *disease effect* and *disease effect despite treatment* are categorized according to our definitions and the definition proposed in table 1 by other authors. Furthermore, in a separate section, we will elaborate on the *disease effect* related to a VDS before and after 21 days pp. Although a manifestation like VDS is not a disease per se, we consider it a *disease effect*.

We have gathered the best available scientific evidence of the *disease effects* of PM, clinical metritis (CM), clinical endometritis (CE; or purulent vaginal discharge), and subclinical endometritis (SE) on two important performance indicators, milk production and reproduction performance (Table 2). The data is from trials with negative control groups that provided estimates of *disease effect*. A review of epidemiological retrospective studies that provided estimates of *disease effect despite treatment* are described separately. Studies on RP are omitted, but the importance of RP as a risk factor for other genital diseases will be discussed below. We also present the diagnostic criteria used in the respective trials to classify diseased and non-diseased cows. Table 2 mainly includes studies, where pregnancy rates as outcome for reproduction performance were based on time-to-event analysis; this methodology reduced the risk of bias due to management decisions (e.g., selection bias) and lack of data (LeBlanc, 2010). In summary, early (≤ 21 days pp) genital diseases can negatively affect both milk production (app. 300 kg milk over a 305 days period with a considerable variation) and reproduction performance (non-diseased app. 3 times as likely to conceive at first service and reduced pregnancy rate). In the later postpartum period (22-60 days pp), disease entities were rarely analysed for associations with milk loss, so no scientific evidence of milk loss are available. However, the evidence is consistent that a late genital disease had a negative disease effect on reproduction performance with a reduced pregnancy rate between 25-60%.

Table 2. Studies on the *disease effects* of bovine genital diseases. Disease entities are defined according to the criteria proposed by Sheldon *et al.* 2006. The table should not be regarded as a complete review, because it is restricted to studies that used non-diseased and untreated diseased groups in comparisons. Cells in the table marked with grey shading refer primarily to disease entities before 21 days pp.

Disease entity (as defined in Table 1)	Examination procedure	Days postpartum	Diagnostic indicators	Estimate of DISEASE EFFECT (untreated diseased cows vs. non-diseased cows)	Reference
Clinical metritis and puerperal metritis	Vaginal and transrectal exploration	5-14	Flaccid, non-retractable uterus, cervical diameter >75 mm, and watery or purulent, fetid vaginal discharge. Cows with retained placenta were not included in trial.	Milk yield: difference of 337 (std.dev 145) kg at 305 days ME* Reproduction: odds ratio of 2.7 for conception at first service for non-diseased cows. Pregnancy rate differed significantly (<0.05), but estimates not reported	(Goshen and Shpigel, 2006)
Puerperal metritis	Farmers' observations	≤20	foul-smelling and brown-red, watery vaginal discharge with RT >39.5°C	Milk: difference of 259 kg projected at 305 days yield for parity**>1 (Retained placenta had additive negative effect of app.700 kg)	(Dubuc <i>et al.</i> , 2011) (based on trial data with no significant treatment effect)
Clinical metritis and clinical endometritis	Vaginocopy	15-60	Mucopurulent or purulent vaginal discharge (flecks of pus not indicative)	Reproduction: 61% reduction in pregnancy rate for diseased cows	(Gautam <i>et al.</i> , 2009)
Clinical endometritis	Vaginocopy	20-33	Purulent uterine discharge or cervical diameter 7.5 cm after 20 days pp or mucopurulent discharge after 26 DIM	Reproduction: 27% reduction in pregnancy rate for diseased cows	(LeBlanc <i>et al.</i> , 2002a; LeBlanc <i>et al.</i> , 2002b)
Clinical (Purulent vaginal discharge) and subclinical endometritis	Vaginal examination and endometrial cytology	35-56	Mucopurulent (or worse) vaginal discharge and >6% neutrophils, respectively.	Milk yield: no difference in yield at 305 days. Reproduction: 28-36% reduction in pregnancy rate for diseased cows	(Dubuc <i>et al.</i> , 2010a; Dubuc <i>et al.</i> , 2011)
Subclinical endometritis	Endometrial cytology and ultrasonography	20-33 34-47	>18% neutrophils or fluid in uterus >10% neutrophils or fluid in uterus	Reproduction: 41% reduction in pregnancy rate at 20-33 days pp and 51% reduction at 34-47 days pp for diseased cows	(Kasimanickam <i>et al.</i> , 2004)
	Endometrial cytology	40-60	>5% neutrophils	Pregnancy rate differed significantly (p<0.0001), but estimates not reported. Median days open was 206 for diseased cows or 118 for non-diseased cows	(Gilbert <i>et al.</i> , 2005)
Cervicitis and subclinical endometritis	Cytology	21-60	Cervicitis: ≥5% neutrophils Subclinical endometritis: ≥6% neutrophil	Reproduction: Pregnancy hazard ratio was 3.4 higher for non-diseased cows than for cows affected by both conditions	(Deguillaume <i>et al.</i> , 2012)

* ME=mature equivalent

In a review of retrospective epidemiological studies on genital disease, the *disease effect despite treatment* on milk yield was inconclusive (Fourichon *et al.*, 1999). This may be due to the inherent problems with differences in diagnostic criteria and the use of non-randomized *treatment records* to estimate the *disease*

effect despite treatment. A similar analysis of genital diseases showed a clearly negative *disease effect despite treatment* on reproduction performance (Fourichon et al., 2000).

It also follows, that a disease effect from the early genital diseases on reproduction may be mediated through the late diseases. The causal diagram described above however, indicates that the early and late disease entities are closely related. That is, an early entity can precede a late disease, but this is not always the case. Also, multiple different disease entities may or may not appear concurrently. Moreover, the effects of different concurrent diseases entities may be additive. The 'true' causal relation between each disease entity and an effect measurement is difficult to establish from field data, because all disease entities are seldom recorded systematically, concurrently, and validated in the same study.

The *disease effect* on selected welfare indicators, like clinical condition and culling, will be reviewed in the subsequent section. The systemic manifestations of PM are described in Table 1 as part of the definition of disease (e.g., RT, dullness, toxaemia). The clinical indicators of metritis can be useful indicators of welfare, but the criteria are far from uniformly defined. Of the systemic indicators, RT may be particularly useful, because it can be measured on a well-defined scale. In fact, RT is widely used as a clinically relevant indicator of *disease effect* (and clinical treatment effect). However, Benzaquen et al. showed that not all cows with abnormally enlarged uterus and fetid, watery, red-brown VD have increased RT (Benzaquen et al., 2007). Therefore, we need estimates of the relationship between RT and the direct manifestations of metritis, and we need to know what potential changes occur in these relationships during the involution period. We also need to define accurate scales for measuring dullness, appetite, and other clinical manifestations mentioned in Table 1. To detect a possible effect on more direct indicators of welfare, we need very detailed records, like activity measurements, regurgitation measurements, or rumen fill. Fortunately, these types of intensive cow-level measurements are available in a growing number of commercial herds. Consequently, there are practical options for describing the relationships between a range of clinical manifestations of genital diseases and indicators of welfare (e.g., pain and discomfort). These descriptions can be used to select appropriate cut-off values to distinguish between relevant and irrelevant clinical manifestations ('severe' and 'non-severe').

Early culling may be caused by a disease that causes poor appetite, reduced milk yield, and extraordinary loss of body condition, which may also cause poor fertility. Consequently, one effect of genital disease (e.g., PM) may be premature culling. With a uniform definition of genital disease, one can estimate the relationship between the occurrence of genital disease and the time of culling. This relationship can be used as an indicator of reduced welfare. A genital disease that causes premature culling is also financially relevant. Knowledge about the culling profile is also important in the evaluation of other indicators,

because culling may cause selection bias. For example, we may want to estimate the relationship between genital disease and milk yield. For some (severe) genital diseases that directly or indirectly cause culling shortly after calving, the diseased cows may not contribute to milk yield records. In that case, we will underestimate the effect of genital diseases on long term estimates of milk production (e.g., 305-day yields). Analytical methods to predict lactational yield can be used to circumvent this problem.

In a recent study on genital disease effects on culling, *treatment effect* was accounted for with analytical control (e.g., including a treatment variable (yes or no) in the analytical models). That led to the conclusion that culling rates were not directly affected by RP, PM, CM, CE, or SE. However, indirectly, the pregnancy rate was associated with increased culling rates; thus, it followed that premature culling could be mediated through genital disease (Dubuc et al., 2011). Another study concluded that parity might influence the association between early genital disease and culling, because multiparous cows with metritis were more likely to be culled within 305 days pp than cows without metritis. This association was not found in first parity cows (Wittrock et al., 2011).

Vaginal discharge as the only indicator of genital disease before 21 days pp

If the definitions in Table 1, and the VDS in the Danish HHMP (appendix A) are compared, a VDS of 7-9 indicates PM, and a VDS of 5-6 indicates CM. However, at present, the Danish VDS system does not explicitly distinguish systemic from non-systemic disease. VDS, by definition, is purely based on a description of the discharge. Scientific evidence indicate that cows with a VDS ≥ 4 (mucopurulent discharge) have impaired reproduction (Elkjær, 2012). As we see it, clinical definition of this condition is lacking in the definitions presented in Table 1. Furthermore, the Danish HHMP operates with a non-validated ordinal scoring system [0-9] of vaginal tear (lr.dk, 2012), which might distinguish diseases in the vagina from diseases in the uterus. No studies have been performed on the validity of these vaginal tear scores with regard to diagnosis or consequences.

The association between VD and milk production is highly relevant from a HHMP point of view. A randomized, clinical field trial with a negative control group diagnosed metritis based on criteria of a fetid, watery or purulent VD, detected between 5-14 days pp, combined with the cervix size and uterus position (Goshen and Shpigel, 2006). That study indicated that the *disease effect* of metritis on the 305-day milk production (and reproduction) was reduced by medical treatment in cows above the first parity. The study also demonstrated the importance of distinguishing between metritis with and without a preceding RP; in a separate analysis, cows with RP showed differences in treatment effect on milk production compared to metritic cows without RP treated with the same protocol. Another study also demonstrated additive

negative effects of RP and metritis in multiparous cows (Dubuc et al., 2011). In conclusion, the interactions between RP and metritis, and their effects on milk production are complex. These interactions should be accounted for in analyses of the *disease effect* of genital diseases and in handling the disease entities in practice.

An increase in the VD index over the first 6 weeks postpartum, based on clinical parameters of temperature, odour, colour, and volume (as used in the Danish VDS), was associated with delayed uterine involution (Gorzecka et al., 2011). However, the associations between the VD index and milk production was not evaluated. The advantage of an index that combines multiple signs into one entity is that the evaluation of the index resembles a clinical decision-making process on the cow level (e.g., combining and weighing different diagnostic sign depending on the relative importance in the specific case). Methods that combine several clinical signs could potentially increase the positive predictive value of the diagnostic procedure (e.g., increase the likely hood that a diagnosis positive cows will are 'true diseased'). The disadvantage of an index is that we lose understanding, because we cannot tell which underlying variables (e.g., RT or VD) had the major impact on the outcome of interest. Also, some practical problems are related to the implementation of an index (when complex) in practice.

Treatment effect caused by medical treatment of puerperal and clinical metritis

Above, we found some evidence of *disease effects* for different genital disease entities that may be important for herd management, due to their negative effects on milk yield and fertility. PM and CM affect both milk production and reproduction; CE and SE affect reproduction performance. The effects of these diseases on culling are less evident. Here, we will only focus on the relationships between genital diseases, medical treatment of genital diseases, and milk yield or reproduction outcomes.

To reduce the *disease effect*, two HHMP objectives are clear: (1) Treat the affected cows and (2) prevent future disease. Because our focus is only on the first objective, we will only consider the intervention (therapeutic treatment of PM and CM) and its related *treatment effect* or *difference in treatment effect*. For completeness in the evaluation of *treatment effects*, in general, we note that multiple trials have studied *treatment effects* of antibiotics and hormones given against RP and applied later than 21 days pp (Drillich et al., 2007a; Drillich et al., 2006a; Drillich et al., 2003; Drillich et al., 2006b; Kasimanickam et al., 2005; LeBlanc et al., 2002b; Sheldon and Noakes, 1998; Sheldon et al., 2004). The results of those studies will not be evaluated or reviewed here.

Therapeutic treatments for metritis can be categorized into four main categories, as follows:

1) *Antibiotic treatments*, local intrauterine or systemic, aim to reduce the bacterial load in the postpartum genital tract, and hence, prevent progression of disease (Smith and Risco, 2002a).

2) *Hormonal treatments*, with prostaglandins or gonadotropin releasing hormones, aim to induce oestrus, and thus, increase uterine discharge evacuation and increase mucus production via host defence compounds (Smith and Risco, 2002a).

3) *Irritating intrauterine treatments*, with antiseptic irritating intra-uterine solutions (e.g., chlorhexidine, iodine etc.), aim to increase tone, blood flow, and defence mechanisms (Smith and Risco, 2002a).

4) *Anti-inflammatory treatments*, with non-steroidal drugs, aim to reduce inflammation and related tissue damage, and increase animal well-being (Drillich et al., 2007b).

Recent (after 1998) scientific trial data on genital disease-related *treatment effects* and *differences in treatment effects* are summarized in Table 3. In this review, most studies used treatments based on the antibiotic group of cephalosporins, applied either parenteral or via the intrauterine route. The reason for this 'preferred drug choice in test' could be due to the recent availability on the market, sponsored interests in studies, or consideration of the advantages of this drug group (e.g., short milk withdrawal period). All the reviewed studies were designed and analysed as superiority trials (in contrast to non-inferiority or equivalence trials). Consequently, the conclusions from these studies could only imply that there was or was not statistical evidence for a difference between treatment groups (e.g., the null-hypothesis could or could not be rejected). That is, they did not provide statistical evidence that any treatment was equally good (or bad) or that it was no worse than another treatment (Altman and Bland, 1995). The studies were selected based on their relevance to a HHMP context (e.g., practical diagnostic criteria) and, to some degree, based on our judgment of scientific evidential value related to the design, sample size, and analytical methods.

In summary, the reviewed studies in table 3 indicate some general evidence that intrauterine tetracycline and cephalosporin administered to cows with PM and CM can reduce milk production and reproduction loss for some parities and some levels of retained placenta based on larger multi-herd studies. Effects of intra-muscular antibiotic treatments were not tested in negative controlled trials with relevant results measurements for a HHMP (e.g., long term milk yield or pregnancy rate). Reports on the effects of hormonal and anti-inflammatory treatments suffer from the same problems as stated above.

Table 3. Selected studies on *treatment effectiveness* and *difference in treatment effectiveness* for puerperal and clinical metritis. Table continued on the next page.

Effect measurement	Inclusion criteria	Treatment groups, Intervention type, and simplified protocol	Estimates of treatment effect or treatment difference	Study design	Reference & comments
Milk yield (kg at 305 days mature equivalent) Reproduction	Flaccid, non-retractable uterus, cervical diameter >75 mm, and fetid, watery or purulent vaginal discharge.	1. Diseased treated: (Tetracycline IU) 2. Diseased non-treated controls 3. Non-diseased controls	Milk yield: No significant difference between non-diseased controls and diseased treated, for multiparous cows. Treatment effect of ~350 kg milk per lactation in diseased treated cows compared to diseased non-treated controls. Reproduction: No significant difference in pregnancy rate between non-diseased controls and diseased treated cows. Treatment effect of ~25 days reduction in days open in diseased treated cows compared to diseased non-treated controls.	Negative controlled, pseudo randomized, non-blinded, multi-herd (5) Ntotal~2320	(Goshen and Shpigel, 2006) Cows with RP were analyzed separately. RP- results not discussed here.
Milk yield at 1-12 days pp	Febrile, RT >39.2°C, enlarged uterus and cervix, and fetid VD, reduced milk yield	1. Diseased treated: Penicillin IM 2. Diseased treated: Penicillin IM + oxytetracycline IU 3. Diseased treated: Ceftiofur IM	No significant difference in milk yield at 1-12 days pp between the 3 groups	Active controlled, randomized (method not described), non-blinded, single-herd Ntotal~50	(Smith, 1998) RP not accounted for in analysis
Reproduction Culling	Febrile, RT>39.5°C, Fetid, red-brown VD	1. Diseased treated: Ceftiofur IM 2. Diseased treated: Ampicillin IM + ampicloxacin IU 3. Diseased treated: Ceftiofur IM + ampicloxacin IU	No statistical evidence for a difference between the 3 groups in the proportions of cows inseminated, days to first insemination, the risk of conception at first service, days open, risk of pregnancy, or culling before 200 days pp	Active controlled, pseudo randomized, non-blinded, single-herd Ntotal~230	(Drillich et al., 2001)
Reproduction	RP and subsequent fetid VD and enlarged uterus	1. Diseased treated: Ceftiofur IM + PGF2 2. Diseased treated: Ceftiofur IM	Significant difference in conception risk at first service (OR=4.15; 95%CI: 1.05–16.5) for first parity cows in treatment group 1 compared to group 2	Active controlled, randomized, blinded, single herd Ntotal~200	(Melendez et al., 2004)
Reproduction	Large, flaccid uterus, and watery, fetid VD	1. Diseased treated: Tetracycline IU + GnRH + 2xPGF2 2. Diseased treated:	Significant difference in risk of conception at first service between groups (treatment effect of ~15% improvement), but no significant difference in overall risk of pregnancy	Active controlled, randomization procedure not described, non-blinded, and single-	(Janowski and Zdunczyk, 2001) Univariate

		Tetracycline IU		herd Ntotal~70	analysis
Milk yield within 6 d after the first treatment Reproduction	Febrile, RT >39.5°C, Fetid, red-brown VD	1. Diseased treated: Ceftiofur IM 2. Diseased treated: Ceftiofur IM + Flunixin meglumine 3. Non diseased, untreated controls	No statistical difference between groups 1 and 2 in milk yield at 0-6 days post treatment, risk of conception at first service, risk of pregnancy within 200 days pp, or days open. Poor reporting of the difference between non-diseased controls and groups of diseased treated cows.	Active controlled, pseudo randomized, non-blinded, and single-herd Ntotal~230 (+ 9 non-diseased controls)	(Drillich et al., 2007b)
Reproduction	Purulent or mucopurulent VD at 7-28 days pp in seasonal breeding herd Inclusion criteria based on high risk prognostic factors; e.g., RP	1. Diseased treated: Cephapirin IUx1 2. Diseased, untreated controls 3. Non-diseased controls	Significant improvement in risk of conception at first service (OR~1.5-2 95%CI ~1-4) and pregnancy hazard rate (HR~1.4; 95%CI ~1-2) at post mating start date for diseased treated cows compared to diseased non-treated controls. No significant difference between non-diseased controls and diseased treated cows.	2 x Negative controlled trial, pseudo randomized, and non-blinded multi-herd NItotal~400 cows, 6 herds NItotal~1150 cows, 17 herds	Study I: (Runciman et al., 2008a). Study II: (Runciman et al., 2009)

Abbreviations: RP=retained placenta; RT=rectal temperature; IM=intra muscular; IU= Intrauterine; VD= Vaginal discharge; N_{total}= total sample size for analysis

Information from studies earlier than 1998 were reviewed by Hoedemaker and Smith & Risco (Hoedemaker, 1998; Smith and Risco, 2002a). Their most prominent recommendations are summarized in the following remarks (in italic font), and our considerations are added (in regular font):

- *Acknowledgement of the occurrence and importance of spontaneous recovery.* This aspect is also recently studied in cows examined between 15 and 60 days. Spontaneous recovery was seen in app. 75% of case (Gautam et al., 2010). Clearly, we acknowledge that spontaneous recovery is very important for rational clinical decision-making. Two issues are noted: (1) when the occurrence of spontaneous recovery is ignored, overtreatment is the logical consequence on the herd and national levels; (2) some cases might be explained by problems in the diagnostic procedure; for example, diagnostic imprecision (e.g., random error in the measurement scales) or an unknown or accepted suboptimal positive predictive value (PPV: the probability that, given a positive veterinary diagnosis, the cow would actually benefit from the treatment) (Dohoo et al., 2003).
- *Individual cow-level anamnesis should be accounted for.* This issue is particularly important in severe cases with a potentially fatal outcome. In a HHMP context, in more or less industrialized

herds, some standard diagnostic procedures are highly relevant as part of a screening program to detect disease that could negatively affect key performance measurements. Special care for the very diseased cows should follow this initial screening (or be provided at first occasion whenever it must occur).

- *Treatment effects differ from herd to herd.* The review showed that this statement is seldom discussed intensively. Focus in research is often on detecting general associations, but the study contexts and the applicability beyond these are seldom discussed. Arguments of the importance of herd differences are supported by the ‘within herd multivariable analysis of risk factors’-approach in an Israeli context (Nir, 2008). Similar approach can be used elsewhere to estimate local treatment effects based on systematically collected herd data. As both the disease and treatment effect could depend on the herd, the parity, and other prognostic factors (like RP); we propose that both retrospective and, potentially, prospective herd data analyses should be conducted regularly to refine diagnostic procedures and treatment protocols for genital diseases.
- *The least harmful treatment should be selected. Caution is needed to prevent inefficient intrauterine application of antibiotics and antimicrobial residues due to high-dose or off-label doses, and inappropriate administration routes.* These comments are highly relevant in the Danish context. In Denmark, few antibiotics are registered for intrauterine application. Also, the dairy factories have increased their focus on milk withdrawal and antimicrobial residues, and the authorities emphasize the prudent use of antibiotics. These circumstances limit the therapeutic choices for veterinarians in practice, and it may be disputed whether the ‘best available scientific evidence’ is or can be implemented legally in Danish practice. For instance, the only treatments found effective in negative controlled trials in the reviewed studies were intrauterine tetracycline in high doses and intrauterine cephalosporin. However, these options are illegal under Danish conditions, which leave Danish veterinarians with little scientific evidence to support their choice of therapeutics for metritis in practice.
- *The recommendations regarding choice of administration route (IM vs. IU) differ.* In the reviewed literature, we found two studies that showed treatment effects of antibiotics applied inside the uterus. However, none of the 5 reviewed studies using IM and IM+IU protocol are designed with a negative control group. Therefore the treatment effectiveness of these protocol cannot be evaluated, only difference in treatment effect. We have found no trials testing these active control protocols against negative control groups. Consequently, we found no general evidence the benefit of using a IM or an IM + IU protocol (which often used in Danish conditions (Lastein, 2012))

Furthermore, we would like to address briefly the problems related to trials with positive control groups. These trials are only recommended when (1) it is ethically irresponsible to conduct the same trial with a negative (or placebo) control group and (2) a validated treatment protocol is available (EMEA, 2001). Otherwise, the results of a superiority study (comparing two active treatments) would be difficult to evaluate. In case of a non-significant difference between treatment groups, **no** conclusions can be drawn as to whether they are equally 'bad' or 'good'. Inclusion of a third, comparable, non-diseased group can facilitate an analysis to demonstrate whether one or the other active treatment differs in outcome from the non-diseased groups. A vague attempt at this approach was performed by Drillich et al. (Drillich et al., 2007b), but it was better implemented by Goshen et al. (2006) together with the 'negative control group' (diseased un-treated).

The reported findings indicated that there are a few other issues of relevance to the Danish HHMP: (1) spontaneous recovery can be substantial, (2) the validity of VD as the only clinical diagnostic indicator of disease warrants more research, (3) when the HHMP aims to improve reproduction performance, a postponed (beyond 21 days pp) systematic examination could be useful, and (4) considerations are warranted of the *treatment effects* of RP and subsequent VD.

Discussion and future research

The definitions of genital disease (Table 1) appear to be derived from a (qualitative) pathological and physiological perspective, where the aim was to understand the disease. That is, they focused on the pathogenesis and disease processes. That understanding is very useful for selecting treatments and formulating a prognosis for a single, complex patient. That way of diagnosing disease is typical for veterinarians working in a hospital setting, where few animals might require intensive, costly treatment, and where expensive, complicated diagnostic tools are available. In contrast, veterinarians working in industrialized dairy herds with numerous cows face practical, logistic, and financial constraints that require different ways of thinking and acting. In the relatively rare situation of a very sick cow, a detailed clinical examination is warranted in either setting; that is, the veterinarian/farmer must decide whether a (possibly costly) treatment is justified to promote welfare and profitability, or whether killing would be the better choice. This decision is context-dependent and qualitative in nature, because it involves human attitudes and values in addition to strictly financial considerations. A series of systematic case studies may provide sufficient evidence to develop a context-based, best veterinary practice recommendation for individual severe cases. The perspective of pathologists or physiologist described above is difficult to apply to the situation in HHMPs, where high numbers of cows are examined in the shortest possible time. In that case, a

frequent event (e.g., occurrence of VD) would initiate a decision-making process to determine whether some more or less distinct manifestations of disease should be handled with a standard intervention (e.g., supplementary examination or medical treatment). The literature review above demonstrated that very few out of a large number of scientific publications provided convincing evidence to support a standard decision of whether to treat based on typical types of VD.

To make financially rational decisions in a dairy herd, an essential prerequisite is to implement a uniform and explicit definition when recording the clinical manifestations of disease in all cows. This requirement is met with some of the manifestations reviewed here (e.g., RT). Others, like dullness or the ability to retract the uterus, require substantial calibration efforts to ensure uniformity. Kristensen et al. (2006) described the diversity in body condition scores recorded by practicing veterinarians; they found that considerable effort was required to achieve uniformity (Kristensen et al., 2006). Kristensen et al. (2006) and Lastein et al. (2009) demonstrated the problems associated with recording genital diseases. The lack of uniformity makes it problematic to compare results among the studies in this review. Consequently, it is difficult to apply the results from one study to a specific herd (limited external validity). However, these studies provided detailed information about clinical manifestations, which are needed to create more efficient measurement scales for practical use.

The above definitions (Table 1) described clinical diagnoses of uterine diseases, often based on VD, with the assumption that a discharge of uterine origin drained to the cranial vagina. However, this assumption might not be valid, because VD can derive from the uterus, the cervix, the vagina, or the urethra (urovagina). In the Danish HHMP, the VDS and subsequent treatment criteria are also based primarily on the assumption of a direct correlation between uterine health and VD. Further work on this issue is warranted; however, from a very practical and pragmatic point of view the origin of the discharge is less relevant if treatment and prognosis are somewhat alike.

Essentially, to make a complex herd-level decision, a supportive model, like the above-mentioned SIMHERD model is useful. However, the model requires estimates of *disease effects* and *treatment effects* (as shown in Table 4). To fill in the blanks, this review cannot provide reasonable estimates of *disease effect before day 21 pp* (2 studies on milk loss, 2 studies on reproduction impairment – table 3). This review can provide reasonable *disease effect* estimates for genital diseases that appear after 21 days pp (5 studies on reproduction impairment in table 3). The *treatment effect* estimates for late disease entities are deliberately omitted here due to the large number of studies and the irrelevance in the Danish HHMP context, but they can be found in peer-reviewed articles. A major reason for this difference in the use of negative control groups before and after 21 days pp might be that systemic disease manifestations are

absent in the late period; this facilitates the realization of efficient trials, due to fewer ethical and financial constraints. In contrast, in the early postpartum period, systemic disease manifestations impose ethical, and perhaps more obvious, financial constraints on the trial design (e.g., milk production loss is ‘obvious’, but impaired reproduction is ‘hidden’). In fact, we could only find one *treatment effect* study that was conducted before 21 days pp (Goshen and Shpigel, 2006) and met the essential trial criteria; i.e., a systematic examination was conducted by experienced (unbiased) personnel (not farmers), a negative control group was included, and randomization was implemented. However, due to inconsistent definitions of the measurement scales, none of the reviewed studies allowed the separation of severe and non-severe clinical manifestations of metritis in the early postpartum period. This problem was exemplified in a trial that excluded cows with toxic diseases (Melendez et al., 2004). Some older textbooks might provide more information about *disease effects* in these severe cases and additional information about associated diseases, like (uro)-vaginitis and cervicitis.

Table 4. Schematic classification of bovine genital disease to support financially rational decisions regarding key performance indicators (for instance milk yield). The cells should contain the decision-maker’s best estimates of *disease effects*; cells are left open for future research. Shaded cells represent combinations that are practically irrelevant.

Manifestations of uterine diseases		5 to 21 days pp		Later than 21 days pp	
		Treated	Non-treated	Treated	Non-treated
Clinical*	High risk of fatal*** outcome				
	Low risk of fatal outcome				
	Local				
Subclinical**					

* Disease manifestations are visible or detectable with inexpensive, cow-side tests that give immediate results in the barn. ** The diagnosis requires laboratory equipment. *** Killing or very intensive medical intervention (as opposed to ‘standard treatment’) is required for ethical reasons and thus the effect on milk yield is dramatic and, therefore, obvious.

The distinction between high and low risk of fatal outcome is probably straightforward. Signs of shock and severe pain should be readily apparent to skilled farm personnel, and certainly, to veterinarians. These ‘near-fatal’ cases are probably quite rare, as discussed above, but further studies might be useful on the occurrence and the best practice in handling this type of genital disease. The distinction between local and systemic manifestations is important in this context, because it is unlikely that purely local manifestations cause reduced food intake, reduced milk yield, or impaired body condition. It follows that evaluations of metritis and *treatment effects* on milk yield should be restricted to cases with systemic manifestations. However, it is also likely that purely local manifestations might cause delayed involution and subsequent impaired reproduction performance (Elkjær, 2012). It is debatable whether local and systemic

manifestations can be distinguished based solely on one diagnostic parameter (i.e., VD); however, this distinction is highly relevant in the process of screening cows for disease (or supplementary examination) and/or treatment. A supplementary (to the initial screening), fully systematic, clinical examination of the entire cow would clarify the distinction between local and systemic manifestations.

Appendix A

Table of vaginal discharge score (VDS) in the Danish Herd health management program

VDS	Odour	Discharge description (volume, contents/colour)
0	No	No or minimal volume of clean mucous discharge
1	No	Minimal volume of bloody mucous discharge
2	No	Small volume of bloody mucous discharge
3	No	Considerable volume of bloody sero-mucous or mucopurulent discharge
4	No	Considerable volume of mucopurulent discharge (yellow)
5	Abnormal	Minimal to plenty amounts of purulent discharge (yellow)
6	Abnormal	Considerable volume of purulent discharge (yellow)
7	Fetid	Considerable volume of purulent/haemorrhagic discharge (yellow, red, brown)
8	Fetid	Plenty volume of watery haemorrhagic discharge (red, brown, grey)
9	Fetid	Large amounts of watery discharge and debris (red, black)

References

- Altman, D., Bland, J., 1995. Statistics notes: Absence of evidence is not evidence of absence. *BMJ* 311(485).
- Anonymous, 2010. Bekendtgørelse om sundhedsrådgivningsaftaler for kvægbesætninger [in Danish]. <https://www.retsinformation.dk/Forms/R0710.aspx?id=132648>. Assessed 20.09.2012.
- Barlund, C.S., Carruthers, T.D., Waldner, C.L., Palmer, C.W., 2008. A comparison of diagnostic techniques for postpartum endometritis in dairy cattle. *Theriogenology* 69(6), 714-723.
- Benzaquen, M.E., Risco, C.A., Archbald, L.F., Melendez, P., Thatcher, M.J., Thatcher, W.W., 2007. Rectal temperature, calving-related factors, and the incidence of puerperal metritis in postpartum dairy cows. *J. Dairy Sci.* 90(6), 2804-2814.
- Bruun, J., Ersboll, A.K., Alban, L., 2002. Risk factors for metritis in Danish dairy cows. *Prev Vet Med* 54(2), 179-190.
- Collins, J.A., Fauser, B.C.J.M., 2005. Balancing the strengths of systematic and narrative reviews. *Human Reproduction Update* 11(2), 103-104.
- Deguaillume, L., Geffré, A., Desquilbet, L., Dizien, A., Thoumire, S., Vornière, C., Constant, F., Fournier, R., Chastant-Maillard, S., 2012. Effect of endocervical inflammation on days to conception in dairy cows. *J. Dairy Sci.* 95(4), 1776-1783.
- Dohoo, I.R., Martin, W., Stryhn, H., 2003. Veterinary epidemiologic research. AVC Inc., Charlottetown, Prince Edward Island, Canada.
- Drillich, M., Beetz, O., Pfützner, A., Sabin, M., Sabin, H.J., Kutzer, P., Nattermann, H., Heuwieser, W., 2001. Evaluation of a systemic antibiotic treatment of toxic puerperal metritis in dairy cows. *J. Dairy Sci.* 84(9), 2010-2017.
- Drillich, M., Klever, N., Heuwieser, W., 2007a. Comparison of two management strategies for retained fetal membranes in small dairy farms in Germany. *J. Dairy Sci.* 90(9), 4275-4281.
- Drillich, M., Mahlstedt, M., Reichert, U., Tenhagen, B.A., Heuwieser, W., 2006a. Strategies to improve the therapy of retained fetal membranes in dairy cows. *J. Dairy Sci.* 89(2), 627-635.
- Drillich, M., Pfützner, A., Sabin, H.J., Sabin, M., Heuwieser, W., 2003. Comparison of two protocols for the treatment of retained fetal membranes in dairy cattle. *Theriogenology* 59, 951-960.
- Drillich, M., Reichert, U., Mahlstedt, M., Heuwieser, W., 2006b. Comparison of two strategies for systemic antibiotic treatment of dairy cows with retained fetal membranes: Preventive vs. Selective Treatment. *J. Dairy Sci.* 89(5), 1502-1508.
- Drillich, M., Voigt, D., Forderung, D., Heuwieser, W., 2007b. Treatment of acute puerperal metritis with flunixin meglumine in addition to antibiotic treatment. *J. Dairy Sci.* 90(8), 3758-3763.
- Dubuc, J., Duffield, T.F., Leslie, K.E., Walton, J.S., LeBlanc, S.J., 2010a. Definitions and diagnosis of postpartum endometritis in dairy cows. *J. Dairy Sci.* 93(11), 5225-5233.

Dubuc, J., Duffield, T.F., Leslie, K.E., Walton, J.S., LeBlanc, S.J., 2010b. Risk factors for postpartum uterine diseases in dairy cows. *J. Dairy Sci.* 93(12), 5764-5771.

Dubuc, J., Duffield, T.F., Leslie, K.E., Walton, J.S., LeBlanc, S.J., 2011. Effects of postpartum uterine diseases on milk production and culling in dairy cows. *J. Dairy Sci.* 94(3), 1339-1346.

Elkjær, K., 2012. Reproduction in the post partum dairy cow - influence of vaginal discharge and other possible riskfactors. Ph.D. Thesis. Science and Technology. Aarhus University, Denmark.

EMA, 2012. Choice of control group in clinical trials.

http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002925.pdf. Assessed 18-6-2012.

Fourichon, C., Seegers, H., Bareille, N., Beaudeau, F., 1999. Effects of disease on milk production in the dairy cow: a review. *Preventive Veterinary Medicine* 41(1), 1-35.

Fourichon, C., Seegers, H., Malher, X., 2000. Effect of disease on reproduction in the dairy cow: a meta-analysis. *Theriogenology* 53(9), 1729-1759.

Gautam, G., Nakao, T., 2009. Prevalence of urovagina and its effects on reproductive performance in Holstein cows. *Theriogenology* 71(9), 1451-1461.

Gautam, G., Nakao, T., Koike, K., Long, S.T., Yusuf, M., Ranasinghe, R.M.S.B., Hayashi, A., 2010. Spontaneous recovery or persistence of postpartum endometritis and risk factors for its persistence in Holstein cows. *Theriogenology* 73(2), 168-179.

Gautam, G., Nakao, T., Yusuf, M., Koike, K., 2009. Prevalence of endometritis during the postpartum period and its impact on subsequent reproductive performance in two Japanese dairy herds. *Animal Reproduction Science* 116(3-4), 175-187.

Gilbert, R.O., Shin, S.T., Guard, C.L., Erb, H.N., Frajblat, M., 2005. Prevalence of endometritis and its effects on reproductive performance of dairy cows. *Theriogenology* 64(9), 1879-1888.

Gorzecka, J., Friggens, N.C., Ridder, C., Callesen, H., 2011. A universal index of uterine discharge symptoms from calving to 6 weeks postpartum. *Reproduction in Domestic Animals* 46(1), 100-107.

Goshen, T., Ben-Gera, J., Koren, O., Bdolah-Abram, T., and Elad, D. The effects of bovine necrotic vulvovaginitis on reproductive and production performance of Israeli 1st calf heifers. *Theriogenology* 77(6), 1178-1185.

Goshen, T., Shpigel, N.Y., 2006. Evaluation of intrauterine antibiotic treatment of clinical metritis and retained fetal membranes in dairy cows. *Theriogenology* 66 (9), 2210-2218.

Habicht, A., 2011. *Vurder selv evidens* (in Danish). Munksgaard, Denmark. ISBN: 978 87 628 1111 9

Hoedemaker, M., 1998. Postpartal pathological vaginal discharge: to treat or not to treat? *Reproduction in Domestic Animals* 33(3-4), 141-146.

Janowski, T., Zdunczyk, S.M.E.S., 2001. Combined GnRH and PGF2 alpha application in cows with endometritis puerperalis treated with antibiotics. *Reproduction in Domestic Animals* 36(5), 244-246.

Kasimanickam, R., Duffield, T.F., Foster, R.A., Gartley, C.J., Leslie, K.E., Walton, J.S., Johnson, W.H., 2004. Endometrial cytology and ultrasonography for the detection of subclinical endometritis in postpartum dairy cows. *Theriogenology* 62(1-2), 9-23.

Kasimanickam, R., Duffield, T.F., Foster, R.A., Gartley, C.J., Leslie, K.E., Walton, J.S., Johnson, W.H., 2005. The effect of a single administration of cephalixin or cloprostenol on the reproductive performance of dairy cows with subclinical endometritis. *Theriogenology* 63(3), 818-830.

Kristensen, E., Dueholm, L., Vink, D., Andersen, J.E., Jakobsen, E.B., Illum-Nielsen, S., Petersen, F.A., Enevoldsen, C., 2006. Within- and across-person uniformity of body condition scoring in Danish holstein cattle. *J. Dairy Sci.* 89(9), 3721-3728.

Kristensen, E., Ostergaard, S., Krogh, M.A., Enevoldsen, C., 2008. Technical indicators of financial performance in the dairy herd. *J. Dairy Sci.* 91(2), 620-631.

Krogh, M.A., 2012. Management of data for herd health performance measurements in the dairy herd. Ph.D. Thesis. Faculty of Health and Medical sciences, University of Copenhagen, Denmark .

Lastein, D.B., 2012. Herd-specific randomized trial - an approach for effect evaluation in a dairy herd health management program. Ph.D. Thesis. Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.

Lastein, D., Vaarst, M., 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 Veterinaria Scandinavica* 51(1), 36.

LeBlanc, S., 2010. Assessing the association of the level of milk production with reproduction in dairy cattle. *J. Reprod Dev.* 56, Suppl: 1-7

LeBlanc, S.J., Duffield, T.F., Leslie, K.E., Bateman, K.G., Keefe, G.P., Walton, J.S., Johnson, W.H., 2002a. Defining and diagnosing postpartum clinical endometritis and its impact on reproductive performance in dairy cows. *J. Dairy Sci.* 85(9), 2223-2236.

LeBlanc, S.J., Duffield, T.F., Leslie, K.E., Bateman, K.G., Keefe, G.P., Walton, J.S., Johnson, W.H., 2002b. The effect of treatment of clinical endometritis on reproductive performance in dairy cows. *J. Dairy Sci.* 85(9), 2237-2249.

LeBlanc, S.J., 2008. Postpartum uterine disease and dairy herd reproductive performance: A review. *Vet J* 176(1), 102-114.

Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, et al. (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med* 6(7): e1000100. doi:10.1371/journal.pmed.1000100

lr.dk, 2012. Vejledning til kliniske registreringer [in Danish]. http://www.landbrugsinfo.dk/Kvaeg/Filer/nysr_vej_klinisk.pdf . Assessed 30-7-2012.

McDougall, S., Macaulay, R., Compton, C., 2007. Association between endometritis diagnosis using a novel intravaginal device and reproductive performance in dairy cattle. *Animal Reproduction Science* 99(1-2), 9-23.

Melendez, P., McHale, J., Bartolome, J., Archbald, L.F., Donovan, G.A., 2004. Uterine involution and fertility of holstein cows subsequent to early postpartum PGF2 [alpha] treatment for acute puerperal metritis. *J. Dairy Sci.* 87(10), 3238-3246.

Nir, O., 2008. The multifactorial approach to fertility problems in dairy herds. *WBC XXV. Hungarian veterinary journal. Suppl 1*, 77-81.

Pleticha, S., Drillich, M., Heuwieser, W., 2009. Evaluation of the Metricheck device and the gloved hand for the diagnosis of clinical endometritis in dairy cows. *J. Dairy Sci.* 92(11), 5429-5435.

Runciman, D.J., Anderson, G.A., Malmo, J., 2009. Comparison of two methods to detecting purulent vaginal discharge in post partum cows and effect of uterine cephalixin on reproductive performance. *Australian veterinary Journal* 87(9), 369-378.

Runciman, D.J., Anderson, G.A., Malmo, J., Davis, G.M., 2008a. Effect of intrauterine treatment with cephalixin on the reproductive performance of seasonally calving dairy cows at risk of endometritis following periparturient disease. *Australian veterinary Journal* 86(7), 250-258.

Runciman, D.J., Anderson, G.A., Malmo, J., Davis, G.M., 2008b. Use of postpartum vaginoscopic (visual vaginal) examination of dairy cows for the diagnosis of endometritis and the association of endometritis with reduced reproductive performance. *Australian veterinary Journal* 86(6), 205-213.

Sheldon, I.M., Lewis, G.S., LeBlanc, S., Gilbert, R.O., 2006. Defining postpartum uterine disease in cattle. *Theriogenology* 65(8), 1516-1530.

Sheldon, I.M., Noakes, D.E., 1998. Comparison of three treatments for bovine endometritis. *Vet Rec.* 142(21), 575-579.

Sheldon, I.M., Noakes, D.E., Rycroft, A.N., Dobson, H., 2004. Effect of intrauterine administration of oestradiol on postpartum uterine bacterial infection in cattle. *Animal Reproduction Science* 81(1-2), 13-23.

Sheldon, I.M., Cronin, J., Goetze, L., Donofrio, G., Schuberth, H.J., 2009. Defining postpartum uterine disease and the mechanisms of infection and immunity in the female reproductive tract in cattle. *Biology of Reproduction* 81(6), 1025-1032.

Sheldon, I.M., Williams, E.J., Miller, A.N.A., Nash, D.M., Herath, S., 2008. Uterine diseases in cattle after parturition. *The Veterinary Journal* 176(1), 115-121.

Sloth, K.H., Trinderup, M., Ancker, S., Østergaard, S., 2008. Børscorer og behandling for børbetændelse i NySR [in Danish]. Kvæginfo nr.1912. https://www.landbrugsinfo.dk/Kvaeg/Sundhed-og-dyrevelfaerd/sundhedsraadgivning/Sider/Boerscorer_og_behandlinger_for_boerbetetae.aspx. Assessed 20-09-2012.

Smith, B.I., 1998. Comparison of various antibiotic treatments for cows diagnosed with toxic puerperal metritis. *J. Dairy Sci.* 81(6), 1555-1562.

Smith, B.I., Risco, C.A., 2002a. Therapeutic and management options for postpartum metritis in dairy cattle. *Compendium* 24(10), 92-100.

Smith, B., Risco, C., 2002b. Predisposing factors and potential causes of postpartum metritis in dairy cattle. *Compendium* 24(9), 74-80.

Thompson, 2011. Thompson's special veterinary pathology. Mosby, St.Louis, USA. ISBN:

Thorpe, K.E., Zwarenstein, M., Oxman, A.D., Treweek, S., Furberg, C.D., Altman, D.G., Tunis, S., Bergel, E., Harvey, I., Magid, D.J., Chalkidou, K., 2009. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *Journal of Clinical Epidemiology* 62(5), 464-475.

Vaarst, M., Paarup-Laursen, B., Houe, H., Fossing, C., Andersen, H.J., 2002. Farmers' choice of medical treatment of mastitis in Danish dairy herds based on qualitative research interviews. *J. Dairy Sci.* 85(4), 992-1001.

Wittrock, J.M., Proudfoot, K.L., Weary, D.M., von Keyserlingk, M.A.G., 2011. Short communication: Metritis affects milk production and cull rate of Holstein multiparous and primiparous dairy cows differently. *J. Dairy Sci.* 94(5), 2408-2412.

3.4 Diagnostic procedures and medical treatments for bovine genital disease in Denmark - a qualitative analysis of the potential for implementing herd-specific randomized trials in a herd health management program

Manuscript III

D. B. Lastein^a, Mette Vaarst^b & C. Enevoldsen^a

^aDepartment of Large Animal Sciences

Faculty of Health and Medical Sciences

University of Copenhagen

Grønnegårdsvej 2, DK-1870 Frederiksberg C

Denmark

^bDepartment of Animal Health

Welfare and Nutrition

Faculty of Agricultural Sciences

Research Centre Foulum

University of Aarhus

P.O. 50, DK-8830 Tjele

Denmark

Diagnostic procedures and medical treatments for bovine genital disease in Denmark - a qualitative analysis of the potential for implementing herd-specific randomized trials in a herd health management program

D. B. Lastein ^{a*}, Mette Vaarst ^b & C. Enevoldsen ^a

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

^bDepartment of Animal Health,Welfare and Nutrition, Faculty of Agricultural Sciences, Research Centre
Foulum, University of Aarhus, P.O. 50, DK-8830 Tjele, Denmark

*Corresponding author: dorte.bay@gmail.com (Lastein, D.B.): phone 0045-20641151

Abstract

Background

Decision-making in a Herd Health Management Program (HHMP) should be supported by valid recommendations for diagnostic procedures, treatment thresholds, and treatment protocols. Genital diseases in a Danish HHMP are diagnosed with systematic clinical examinations of all or a majority of cows, 5-21 days postpartum, including a vaginal discharge score (VDS). This study addresses the potential of combining this systematic approach to diagnosis with a practical herd-specific trial approach to evaluate effect of interventions. Inferences from these trials might potentially support clinical decision-making in regard to treating genital diseases. This concept may also be applied to other diseases and management interventions.

Results

Based on semi-structured interviews with and observations of 12 veterinarians in the HHMP we found coherent patterns, but a wide range of procedures, actions, and perceptions among veterinarians. Action patterns were linked to the individual veterinarian's perception of focus point, aim of treatment aim, and rationale for VDS scoring. With a tool designed to structure trial development (PRECIS), we linked the empirical data describing actions and perceptions to a pragmatic-explanatory continuum to identify conceptual trial designs that had potential for implementation in cattle practice and to appropriate practical trial designs. The results indicated the potential for implementing trials with pragmatic designs. That is, estimates of treatment effectiveness were more informative in a HHMP context than evaluating treatment efficacy.

Conclusion

Due to literature discrepancies and shortages of trials conducted before 21 days postpartum, little scientific evidence exists to justify current HHMP procedures. Clinical field trials within the pragmatic-explanatory continuum should be carefully adapted to individual veterinarians/practices and conducted by highly motivated participants to ensure success. We suggest three types of trial designs: (1) a moderately explanatory within-herd trial with a clinical focus, (2) a pragmatic within-herd trial with a production focus, and (3) a pragmatic multi-herd/within-practice trial with a production focus. Furthermore, we propose a practical approach, non-blinded treatment, group allocations by ear-tag number, and trial designs that include active or negative-controlled parallel groups, modified cross-over treatments, or factorial groups, depending on the presently applied protocols in the herd. Superimposing the practical design with a group-sequential analysis and criteria for a clinical cure could provide the potential of 'early stopping'. An 'intention to treat' analysis (superiority testing) was proposed to provide a pragmatic, conservative estimate of the effect of changing standard procedures.

Keywords

Bovine, dairy, genital, trial, effectiveness, herd health management, diagnosis, metritis, pragmatic trial

Introduction

Decision-making related to medical treatment of dairy cows in veterinary practice and HHMPs requires consideration of both animal welfare and cost-benefit issues. Numerous context-related, farmer-specific, and often, implicit concerns influence the choice of the threshold for treatment (Vaarst et al., 2002). The veterinarian working within a HHMP could potentially benefit from a systematic approach for selecting a treatment threshold and/or a treatment protocol. The selection could be based on validated, explicit criteria, and the estimated effect on welfare or production. These choices may serve the joint interests of both animals and humans by minimizing unnecessary medical treatments and avoiding potential antibiotic resistance. The present study addresses the potential of combining a systematic approach to diagnosis with clinical decision-making, based on a systematic evaluation of the effect of treatment applied in veterinary practice.

In the specific case of genital diseases in dairy cows, considerable controversy exists regarding definitions, the need for medical treatments, the choice of diagnostic indicators, and the treatment threshold. Only few randomized clinical trials with untreated control groups have been reported (LeBlanc, 2008). Often, published trials focused on the effects of treatment on reproductive performance; only rarely have they

focused on milk production or animal welfare. Consequently, validated practical recommendations are sparse for medical treatment of genital diseases in veterinary practice, especially in the early postpartum period (Lastein, 2012). This also applies to an extended Danish HHMP (Ministry of Food and Agriculture and Fisheries, 2006). One focus of this HHMP is early detection and treatment of genital diseases (5-21 days postpartum), based on intended systematic clinical examinations of all cows in the HHMP. However, the diagnostic criteria and the applied treatment protocols were never evaluated in a trial setting with untreated control groups. Recently, retrospective observational studies evaluated the consequences of the HHMP examination procedures on milk production and reproduction, and the results called into question the HHMP procedures (Elkjær, 2012; Krogh, 2012).

The following quotation formulates very precisely the challenges faced by the practicing veterinarian:

“In the absence of evidence as to the efficacy and safety of animal-health products and procedures derived from controlled trials, practitioners are left in the unenviable position of making decisions about their use based on extrapolation of data from studies carried out under artificial (laboratory) conditions or on their own limited, uncontrolled experience” (Dohoo et al., 2003).

Moreover, strategies extrapolated in this manner might not be applicable or effective in a specific herd context. These problems increase the uncertainty for the practicing veterinarian. We hypothesized that one way to bridge this knowledge gap, in support of clinical decision-making and evaluations of treatment effects in modern dairy production, might be to combine locally collected data and herd analyses with experience and general knowledge – an approach quite similar to that of ‘evidence based medicine’ (Sackett et al., 1996).

A previous analysis was performed to determine veterinarians’ motivations for collecting data in the Danish HHMP. The results indicated that an individual veterinarian’s perceptions of data and data quality interacted with the veterinarian’s method of working, which could be classified as experience-based or analysis-based (Lastein et al., 2009). Therefore, we reasoned that a veterinarian’s personal involvement in an experiment or clinical field trial could potentially validate specific diagnostic procedures or protocols; subsequently, this experience could promote changes in patterns of action (e.g., changes in treatment criteria or treatment protocols). Furthermore, we proposed that the development and implementation of a financially uncompensated, controlled, and practically feasible trial design could be integrated into the HHMP. This would open up the possibility of testing treatment criteria and treatment regimes that are specifically applicable to the veterinarian or the herd (Kristensen, 2008). To develop a trial design suitable for veterinarians working in the field, we must construct a design that is both practical and adjustable. To maximize the feasibility of a trial design, the procedures in current use could be identified and integrated into practical parts of the future trial design. Furthermore, the primary goals and aims of the trial should be

consistent with the needs of the end-users (veterinarians and farmers). That is, “does the trial address the relevant question?”

Trial theory, as described by Thorpe and co-workers emphasizes the recognition of a multidimensional continuum between explanatory and pragmatic trials. We will refer these terms as a trial’s conceptual design. Explanatory trials are set up to assess intervention effects (like efficacy) under controlled conditions. Hence, they show causal biological associations. Pragmatic trials are set up to illustrate the effect (effectiveness) of an intervention in ‘the real world’. That is, to support decision-making in practice, when choosing between different treatment policies (Thorpe et al., 2009). Thorpe et al. (2009) previously developed a trial planning tool (PRECIS) to facilitate a coherent relationship between the aim of a trial and the needs of the end-user in the development of a trial design. We used this tool as a guide to ensure consistency between the aims of the trials and the requirements of participating herds, farmers, and veterinarians. We also used the tool to guide the design of different trial components during the development phase.

By means of qualitative research methodology, the objectives in the study were to:

- Describe and theoretically justify presently applied action patterns related to diagnosis and treatment of bovine genital disease before 21 days postpartum within a Danish HHMP.
- Combine the applied procedures with potential trial designs to exemplify and develop a method for systematic effect evaluation for herd-specific problems in HHMP.

Material and methods

The context of Danish Herd Health Management legislation

Legislation concerning the voluntary dairy HHMP, with mandatory systematic data collection, was introduced in Denmark in 2006. The program aimed to improve the detection and documentation of some important health disorders. Dairy herds were visited weekly/fortnightly, and data were intended to be collected according to a standard procedure, on all or a majority of cows, during specified risk periods. Clinical examinations of all cows (by intention) or a majority of cows (in practice) were performed at drying off and at calving (5-21 days postpartum). Body condition, vaginal discharge, and udder condition were some of the mandatory health indicators. All treatments and scores related to genital diseases were to be recorded according to an official scoring manual for vaginal discharge (called vaginal discharge score = VDS) (Lastein et al., 2009). No distinction between different genital diseases was attempted (e.g., metritis and cervicitis); however, different vaginal lesions were scored separately. No official treatment threshold was linked to the VDS. At the time of this study (2008), Danish legislation restricted the use of antibiotics applied by farmers. Only drugs that were prescribed by a veterinarian to a specific cow with a specific

genital disease could be used, and all treatments had to be recorded at the level of individual cows in a national database. The initial treatment with antibiotics and all hormone treatments for genital disease were to be administered by a veterinarian. Off-label use of registered drugs was illegal. After the time that the study was conducted, the regulations on the use of antibiotics by farmers and the requirements for registration were relaxed. This could influence some aspects of the future implementation of herd-specific randomized clinical field trials in the HHMP, particularly in relation to data management and data quality, as discussed later.

Selection of participants to interview

From the Danish authorities, we obtained a list of veterinarians within three geographical regions that worked with a minimum of two herds within the HHMP. The procedures used to create the list were unknown to the authors. The first twelve participants at the top of the list were phoned, and all agreed to participate in the present study. These veterinarians worked in 2-15 herds (median: 4 herds) within the HHMP, and they had 3-30 years of experience in dairy cattle practice. Only one veterinarian per veterinary practice was included. Anonymity was guaranteed to promote openness and confidentiality.

Participant observation

The first author [DBL] performed observations of veterinarians at work on farms and conducted the interviews, from January to March 2008. The veterinarians were observed for one day at 1-4 herd visits, when they performed practical scoring of cows and medical treatments; in addition, during the visits, the observer recorded conversations with the farmer and veterinarian in the barn and in the car. Notes on observations were recorded during the herd visits.

Qualitative semi-structured research interviews

DBL used semi-structured research interview techniques to interview all the veterinarians about work related to diagnosis and treatment of genital diseases (Kvale, 1994). Each interview (½ hour to 1¼ hour) followed an interview guide. The conversation was directed through the themes in the interview guide, and questions were followed up for clarification or elaboration by the interviewed veterinarian. Details on the interview guide were described in related work (Lastein et al., 2009).

Data management and analysis

All interviews were recorded with a digital voice recorder and transcribed in full length. The transcripts and the notes on observations and discussions were reviewed. From this, we identified a variety of practical diagnostic methods, methods for determining a threshold for treatment, and treatment protocols. Coherent patterns in the described procedures were identified and ranked, based on the principles of

analysis described previously (Lastein et al., 2009). These descriptions of specific procedures were compared to relevant literature about diagnosis and treatment of genital diseases and practical trial design.

Table 1. The 10 domains within the pragmatic-explanatory continuum. These are fictitious extremes within the field trials concerning genital diseases. The observed range of procedures, actions, and perceptions in a Danish herd health management program (HHMP) are illustrated with examples and quotes. Inspired from Thorpe et al., (2009).

Question	Explanatory trial - 'fictitious extremes'	Observed range of present procedures, actions, and perceptions, illustrated by examples and quotes.		Pragmatic trial - 'fictitious extremes'
	Efficacy – can the intervention work?			Effectiveness – does the intervention work when used in practice?
Domain		Most explanatory observed (E _{obs max})	Most pragmatic observed (P _{obs max})	
1. Eligibility criteria (herd and national levels)	One or a few herds, perhaps university/research herds. If, private herds intensive exclusion of herds with non-motivated farmers to reduce drop out of farms	Local recommendations at the herd level. A veterinarian refused to compromise his procedures in the HHM program. He strongly favoured the farmers that appreciated his methods	General recommendations at national level	Multi-centre trial with no specific selection of participating herds. Random samples
2. Flexibility (experimental intervention at the cow level)	Very detailed protocol (diagnostic criteria and treatment); e.g., disease index based on several diagnostic indicators. Restrictions on co-interventions; e.g., cows must not receive other treatment during trial period. Intensive exclusion criteria at the cow level; e.g., treat only cows with high risk of milk yield loss	"We will not treat cows that we cannot cure" (quote from a veterinarian); this was applied in combination with a fluctuating cow dependent threshold	No exclusion criteria -used a simple threshold strategy (treat all cows with VDS above 5 in all herd)	Protocol of intervention resembled the 'usual practice' e.g., 50 ml of penicillin to all cows, despite size differences. No restrictions on co-interventions. No exclusion criteria at the cow level. The most extreme would be a farmer's diagnosis without confirmation from a gynaecological examination
3. and 5. Veterinarian's/farmer's expertise (experimental and comparison interventions)	Intensive training of veterinarians and VDS-calibration before and during trial period. Intensive instruction for farmers	Within one practice, clinical assessments were calibrated, mainly on body condition scoring	A veterinarian expressed little knowledge of the official descriptive scale in the manual	No expertise (experience or training) required of veterinarian or farmer with regard to diagnosis and treatment
4. Flexibility (comparison intervention)	Untreated or a placebo control group	One veterinarian had changed treatment threshold due to the suspicion of treating false positive cows	"We will not try a zero therapy solution" (quote from a veterinarian)	"Usual practice" with an active control group
6. Follow-up intensity (cow and herd levels)	Intensive collection of data to determine clinical outcome; e.g., daily rectal temperature or blood sampling. Intensive evaluation of milk production data in relation to the treated cows.	Follow-up examinations and scoring on all treated cows to assess clinical treatment effect. A veterinarian described his methods for evaluating effect by data inspection (no statistics)	Farmer's observation	No specified follow-up strategy for either clinical or milk production measurements. No interest in the welfare of treated animals from either the veterinarian or the farmer.
7. Primary trial outcome	Short-term outcome, not directly relevant for 'trial end-users'; e.g., blood indicators of infectious disease	Improvement in vaginal discharge within 1-2 weeks post treatment	Milk yield – no specifications on peak or lactational yield was given; reproduction performance not assessed	Long term outcome, e.g., database registration of milk yield and reproduction performance
8. Farmer adherence to protocol (at the cow level) NOTE: Danish farmers applied additional treatments	Detailed verification of data quality. Detailed strategies to improve data quality (reduce non-adherence).	In general, veterinarians assumed that farmers treated cows as prescribed and recorded data correctly in database.	A farmer recounted how he deliberately and illegally prolonged the prescription period and recorded another disease in the database. He claimed that the veterinarian knew and blindly accepted this breach in protocol. A veterinarian allows farmers to collect data within the HHM program (not VDS).	Non-adherence is considered a reality in practice; i.e., no special efforts are made to register cows that do not fit the protocol for whatever reason or verify data quality.
9. Veterinarian's and farmer's adherence to protocol. NOTE: Veterinarians diagnosed and initiated all treatment	Recording of all errors in the protocol. Feedback and improvement strategies are important.	Intended use of both standard herd and practice protocols thought to increase likelihood of adherence	A veterinarian used whatever drug he had in the car on the day of the visit	Veterinarian's and farmer's 'errors' were accepted and were not recorded; e.g., allocating a cow to the wrong treatment; choosing treatment protocols at random
10. Analysis	Exclusions of cows and practitioners, due to non-compliance and non-adherence; e.g., erroneous registration of re-treatments and poor VDS-calibration results. Intention to treat analysis (ITT) and per protocol (PP) analysis possible.	Focus on the individual cow. Desired fast answers. Vaginal discharge scoring used as a decision making tool.	Focus on the herd or national strategies. Some acceptance of waiting for long periods before effects can be evaluated.	All cows included, irrespective of non-compliance to protocol; intention to treat analysis (ITT).

VDS - vaginal discharge score obtained between days 5 to 21 postpartum. Ordinal scale [0;9]

ITT - intention to treat (all available data included in analysis, except loss to follow up)

PP – per protocol (only per protocol data included in analysis)

The research tool known as PRECIS was used to frame the requirements of the ‘conceptual trial design’ and describe the context in which we seek to implement a trial design. PRECIS describes 10 domains that represent different areas related to trial design that are plotted in a ‘spider graph’, which indicates the pragmatic-explanatory continuum (Fig. 1). The centre of the graph represents the extremes of explanatory trials, and the outer circumference represents the extremes of pragmatic trials. A detailed description of the domains is presented elsewhere (Thorpe et al., 2009).

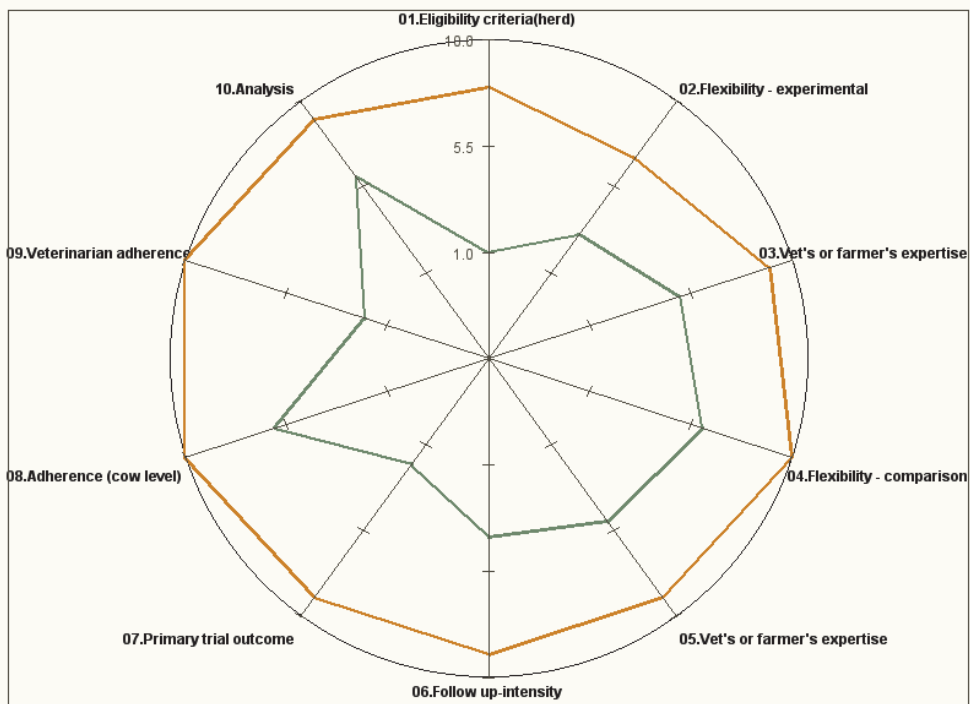


Figure 1. PRECIS spider graph with 10 domains (01-10). The observed range of procedures among 12 interviewed Danish veterinarians is plotted in the pragmatic-explanatory continuum (‘angular’ lines). The outer lines represent the observed pragmatic extreme, Pobs, max. The inner lines represent the observed explanatory extreme, Eobs, max. Domain 3 and domain 5: Veterinarian/farmer expertise on experimental and comparison intervention, respectively are identical in this case.

We used the tool in this study to illustrate how the different action patterns in veterinary practice are in line with the domains in the continuum. First, we described some aspects of ‘extreme pragmatic and extreme explanatory’ clinical field trial approaches designed to evaluate the effects of treatment for genital diseases (Table 1 – left and right column). This description of the extremes helped us to delineate the spectrum of possible designs relevant for the HHMP context, being the identified procedures in the empiric

material. To identify the range of the identified procedures within the domains we subsequently generated supplementary keywords and questions related to each domain that were used to explore the empirical data (Table 2). Then, we identified the range of procedures associated with each domain by searching for the answer of the supplementary questions. We thereby identified the most explanatory procedure/action/perception observed ($E_{\text{obs max}}$) and the most pragmatic procedure/action/perception ($P_{\text{obs max}}$) observed within each domain.

Table 2. Supplementary keywords/questions that were used to identify and link the 10 domains in the PRECIS tool to empirical observations, based on interviews with 12 Danish veterinarians in a herd health management program.

1. Eligibility criteria on the herd level: Is the veterinarian interested in a general or local recommendation? Is the focus point at cow, herd or national level?
2. Flexibility of experimental intervention: How are inclusion and exclusion criteria used in practice?
3. Veterinarian/farmer expertise concerning the experimental intervention: Any use of calibration and training among involved veterinarians on diagnostic procedures? How is knowledge transferred between involved veterinarians/farmers? Is information to herd-specific diagnostic and treatment procedures shared among all veterinarians in a practice that visit the herd?
4. Flexibility of comparison interventions: Any consideration of alternative strategies for diagnostic procedures and treatment, other than those applied? Any doubts about the effects of treatment – did this influence the choice of treatment protocol or treatment threshold?
5. Veterinarian/farmer expertise concerning the comparison intervention: Same as 3.
6. Follow-up intensity: ‘Clinical follow-up examination’ of treated cows to ensure clinical cure? Which procedures are used? Use of observational data analysis to evaluate effect?
7. Primary trial outcome: What is the primary aim of the currently applied treatment?
8. Farmer compliance: Any consideration of current data quality and acceptance of this status?
9. Veterinarian’s/farmer’s adherence to protocol: Current use of standard procedures at the herd level, motivation to be systematic (e.g., to follow a protocol)
10. Analysis: What focus point is dominant? Which time horizon is acceptable for a trial?

The observed extremes of the range were exemplified in descriptions and quotations. To illustrate the ranges graphically on a spider graph, we had to transform the observed veterinary procedures within each domain onto a visual analogue scale with the explanatory extreme on one end (score=0) and the pragmatic extreme on the other end (score=10) (Fig. 2). This rating was then plotted on the spider graph. The

resulting spider graphs were used to illustrate the range of the currently applied procedures within the continuum and to identify potential inconsistencies among the veterinarians' actions, perceptions, and requirements for trial design.



Domain: 2. Flexibility - experimental

$E_{obs\ max} = 3.7$: The use of multiple diagnostic criteria to determine treatment threshold resemble the use of a 'detailed disease index' (assuming same criteria used every time) and also contributes to exclusion of cows with less probability of reduced production

$P_{obs\ max} = 7.8$: The use of a fixed treatment threshold in all herds due to ease and practicality

Figure 2. A visual analogue scale used to score the observed range of procedures and actions. Procedures were rated on a scale from completely explanatory (0 cm) to completely pragmatic (10 cm). This example shows the rating for Domain (2): Flexibility. $E_{obs,max}$ represents the most extreme explanatory procedure; $P_{obs,max}$ represents the most extreme pragmatic procedure used among 12 interviewed Danish veterinarians. The ratings were later integrated into the PRECIS spider graph.

Results

Description of applied procedures

The main results of the 12 interviews with the veterinarians in the HHMP are presented in Table 3. The procedures identified in the transcripts were listed from left to right to represent the range of the applied procedures. In this context, "data quality" refers to the extent to which data collection was consistent and whether there was sufficient comparability between groups of cows for valid quantitative trial analysis. The qualitative analysis of transcripts identified some interactions between the focus point, the personal reasons for scoring, the aim of treatment, and the practical use of the VDS. These findings indicated that different focus points were reflected in the veterinarians' reasons for using the VDS. This issue was discussed in detail by Lastein and co-authors (2009). These interactions were not evident in all veterinary actions; therefore, they were not applicable to all the described aspects of the procedures.

Table 3. Main results from 12 semi-structured interviews intended to describe the range of diagnostic and treatment procedures related to genital examinations and vaginal discharge scores (VDS, measured 5-21 days postpartum) in a Danish herd helath management program

	Range of procedures
Focus point	cow > herd >veterinary practice >national
Aim of treatment	achieve clinical cure* and welfare > reduce secondary disease>improve milk production/reproduction> reduce non beneficial treatments/antibiotic usage
Reason for scoring	select cows for treatment > evaluate the need of additional treatment > monitor the herd > standardize work/data >follow legislation
Diagnostic indicators	multiple criteria** > few criteria (VDS + RT) > one criteria (VDS)
Treatment threshold	fluctuating: VDS range 4-7 and cow dependent > fluctuating: VDS range 4-7 and herd dependent > fixed: VDS ≥ 5 (odour)
Follow-up procedure (including timing) ***	Clinical effects: farmer control**** > veterinarian occasional control > systematic control (1-2 weeks post 1 st treatment) Production effects: ask farmer about expected yield > check yield/disease data – no statistics > intentions to conduct herd trials
Threshold for repeated treatment (non-cure)	Veterinarian's examination: same as initial examination, but generally, pus in discharge is acceptable on second examination (e.g., VDS=3-4, accepted as cured)
Treatment regimens	case variable (e.g., increasing number of days of antibiotic treatment for higher VDS, certain parities, or later in lactation) > standard herd procedure > standard practice procedure
Medicine – type, dosage etc.	Antibiotics : - narrow or broad spectrum - duration: 1- 5 days - combination of intramuscular and intrauterine treatment (including off-label use of injected medicine for intrauterine use) Prostaglandins: - often in combination with antibiotics - different analogues - one injection on day of visit - varying decisions on use in relation to presence of CL, lactation stage, and lactation number Corticosteroids/NSAID: - either based on indication or a standard procedure Non-cured: - same or different antibiotics than 1 st treatment - prostaglandins
When to change standard treatment protocol?	Herd measures: few non-cured cows / few weeks of 'non successful treatment' > many non-cured cows/several months of 'non successful treatment' >never (total focus on preventive measures) Decision support: 'gut feeling'> bacteriology > national guidelines > personal ideology
Use of data	None > seldom, due to lack of time or skills > regularly

* Some veterinarians worked under the assumption that a strong association existed between reduction of clinical signs (lower vaginal discharge score [VDS] in the weeks after treatment) and improved milk production and reproduction.

** Criteria included in addition to VDS: rectal temperature (RT), general condition, observed vs. expected yield, ketosis, rumen fill.

*** Assessing treatment effect based on clinical cure/non cure was very dependent on the veterinarian's perception of time investment and value of control procedure.

**** When the farmer's observations were used for assessing effects of treatment, the cows in question would be presented for the veterinarian's examination at next visit, due to legislation on drug usage, or the farmer could be tempted to use drugs illegally (example described by a farmer in the barn during a herd visit).

Procedures related to the clinical examination

The practical examination procedures differed among veterinarians with regard to the cleaning procedure (washing or alcohol towel) and the (re)use of examination gloves between cows. The HHMP manual suggested a vaginal or rectal examination at 5-21 days postpartum to evaluate the genital health. Only few veterinarians in this study used a rectal examination to supplement the vaginal examination. No veterinarians used technical equipment (e.g., ultrasonography), and bacteriological testing in a laboratory

was only rarely mentioned as supplement to the clinical examination. Consequently, those technical and laboratory techniques were considered irrelevant in our study context.

Diagnostic indicators and treatment thresholds

The HHMP manual suggested the use of the VDS (range, 0 - 9) for rating the severity of disease. A VDS of 5 or above indicated a pathological odour. The use of other diagnostic indicators for genital disease was voluntary. All interviewed veterinarians applied the VDS, but they used the result in different ways to determine a threshold for treatment. Some veterinarians consistently used a VDS of 5 and above as the only diagnostic indicator. We called this a 'fixed treatment threshold' (Table 3). Other veterinarians combined a 'fixed' or a 'fluctuating' VDS with other cow-level criteria, including temperature, rumen fill, general appearance of the individual cow, or implicit herd-level indicators, like the apparent level of disease surveillance provided and the general immune status. An interaction between the farmer's level of disease surveillance and the veterinarian's decision making process was exemplified by a veterinarian's description of his action patterns in farms, which, in his opinion, had low management levels: 'I decreased the treatment threshold in these herds, so the metritic cows were not left in the corner to die'.

Treatment protocols

Two different perspectives on treatment protocols were identified in this study: (1) The treatment decision varied depending on the case, at the cow and/or herd level; for instance, a higher VDS was treated with a longer duration of antibiotics. (2) A standard herd or standard practice procedure for treatment was applied to all cases. The treatment protocols also differed in the combinations of parenteral and intrauterine antibiotic treatments. Differences in action patterns were identified regarding the timing and reasoning involved in changing a protocol when treatment failed. These different action patterns could be linked to potentially relevant practical trial designs, because they represented different treatment strategies (discussed below).

The choices of protocol and medicine were governed by practicalities, tradition, belief in or experience with an effect, financial costs (initial price and withdrawal period), concern about antimicrobial resistance, recommendations from the veterinary authorities, and, to a limited degree, scientific literature and the individual's herd-specific analysis. A detailed evaluation of the individual treatment protocols was beyond the scope of this study.

Evaluation of effect

The range of procedures used to assess the effect of treatment could be divided into clinical- and production-focused evaluations. No official recommendations were given on whether or when veterinarians should perform post treatment examinations to evaluate the clinical treatment effects. The

clinical post treatment evaluation (subsequently called the follow-up procedure) ranged from a systematic vaginal examination of all treated cows, conducted by the veterinarian at 1-2 weeks post treatment, to the farmer's observation of a clinical non-cure, based on an individual's perception of clinical normality. In a herd with the latter follow-up procedure, an interaction was identified between the farmer's observations/actions and the quality of the data records. One farmer described how he deliberately gave antibiotics prescribed for claw disease to treat cows with clinical metritis, when he considered the initial veterinary treatment was not successful. He recorded these treatments as claw treatments to circumvent the legislation, because it was illegal for a farmer to initiate treatment of genital diseases.

Evaluations of the effects on milk or reproduction performance were rarely performed among the interviewed veterinarians, and no veterinarian included any statistical methods for evaluation. Evaluation of a key performance indicator (e.g., expected vs. observed milk yield) was described. Nevertheless, some veterinarians clearly stated that they were convinced that there was a positive effect of treatment on both reproduction and milk yield, based on their personal experience (a subjective, qualitative judgment).

Requirements for trial design development

Conceptual and practical trial design

The main results associated with the requirements for trial design and PRECIS tool applications are shown in Table 1 (middle columns). The range of the identified procedures, actions, and perceptions related to the 10 domains are presented graphically in Figure 1. The results showed that the procedures covered a wide spectrum of both explanatory and pragmatic elements. This range included consistency and inconsistency among the domains for individual veterinarians. These findings implied that trial designs should be individually adapted to each veterinarian or practice. The process of extracting the essence of veterinarian's individual procedures to suit a design will require a thorough understanding of all trial design elements, including the elements that cannot be included and the underlying rationale for this. The widest ranges of procedures were observed in domains 7 and 9. An example of how a broad range of procedures can affect trial design is given below for those domains.

Domain 7; Primary trial outcome: The explanatory end of the range described the outcome as a clinical cure, and the pragmatic end of the range described the outcome as a reduction in the loss of milk production due to disease. According to the HHMP perspective, the pragmatic approach appeared to be highly relevant. That is, a clinical cure was less relevant when the outcome was not directly associated with the risk of culling (welfare issue) or with reduced production/reproduction performance.

Domain 9; Veterinarian's/farmer's adherence to protocol: The explanatory end of the range described a standard practice protocol, and the pragmatic end described an individual, case-dependent protocol. In the

context of the HHMP and the trial, systematic data collection and adherence to a protocol (explanatory approach) would increase the comparability between intervention groups.

These two descriptions showed that both explanatory and pragmatic trial design elements can be relevant for trials designed to be implemented in clinical veterinary practice as an integrated part of a HHMP. Overall, the requirements of the end-users (veterinarians, practices, and farmers) must be considered in the conceptual design of a HHMP trial; thus, the design should include pragmatic elements to ensure that the overall aim, which is the effectiveness of procedures, is relevant for the ultimate end- user (the farmer) in real life settings. However, the fundamentals of trial theory must be acknowledged; that is, the most 'explanatory inspired' data collection methods and the most comparable intervention groups will provide the most reliable estimates of causal biological effects. Also, the 'intention to treat' principle should be acknowledged; that is, the most 'pragmatically inspired' data collection methods (e.g., high levels of non-compliance, divergence in inclusion criteria, and protocols that represent the 'real world' situation) will provide the most unreliable estimates of the biological causal effects. However, the latter will also provide the most reliable estimates of the effects of the decisions made (and the intentions behind them) in practice.

Clearly, veterinarians and farmers that participate in a HHMP trial would be interested in the herd-specific trial result. However, it became evident during the interviews that some veterinarians were eager to use highly standardized procedures across cows and herds. Consequently, it was relevant to implement and analyse multi-herd trials that could provide reasonable general estimates of effects. Based on these considerations and the results obtained with the PRECIS tool assessment, we propose that the following three overall trial designs could be used as a template for further development of trials:

1. Explanatory, within-herd clinical field trial with a focus on individual cows and a clinical cure;
2. Pragmatic, within-herd clinical field trial with a focus on the herd and milk production or reproduction performance;
3. Pragmatic, multi-herd or within-practice clinical field trial with a focus on milk production or reproduction performance.

In this study, all three of these types of trials were referred to generally as a 'clinical field trial'. We do deliberately not include the term randomized as the practical circumstances of trials in HHMP might hinder the 'optimal' randomization procedure as discussed below. The procedures for follow-up treatments and the actions taken after changing the standard treatments (Table 3) were linked to the 'practical' trial design, as described by the European Medicines Agency (EMA, 2001). We use the term 'practical' trial design to describe the structural design e.g., parallel group, factorial, etc. For instance a parallel group

design would be well suited for veterinarians that used the same protocol at the first and subsequent treatments (with the option of stratifying on parity and/or on the retained placenta). However, some veterinarians changed antibiotics when they deemed a non-cure after the initial treatment. In that case, the procedures would be easier to test in a cross-over design, which permits sequential combinations of different medical treatments. However, the full implementation of a classical cross-over design would require long disease periods and washout periods between treatments; therefore, that design was not suited for this context. Instead, a 'modified' cross-over design could be developed with no washout period between the initial and follow-up treatments. Other veterinarians combined different simultaneous medical treatments; e.g., parenteral and intrauterine antibiotics or antibiotics and prostaglandin. In that case, a factorial design might be applicable. Moreover, some veterinarians applied procedures that were more flexible; e.g., they changed a standard treatment after finding a few non-cured cows. This approach could be adapted to the principles of 'adaptive allocation procedures'; thus, the outcome (success or failure, e.g., clinical cure or non-cure) for each cow would determine which treatment the next cow should be given. This procedure ensures that most cows receive a superior treatment (Bjerkset et al., 1997).

Some veterinarians explicitly stated that, in a hypothetical trial situation, they would like the opportunity to interfere when the clinical signs of disease did not resolve quickly. This requirement could be met by a clause in the protocol for an 'early escape' (EMEA, 2001), where individual cows could be discontinued, and/or a sequential design could be implemented with stopping rules included throughout the trial (Whitehead, 1992). Simple trial designs, like a parallel group design, might also be superimposed by a sequential design with clinical stopping rules (e.g., a double triangular test). We have earlier demonstrated by simulation that this procedure could potentially reduce the required sample size (Lastein and Enevoldsen, 2009), as the theory also implies (Whitehead, 1992). Additionally, the interviewed veterinarians often desired to use multiple outcomes to assess the effect of treatment. This goal could be achieved by designing stratified studies with planned secondary analyses (e.g., does the effect of metritis treatment depend on occurrence of retained placenta), analyse multiple endpoints and/or by applying multivariate analytical methods.

Diagnostic protocol

Some interviewed veterinarians expressed concern about what they called 'unnecessary medical treatments' and antibiotic resistance. These veterinarians often exclusively treated cows with systemic signs of disease. They often used multiple diagnostic criteria and a fluctuating threshold relative to the VDS. In contrast, other veterinarians seemed to be concerned about the risk of production loss. This segment of veterinarians implicitly expressed that it was important not to miss potentially diseased cows. Hence, they prioritized minimizing the loss over the risk of treating false positive cows. These veterinarians

predominantly followed very few explicit criteria (e.g., treatment based on a fixed VDS and standard procedures in all herds). Thus, these two groups reflected different procedures related to the design of the diagnostic protocol. The procedures of the former group of veterinarians were suited to the development of a disease index, based on criteria that were explicitly validated. The latter group used standard criteria that could be directly applied to a trial situation.

Motivation for evaluating the treatment effect and participating in a field trial

A note is warranted on the motivation for participating in clinical field trials. The interviewed veterinarians expressed different rationales for conducting the VDS (Table 3). The reasons ranged from using the VDS for selecting cows for treatment to using the VDS for standardizing the work and data collection to facilitate compliance with legislation. During the interviews, we identified some factors that appeared to motivate the veterinarians to use a systematic approach for evaluating the effect of a treatment. These factors included doubt about the currently applied procedures and strategies, a willingness to test their own theories to challenge themselves academically, as part of 'the academic game', and an understanding or 'trust' in epidemiological methodology. Other factors were identified during the interviews that appeared to demotivate the veterinarians to take a systematic approach. Among the factors that prevent motivation were a conviction of the validity of currently applied procedures, a predominantly disease preventive focus, a general scepticism towards statistical analysis, and stress or a lack of time.

Discussion

Choice of methodology and tools

The overall objective of this study was to describe procedures related to diagnosis and treatment of genital diseases that would provide a basis for developing a practice-based clinical field trial design related to this specific herd problem. We used a qualitative study design with the aim of describing the diversity of concepts, perceptions, and procedures in the words of the interviewees. However, we do not know whether we have sufficiently described the complexity. A larger survey of randomly chosen veterinarians may have provided a more representative sample. However, the aim of sampling in a qualitative study is often to maximize the diversity of the study units. This would have required, a priori, substantial knowledge about factors that affected veterinarians' attitudes and perceptions toward our research question (e.g., age and gender), which could have guided the appropriate selection of veterinarians. Then, the inclusion of extreme views might have provided valuable insight into the diversity of the decision-making process, motivation factors, and current principles and procedures for diagnosis and treatment of genital diseases within the HHMP. In contrast, the aim of a quantitative study is to estimate population parameters, like means, proportions, and variance. Those estimates require a representative sample, where extreme values

would be rare. Due to our limited knowledge of the domain, our sampling approach represented a compromise between these two methods. We restricted the sampling to veterinarians that had practical experience with the HHMP to ensure that, at baseline, they had practical knowledge of the HHMP and had accepted the concept. Consequently, our group was more homogeneous than the general population of veterinarians. Then, geographic criteria were used to ensure broad representation of traditions that might be important in the treatment approach. We could not obtain information about the criteria used to select the list of veterinarians that comprised our sample. Consequently, we cannot claim that we had a representative sample. Therefore, because our study was qualitative, we did not present any quantitative estimates, including means, proportions, and variances.

The final number of interviews (sample size) in qualitative research can be determined sequentially, during the on-going interviews. That is, when a new interview does not give substantial new information, the sample size can be deemed sufficient ('data saturation'). For the present study, the 12 initially planned interviews and observations periods of 2-6 h were considered sufficient. Our sample size was in accordance with guidelines for phenomenological research (10 interviews) (Onwuegbuzie and Leech, 2007). The general rule is that sample size can be relatively low in a homogeneous population (as in our case). In contrast, a small random sample will provide less diversity.

We used the PRECIS tool to examine the diversity of the applied procedures and pinpoint possibilities and limitations of developing trial design suitable for HHMPs. This tool framed our observations within the pragmatic-explanatory continuum of conceptual trial designs. We acknowledge that we would have increased the validity of our conclusions by having the transcripts evaluated by more researchers. Furthermore, the tool was originally designed to be used directly by a research team during the trial planning phase to assess the degree to which degree their decisions regarding design (the 10 domains) aligned with the stated purpose of the trial. In the present study, we showed that the tool could also be used to organize empirical observations and statements. The connection between the empiric data and the spider graph had to go through a series of supplementary questions. However, we omitted the use of any metrics calculated based on the visual analogue scale as these did not seem appropriate with our use of the tool. Despite this restriction, we obtained a better understanding of the diversity in the empirical data. Moreover, by implementing the concepts of the pragmatic-explanatory continuum, we were better able to relate the data to the requirements for a trial design than without the tool.

Justification of currently applied procedures and treatment protocols

Timing of examinations

The timing of the initial clinical examination in the HHMP is specified by legislation. Early clinical examinations and potential treatments take place during the involution period. This presents the veterinary practitioner with a differential diagnostic challenge, because she must distinguish between pathological and non-pathological discharge with the aim of minimizing false positive and false negative diagnoses. These aspects of the HHMP have caused some controversy regarding the applied procedures, both within Denmark and internationally.

When we reviewed the literature on diagnostic procedures and treatment protocols for genital disease validated in controlled randomized trials, we found that the recommendations of timing differed, depending on the aim of treatment. When the aim was improvement of reproduction performance (e.g., time to pregnancy) in cows with purulent vaginal discharge, the strategies were: (1) treat both puerperal and clinical metritis before 14 days postpartum (Goshen and Shpigel, 2006) or (2) treat puerperal metritis when diagnosed, and treat clinical endometritis after 27 days postpartum (LeBlanc et al., 2002a; LeBlanc et al., 2002b). From an overall perspective, it is recommended that all clinical and subclinical signs (VDS and inflammation) should be resolved within 35 days postpartum (Dubuc et al., 2011). However, when the aim was to avoid milk loss due to genital diseases, the recommendation were based on findings that cows examined early and initially treated for primary puerperal or clinical metritis within 5-14 days postpartum showed increased milk production for ≥ 2 parity (Goshen and Shpigel, 2006). Those findings indicated that early diagnosis, similar to the Danish HHMP recommendation, and an effective treatment protocol could diminish milk loss related to puerperal and clinical metritis, at least for some parity groups. Furthermore, other studies showed extensive self-cure rates before 60 days postpartum and complex dynamics (persistence and re-occurrence of clinical signs) between 60 and 150 days postpartum (Gautam et al., 2010). Those results indicated that early treatment could introduce a higher level of false positive diagnoses (and unnecessary treatments) compared to a more conservative examination and treatment strategy.

In the context of the Danish HHMP, we noticed several important aspects of the applied methods and the scientific findings. First, when clinical metritis could resolve without significant loss of milk production or impaired reproduction performance, the veterinarian could consider postponing the timing specified in the Danish HHMP for the initial and follow-up examinations. Second, in herds with reproduction impairments, the systematic follow-up vaginal examination in the HHMP (and in a potential trial situation) could be postponed beyond 27 days postpartum, but the veterinarian could confirm negative clinical findings at least by 35 days postpartum. Third, in herds with reproduction impairments due to (sub)-clinical endometritis or

cervicitis, smears and a cytological evaluation could be considered an alternative to the vaginal discharge evaluation.

At the same time, it is important to consider both milk production and reproduction together in each herd-specific context before abandoning the current procedures. The majority of interviewed Danish veterinarians had formed a clinical perception and gained experience that influenced their opinions about the effects of the clinical examination and early treatment. This experience should be considered as 'evidence' in the context of 'evidence based veterinary medicine' as applied in practice (Schmidt, 2007) though it is of little scientific value. However, we believe that more research is needed on the dynamics of genital disease and the justification of diagnostic procedures based on vaginal discharge. Before the program can be altered, studies are needed to investigate the interactions between vaginitis, puerperal/clinical metritis before 21 days postpartum, clinical/subclinical endometritis after 21 days postpartum, and the extent of self-cure situations. Small herd-specific changes in timing and procedures (e.g., changing treatment regime) could be based on evidence obtained in the proposed within-herd clinical field trials.

Diagnostic procedures and treatment threshold

Methods suitable for cow-side diagnosis and decision-making in practice (and in the trial context) must be economical, rapid, and they must require a minimum of robust equipment, because decisions regarding treatment are most often made in the barn. All methods for detecting the presence or absence of discharge in the vagina, like the manual vaginal/rectal examination, the 'metri-check' device, and vaginoscopy meet the necessary requirements (Runciman et al., 2009). Comparable methods are needed for diagnosis of clinical entities, until approximately 30 days postpartum (Pleticha et al., 2009; Runciman et al., 2009). However, the validity of vaginal discharge as a predictor of uterine pathology is a controversial subject in on-going research. From a theoretical and preventive perspective, it is important to identify the pathology, aetiology, and origin (e.g., uterine, cervical, or vaginal) of vaginal discharge. This area has been researched and discussed intensively in recent years (Deguillaume et al., 2012; Dubuc et al., 2010a; Dubuc et al., 2010b; LeBlanc, 2008). However, from a practical and pragmatic (but somewhat controversial) viewpoint, the theoretical considerations could be considered less relevant in the HHMP context for several reasons. First, it would often be practically difficult (or impossible) and expensive to distinguish between disease entities and their potential co-existence. Second, if the effects of any level of discharge (regardless of origin) on clinical appearance or production measurements are comparable and/or additive and the recommended treatment is nearly the same (e.g., systemic antibiotics for uterine, cervical, and vaginal infections); thus, identifying the exact pathological or aetiological diagnosis would have little or no implications for the decisions made in practice. Consistent with our pragmatic approach, Runciman and co-

workers proposed the term 'bovine reproductive tract inflammatory disease (BRTID)' for cases of reproductive tract disease diagnosed with methods that rely only on samples of vaginal content (Runciman et al., 2009).

Theories related to diagnostic testing can be linked to the applied diagnostic procedures and specifically to the VDS. The weighing of sensitivity vs. specificity, the predictive value of the diagnostic procedure(s), and the determination of applied treatment thresholds are based on individual criteria for each veterinarian. In practice, most veterinarians may be unaware of the importance of these issues in relation to data quality and comparability between cases/non cases or among treatment groups in the trial context. Veterinarians that use multiple diagnostic criteria to determine a treatment threshold probably implicitly use 'serial interpretation'; in other words, all criteria must be fulfilled for the cow to be treated. This method of test interpretation increases the overall specificity and reduces the overall sensitivity (Dohoo et al., 2003). The serial interpretation approach corresponds to the viewpoint of veterinarians that consider it important to avoid unnecessary treatments. This is because tests with high specificity and low sensitivity result in relatively high detection of true negative diagnoses and low detection of false positive diagnoses. In contrast, other veterinarians appeared to be concerned about the risk of production loss. This segment of veterinarians implicitly expressed that it was important not to miss potentially 'diseased' cows and it was less important that false positive cows were treated. These veterinarians often used a fixed treatment threshold (e.g., $VDS \geq 5$) and standard procedures in all herds. This procedure reflected the requirement of a test with high sensitivity (finding most true positives). In the development of a diagnostic protocol for future clinical field trials, a disease index, based on multiple disease indicators (e.g., VDS + rectal temperature), can be considered when the aim (expressed by the veterinarian and farmer in a future trial) is to influence the number of cows treated (e.g., reduce false positives). Recently, an index for uterine status (0-42 days postpartum), based on vaginal discharge proportions, odour, and rectal temperature, was associated with the delay of involution events (Gorzecka et al., 2011). However, the index should be associated with production performance before implementation in practice and in a trial context.

During the observations and interviews, the first author was frequently asked: "at what VDS value should I initiate treatment?" Changing the treatment threshold based on a cut-off value on the VDS scale would correspond to changing the sensitivity and specificity of the diagnostic procedures. To answer this question adequately, the VDS should be modelled based on different selected outcomes (e.g., milk or reproduction measurements). A ROC analysis could provide knowledge about the sensitivity and specificity of different VDS cut-off values. Non-gold-standard methods, like Latent class models, could also be considered (Krogh et al., 2011). The best answer available at present is based on a randomized trial with diagnostic methods comparable to those of the Danish HHMP. That study showed a parity-dependent decrease in the 305-day

milk yield and a lower conception rate at first insemination for untreated metritic cows compared to their 'non-metritic herd mates' (cows with retained placenta were not included) (Goshen and Shpigel, 2006). The treatment threshold for giving effective treatment was based on multiple factors, including flaccid, non-retractable uterus; cervical diameter >7.5 cm; watery, purulent, and fetid vaginal discharge within 5-14 days postpartum. That threshold was valid for evaluating both milk yield and reproduction effects. Other researchers proposed a more pragmatic practical approach to a treatment protocol for puerperal metritis. They stated that cows should be given systemic treatment when they had any two of the following diagnostic indicators: retained placenta, fever, dullness, or fetid uterine discharge (LeBlanc, 2008).

The procedure that relied on the farmer's observation to evaluate a clinical treatment effect could potentially lead to the preferential repeated treatment of high-yielding or 'valuable' cows; hence, that approach may produce misleading statistical interactions in the analysis of treatment effects. In the trial context, the procedure that relies on the farmer's observation is highly pragmatic, and it is comparable to 'self-reporting' in human medicine. The problem with misleading inferences due to poor data quality was further illustrated by the identification of a farmer that deliberately misused medical treatment and registered it as 'other disease' to circumvent the legislation. The extent of irregularities in the HHMP at the time of this study is unknown.

Treatment protocols

The use of intrauterine therapy, either alone or in combination with a parenteral injection, was widespread among the interviewed veterinarians, despite some evidence in the literature that indicated that the intrauterine application provided no additional effect (Drillich et al., 2001; Smith, 1998). Some veterinarians justified this choice by the observation of mixed intrauterine bacterial flora. However, the veterinarian's choice of drug in Denmark is strongly influenced by legislation and official recommendation (e.g., preventive treatment and off-label use are illegal, and official recommendations promote the use of narrow spectrum antibiotics and reduced use of cephalosporin and tetracycline). For this reason, most current literature on treatment effects may lack relevance under Danish conditions.

In Denmark, prostaglandin injections cannot (legally) be performed by the farmer, only by the veterinarian. Consequently, treatment most often consisted of a single dose of prostaglandin given by the veterinarian at any given time in the postpartum period. It was often applied as a secondary treatment one week after the antibiotic treatment. Little evidence justifies this procedure. The application of a single dosage of prostaglandin (between 20-33 μg) was shown to be effective for preventing reduced reproductive performance due to subclinical endometritis (Kasimanickam et al., 2005); however, the clinical and production-related effects of single injections given before 20 days postpartum should be investigated.

General considerations on standardization of procedures

The overall findings of a diverse range of procedures might warrant a more strict policy for examination procedures and treatments. However, the necessity for increased regimentation depends on the purpose of data collection and analysis, as discussed intensively in a previous related study (Lastein et al., 2009). For example, under conditions in Israel, it was shown that diverse diagnostic procedures and treatment protocols did not affect milk production when the overall procedures were followed (examination of all cows at risk) (Bar and Ezra, 2005). Those findings justified some individual diagnostic freedom for the veterinarian, within the overall framework of an extended HHM program with clinical registration. In the herd-field-trial context, the importance of consistency in procedures lies entirely at the herd level. In multi-herd analyses, consistency across herds is also required.

Requirements for trial design

Conceptual design

During the PRECIS analysis of the observed range of procedures, we found the broadest diversity among the interviewed veterinarians within Domain (7), the primary trial outcome. We attribute this diversity to the difference between ‘explanatory veterinarians with a cow focus’ and ‘pragmatic veterinarians with a herd focus’. The explanatory extreme defines the primary goal of treatment as a clinical cure; the pragmatic extreme insists on production improvement as the goal of treatment. However, within the range of perceptions, some veterinarians viewed clinical cure as a prerequisite for production effects. This standpoint indicated that the clinical cure could be included as either an acceptable surrogate endpoint or a stopping rule in a sequential trial design with a pragmatic nature.

The observed diversity in Domain (9), the Veterinarian’s and farmers’ adherence to protocols, was attributed to the difference between using systematic procedures for data collection and treatment application versus an acceptance of anarchy in the diagnostic criteria and randomness in the treatment protocol.

We are aware that the concept of a very pragmatic trial design can appear rather controversial to academics educated in the natural science community due to the somewhat multi-dimensional relation between context-specific and general evidence and the explanatory-pragmatic continuum. The results in this study illustrated the need for cooperation between the farmer and the veterinarian in the early phases of planning a clinical field trial to place a meaningful research question within this frame. We recommend a bottom-up approach where veterinarian and farmer formulate their own design to ensure selecting a conceptual design that is relevant for all involved parties, perhaps supported by professional trial managers. Whether to proceed with a more explanatory type or pragmatic type trial in a given herd-specific situation will thus depend entirely on the context. In particular, the overall purpose of

implementing a trial in a HHMP to support decision making related to a herd problem will influence the choice of trial design. The PRECIS tool could be used to ensure consistency in each individual trial (the original purpose of the tool). For instance, some veterinarian with a clinical cow-focused could proceed with an explanatory type single-herd trial with a detailed diagnostic protocol (multiple diagnostic criteria), with strong adherence enforcement to measure treatment effectiveness on clinical cure (VDS and rectal temperature) within 2 weeks post treatment. This approach would produce herd-specific evidence of the biological association between treatment and clinical cure. In contrast, veterinary practice-focused veterinarians could proceed with a pragmatic type multi-herd trial estimating treatment effectiveness on long term milk production with a more relaxed protocol. This approach would result in more general (practice level) evidence of the effect of the decision of treating metritis in the context in these practice settings. In general, we found that veterinarians in the HHMP was very focused on key performance indicators as milk yield (as anticipated) and that there was a tendency to pragmatic elements in their daily routines (also illustrated in the spider graph). Based on these finding we would primarily assume that veterinarians would follow a more pragmatic type trial design. In table 1, the fictitious extremes could be used to exemplify the range of possible trial design relevant in the metritis case.

To maintain motivation throughout the trial period, it is important to have a common aim and adequate knowledge and acceptance of all procedures (Farrell et al., 2010). To ensure the internal validity of any proposed herd trial approach, both veterinarians and farmers must be motivated to improve the quality of data. Although this statement can be regarded as highly explanatory, we find that the better data quality, the more reliable estimates of effects no matter the conceptual trial design. That is, the participants must at least agree to try to adhere to the protocol to avoid missing cows at examination etc. Consequently, further research may be warranted to identify factors that motivate these individuals to evaluate treatment effects in a coherent approach and subsequently perform clinical field trials in the HHMP.

Practical design

Based on the observations of field practice and the interviews, we concluded that, to achieve successful development and implementation of the herd-trial-design, the procedure must be simple to follow. In particular, the practical aspects of the trial must be easy to incorporate in the everyday routines of the HHMP. Otherwise, a clinical field trial in the HHMP context is likely to fail. Unfortunately, this insight obligates the rejection of potentially interesting designs. For instance, an 'adaptive allocation' trial design is attractive, because it resembles the every-day procedures of many veterinarians. This design also satisfies the ethical concerns of caring for the individual cow in a group experiment, because the number of cows allocated to an inferior intervention is minimized. Moreover, the adaptive allocation design permits valid

statistical inferences of differences between intervention groups, similar to other designs (Rosenberger, 1999). However, without expert support, we consider the allocation procedures in the adaptive trial design too complex and impractical to fit into the context of HHMP field trials organized by a veterinary practice.

Ideally, trials aiming at evaluating treatment effectiveness and/or validation of treatment threshold should include a non-treated or placebo-treated control group. However, in the context of clinical field trials in HHMP, it would not be possible to include a non-treated control group in most cases, because financial compensation is not an option. Also, a non-treated control group is considered an explanatory choice for a comparison intervention. A more pragmatic approach would be to include a control group that was treated according to the currently applied protocol in a so-called “actively controlled” trial. However, inferences from that kind of study should be carefully interpreted; it would be important to take into account the analytical purpose of the study (superiority versus non-inferiority/equivalence) and the fact that new treatments should be compared to validated treatments (e.g., it is not recommended to compare two unvalidated treatments) (EMA, 2001). Some validated treatments (e.g., off-label use of intravenous tetracycline) are illegal, and the use of cephalosporin is not recommended by the Danish authorities; therefore, it is currently difficult to select a validated treatment protocol under Danish laws.

Alternatively, we have considered introducing the clinical field trial in the context of an Evolutionary Operation (EvOp) (Box et al., 1978; Schwabe et al., 1977). The EvOp principle involves testing small changes in a continuous cyclic pattern. Implementation of field trials in an EvOp-like fashion could eventually lead toward a more efficient use of medical interventions and optimization of treatment thresholds in the Danish HHMP, despite the small changes observed in each trial cycle.

Future perspectives that influence implementation

Before initiating implementation of the proposed clinical field trials in the HHMP, the on-going changes in legislation in the area and the consequences thereof should be considered in depth. An example of data manipulation as a consequence of present legislation was found in an instance where a farmer deliberately introduced error in his records to circumvent the legislation. In a trial situation, the records of additional treatments in that herd would be useless. Thus, changes in legislation will inevitably influence the way that data can be collected. Other examples of how changes in the Danish legislation might impair data quality are discussed elsewhere (Krogh, 2012); this issue complicates the design of clinical field trials.

Finally, we will address some additional, relevant design elements that should be considered in planning field trials and analyses. The selection of a pragmatic or explanatory conceptual design will influence the design elements listed below, but the detailed implications are beyond the scope of this article.

- *Sample size*: To obtain statistically valid results from a quantitative analysis, sample size (e.g., the number of cows in each intervention group) must be adequate. However, under our

conditions, the sample size will be primarily influenced by several non-statistical factors. One factor is the herd characteristics, like herd size and disease incidence, which is influenced by the choice of treatment threshold. A second factor is the choice of a 'relevant clinical difference' in clinical or production outcome between the intervention groups. Depending on the choice of comparison group (no treatment, placebo, or active control), the sample size will vary; a smaller sample size is required to evaluate the 'true' effect between treated and non-treated or placebo groups, and a larger sample size is required to evaluate the difference between two treatments. A third factor is the length of the trial study period; a protracted trial can infinitely increase the sample size. However, relevance and motivation requires an acceptable time frame. We suggest a maximum study period of 6 months to 1 year.

- *Randomization procedure:* The practical context of the HHMP requires a randomization method that does not include additional paperwork to be completed in the barn. An existing cow identification would be preferable (e.g., even or uneven ear-tag numbers) to avoid additional cow labelling in the barn. Important confounders should be identified, and their equivalence at baseline should be ensured; alternatively, an analytical control should be implemented (Dohoo et al., 2003). However, Dohoo et al. also stresses that such procedures are systematic assignments to groups, but also that the procedures in practice will work equally compared to 'classical randomization (e.g., random number generation)'. As a consequence of systematic assignment or 'pseudo-random' procedures, we have chosen to refer to our proposed trial designs as a 'clinical field trials', not randomized trial. However, we will emphasize trial to be implemented in HHMP should be either 'pseudo-randomized' or randomized to ensure the validity of the results. *Blinding:* Although blinding is considered an explanatory element (Thorpe et al., 2009), it is crucial for reducing bias (EMEA, 2012). However, we assume that blinding veterinarians and farmers in the present context would be unfeasible and unnecessary for clinical field trials in a HHMP context. We speculate that it would be unlikely for the farmer or veterinarian to want to cheat, when they have agreed to participate with informed consent. It would be possible to perform the analysis blinded, if the analysis could be automated or coded to fit this requirement.
- *Analytical design:* Alternatives to superiority testing, like non-inferiority or equivalence testing, should be considered (Piaggio et al., 2006). The choice of data management and analytical method should be consistent with the overall purpose of the study. For pragmatic trials, an 'intention to treat' analysis is considered the most consistent method. In explanatory trials, both 'per protocol' and intention to treat' analyses should be performed (Thorpe et al., 2009).

Conclusion

We found that there was considerable variability in veterinarians' procedures for the diagnosis and treatment of genital diseases within 5-21 days postpartum in an extended Danish HHMP. The procedures were difficult to evaluate and justify, based on the available scientific evidence, due to diverging definitions and objectives of existing studies. According to scientific recommendations, the timing of diagnostic procedures for follow-up examinations designed to improve reproductive performance in the HHMP should be postponed until after 30 days postpartum. The present study identified systematic elements in the applied procedures that could be implemented in field trial protocols. We concluded that a protocol for potential clinical field trials should have a simple design, defined by the veterinarians and farmers in cooperation, to facilitate and optimize adherence to protocol, valid data recording, and subsequent valid results.

Based on the PRECIS tool, we propose three different conceptual trial designs, as follows: (1) An explanatory, within-herd clinical field trial with a clinical focus, (2) a pragmatic, within-herd clinical field trial with a production focus, and (3) a pragmatic, multi-herd clinical field trial with a production focus. We propose that the trial design should include non-blinded procedures, systematic assignment procedures such as for instance ear-tag allocations, and the use of parallel groups, a modified cross-over treatment, or a factorial group, depending on the combination of treatments to be tested. These practical designs can potentially be superimposed on a group-sequential analysis with clinical stopping boundaries, based on the clinical cure as a surrogate endpoint. Negative or active control groups can be chosen, depending on the specific herd situation. 'Intention to treat' analyses are proposed to provide a pragmatic, conservative estimate of the effect of changing standard regimes. 'Per protocol' analyses could be relevant in the explanatory cow-focus trial. We also concluded that the successful implementation of the proposed randomized trials in Danish veterinary practice and the HHMP would depend on a high degree motivation, which was lacking among some of the veterinarians in this study. Further research in the area of motivation is recommended before trials can be successfully integrated into HHMPs.

References

- Bar, D., Ezra, E., 2005. Effects of common calving diseases on milk production in high yielding dairy cows. Israel veterinary medical association 60(4).
- Bjerkset, O., Larsen, S., Reiertsen, O., 1997. Evaluation of enoxaparin given before and after operation to prevent venous tromboembolism during digestive surgery: Play the winner designed study. *W J Sur* 21, 584-589.
- Box, G., Hunter, S., Hunter, W., 1978. *Statistics for Experimenters*. Wiley, USA. ISBN: 0471093157
- Deguillaume, L., Geffré, A., Desquilbet, L., Dizien, A., Thoumire, S., Vornière, C., Constant, F., Fournier, R., Chastant-Maillard, S., 2012. Effect of endocervical inflammation on days to conception in dairy cows. *J. Dairy Sci.* 95(4), 1776-1783.
- Dohoo, I.R., Martin, W., Stryhn, H., 2003. *Veterinary epidemiologic research*. AVC Inc., Charlottetown, Prince Edward Island, Canada. ISBN: 0-919013-41-4
- Drillich, M., Beetz, O., Pftuzner, A., Sabin, M., Sabin, H.J., Kutzer, P., Nattermann, H., Heuwieser, W., 2001. Evaluation of a systemic antibiotic treatment of toxic puerperal metritis in dairy cows. *J. Dairy Sci.* 84(8), 2010-2017.
- Dubuc, J., Duffield, T.F., Leslie, K.E., Walton, J.S., LeBlanc, S.J., 2010a. Definitions and diagnosis of postpartum endometritis in dairy cows. *J. Dairy Sci.* 93(11), 5225-5233.
- Dubuc, J., Duffield, T.F., Leslie, K.E., Walton, J.S., LeBlanc, S.J., 2010b. Risk factors for postpartum uterine diseases in dairy cows. *J. Dairy Sci.* 93(12), 5764-5771.
- Dubuc, J., Duffield, T.F., Leslie, K.E., Walton, J.S., LeBlanc, S.J., 2011. Randomized clinical trial of antibiotic and prostaglandin treatments for uterine health and reproductive performance in dairy cows. *J. Dairy Sci.* 94(3), 1325-1338.
- Elkjær, K., 2012. *Reproduction in the post partum dairy cow - influence of vaginal discharge and other possible riskfactors*. Ph.D.thesis. Science and Technology. Aarhus University, Denmark.
- EMA, 2001. Choice of control group in clinical trials. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002925.pdf . Assessed 18-6-2012.
- EMA, 2012. Guideline on statistical principles for clinical trials for veterinary medical products (pharmaceuticals). http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/01/WC500120834.pdf . Assessed 18-6-2012.
- Farrell, B., Kenyon, S., Shakur, H., 2010. Managing clinical trials. *Trials* 11(78).
- Gautam, G., Nakao, T., Koike, K., Long, S.T., Yusuf, M., Ranasinghe, R.M.S.B., Hayashi, A., 2010. Spontaneous recovery or persistence of postpartum endometritis and risk factors for its persistence in Holstein cows. *Theriogenology* 73(2), 168-179.

Gorzecka, J., Friggens, N.C., Ridder, C., Callesen, H., 2011. A universal index of uterine discharge symptoms from calving to 6 weeks postpartum. *Reproduction in Domestic Animals* 46(1), 100-107.

Goshen, T., Shpigel, N.Y., 2006. Evaluation of intrauterine antibiotic treatment of clinical metritis and retained fetal membranes in dairy cows. *Theriogenol* 66(9), 2210-2218.

Kasimanickam, R., Duffield, T.F., Foster, R.A., Gartley, C.J., Leslie, K.E., Walton, J.S., Johnson, W.H., 2005. The effect of a single administration of cephapirin or cloprostenol on the reproductive performance of dairy cows with subclinical endometritis. *Theriogenology* 63(3), 818-830.

Kristensen, E.L., 2008. Valuation of dairy herd health management. PhD thesis, Faculty of Life Sciences, University of Copenhagen, Denmark.

Krogh, M.A., 2012. Management of data for herd health performance measurements in the dairy herd. Ph.D.thesis. Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.

Krogh, M.A., Toft, N., 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.

Kvale, S., 1994. Interview - en introduktion til det kvalitative forskningsinterview. Hans Reitzels Forlag, [in Danish]. Copenhagen, Denmark. ISBN: 87-412-2816-2

Lastein, D.B., 2012. Herd-specific randomized trial - an approach for effect evaluation in a dairy herd health management program. Ph.D.thesis.Faculty of health and Medical sciences, University of Copenhagen, Denmark.

Lastein, D.B., Enevoldsen, C., 2009. Clinical stopping rules in sequential field trials. *J. Dairy Sci.* 92, E-Suppl 1.

Lastein, D., Vaarst, M., 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 Veterinaria Scandinavica* 51(1), 36.

LeBlanc, S.J., Duffield, T.F., Leslie, K.E., Bateman, K.G., Keefe, G.P., Walton, J.S., Johnson, W.H., 2002a. Defining and diagnosing postpartum clinical endometritis and its impact on reproductive performance in dairy cows. *J. Dairy Sci.* 85(9), 2223-2236.

LeBlanc, S.J., Duffield, T.F., Leslie, K.E., Bateman, K.G., Keefe, G.P., Walton, J.S., Johnson, W.H., 2002b. The effect of treatment of clinical endometritis on reproductive performance in dairy cows. *J. Dairy Sci.* 85(9), 2237-2249.

LeBlanc, S.J., 2008. Postpartum uterine disease and dairy herd reproductive performance: A review. *Vet J* 176(1), 102-114.

Ministry of Food, Agriculture and Fisheries, 2006. Act of New Health Management in cattle herds. Order No 1045 of 20 November 2006 [in Danish].

Onwuegbuzie, A.J., Leech, N., 2007. A call for qualitative power analysis. *Qual & Quan* 41, 105-121.

Piaggio, G., Elbourne, D.R., Altman, D.G., Pocock, S.J., Evans, S.J.W., for the CONSORT Group, 2006. Reporting of noninferiority and equivalence randomized trials: An extension of the CONSORT Statement. *JAMA* 295(10), 1152-1160.

Pleticha, S., Drillich, M., Heuwieser, W., 2009. Evaluation of the Metriceck device and the gloved hand for the diagnosis of clinical endometritis in dairy cows. *J. Dairy Sci.* 92(11), 5429-5435.

Rosenberger, W.F., 1999. Randomized play-the-winner clinical trials: review and recommendations. *Cont Clin Trial* 20(4), 328-342.

Runciman, D.J., Anderson, G.A., Malmo, J., 2009. Comparison of two methods to detecting purulent vaginal discharge in post partum cows and effect of uterine cephalosporin on reproductive performance. *Australian veterinary Journal* 87(9), 369-378.

Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M., Haynes, R.B., Richardson, W.S., 1996. Evidence based medicine: what it is and what it isn't. *BMJ* 312(7023), 71-72.

Schmidt, P.L., 2007. Evidence-based veterinary medicine: Evolution, revolution, or repackaging of veterinary practice? *Veterinary Clinics of North America: Small Animal Practice* 37(3), 409-417.

Schwabe, C., Riemann, H., Franti, C., 1977. Herd health programs. *Epidemiology in veterinary practice*. Lea & Febiger, Philadelphia, USA., 246-248. ISBN: 0-8121-0573-7

Smith, B.I., 1998. Comparison of various antibiotic treatments for cows diagnosed with toxic puerperal metritis. *J. Dairy Sci.* 81(6), 1555-1562.

Thorpe, K.E., Zwarenstein, M., Oxman, A.D., Treweek, S., Furberg, C.D., Altman, D.G., Tunis, S., Bergel, E., Harvey, I., Magid, D.J., Chalkidou, K., 2009. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *Journal of Clinical Epidemiology* 62(5), 464-475.

Vaarst, M., Paarup-Laursen, B., Houe, H., Fossing, C., Andersen, H.J., 2002. Farmers' choice of medical treatment of mastitis in Danish dairy herds based on qualitative research interviews. *J. Dairy Sci.* 85(4), 992-1001.

Whitehead, J., 1992. *The design and analysis of sequential clinical trials*. Wiley, West Sussex, England. ISBN: 0 471 97550 8

3.5 Veterinary decision making in relation to metritis - a qualitative approach to understand the background for variation and bias in veterinary medical records

Manuscript IV

Published: Acta Scandinavica Veterinaria 2009; 51:36

D. B. Lastein^a, Mette Vaarst^b & C. Enevoldsen^a

Department of Large Animal Sciences

Faculty of Health and Medical Sciences

University of Copenhagen

Grønnegårdsvej 2, DK-1870 Frederiksberg C

Denmark

^bDepartment of Animal Health

Welfare and Nutrition

Faculty of Agricultural Sciences

Research Centre Foulum

University of Aarhus

P.O. 50, DK-8830 Tjele

Denmark

“Dogmatism: this is the best way to do it.

Policy: this is the way we do it around here.

Experience: this way worked the past few times.

Whim: this way might work.

Nihilism: it does not really matter what we do.

Rule of least worst: do what you are likely to regret the least.

Defer to experts: how would you do it?

Defer to patient: how would you like to proceed?”

Methods of decision making in veterinary practice as described by Cockcroft (2007) Veterinary Clinics of North America: Small Animal Practice 37(3), 499-520. 2007.

Research

Open Access

Veterinary decision making in relation to metritis - a qualitative approach to understand the background for variation and bias in veterinary medical records

Dorte B Lastein*¹, Mette Vaarst² and Carsten Enevoldsen¹

Address: ¹Department of Large Animal Sciences, Faculty of Life Sciences, University of Copenhagen, Grønnegårdsvej 2, DK-1870 Frederiksberg C, Denmark and ²Department of Animal Health, Welfare and Nutrition, Faculty of Agricultural Sciences, Research Centre Foulum, University of Aarhus, P.O. 50, DK-8830 Tjele, Denmark

Email: Dorte B Lastein* - bay@life.ku.dk; Mette Vaarst - mette.vaarst@agrsci.dk; Carsten Enevoldsen - ce@life.ku.dk

* Corresponding author

Published: 30 August 2009

Received: 18 May 2009

Acta Veterinaria Scandinavica 2009, **51**:36 doi:10.1186/1751-0147-51-36

Accepted: 30 August 2009

This article is available from: <http://www.actavetscand.com/content/51/1/36>

© 2009 Lastein et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Results of analyses based on veterinary records of animal disease may be prone to variation and bias, because data collection for these registers relies on different observers in different settings as well as different treatment criteria. Understanding the human influence on data collection and the decisions related to this process may help veterinary and agricultural scientists motivate observers (veterinarians and farmers) to work more systematically, which may improve data quality. This study investigates qualitative relations between two types of records: 1) 'diagnostic data' as recordings of metritis scores and 2) 'intervention data' as recordings of medical treatment for metritis and the potential influence on quality of the data.

Methods: The study is based on observations in veterinary dairy practice combined with semi-structured research interviews of veterinarians working within a herd health concept where metritis diagnosis was described in detail. The observations and interviews were analysed by qualitative research methods to describe differences in the veterinarians' perceptions of metritis diagnosis (scores) and their own decisions related to diagnosis, treatment, and recording.

Results: The analysis demonstrates how data quality can be affected during the diagnostic procedures, as interaction occurs between diagnostics and decisions about medical treatments. Important findings were when scores lacked consistency within and between observers (variation) and when scores were adjusted to the treatment decision already made by the veterinarian (bias). The study further demonstrates that veterinarians made their decisions at 3 different levels of focus (cow, farm, population). Data quality was influenced by the veterinarians' perceptions of collection procedures, decision making and their different motivations to collect data systematically.

Conclusion: Both variation and bias were introduced into the data because of veterinarians' different perceptions of and motivations for decision making. Acknowledgement of these findings by researchers, educational institutions and veterinarians in practice may stimulate an effort to improve the quality of field data, as well as raise awareness about the importance of including knowledge about human perceptions when interpreting studies based on field data. Both recognitions may increase the usefulness of both within-herd and between-herd epidemiological analyses.

Background

Files with information on animal disease have a variety of applications at both the herd and national level, including monitoring the incidence of animal diseases or medical treatments, analyses of causal relationships, bench marking, estimation of treatment criteria, effectiveness of treatment on production, etc. Such information necessarily must be gathered from multiple observers in a wide range of contexts (e.g., the Danish national cattle database). Both disease detection and criteria for treatment are influenced by human perception, as exemplified by a study of farmers and mastitis [1]. This influence introduces the possibility of both variation and bias (e.g., problems related to intra- and inter-observer agreement). Consequently, consideration of data quality in existing data files becomes essential before any quantitative analysis can be conducted and interpreted. Intra- and inter-observer agreement about the manifestations and criteria for treatment must be estimated (quality control), because different people often judge the same conditions differently, as discussed by Baadsgaard and Jørgensen [2].

Disease manifestations or 'diagnostic data'--e.g., which clinical signs of metritis can be seen or scored--should be clearly distinguished from treatment records or 'intervention data'. In the Danish Central Cattle Data Base, it is now possible to record information about disease--for example, as various types of scores--and medical treatments separately. This option is primarily used in case of metritis in dairy cows in herds participating in a recently implemented herd health programme [3]. The metritis diagnosis is recorded as an ordinal score with values from 0 to 9 (higher score corresponds to a more 'severe' disease). The scores are gathered by veterinarians between 5 and 21 days in milk from all cows calving in the herds. Medical treatments of metritis are also recorded by the practicing veterinarians, because farmers' use of antibiotics is restricted by Danish legislation.

In summary, the individual veterinarian records two distinct variables: 1) Diagnosis, that is, a score based on observed clinical signs of metritis, and 2) Treatment decision, that is, determining treatment or non-treatment based on criteria for treatment classification. The consequence is that disease incidence can be described separately from disease treatment incidence.

In this article, data collection related to metritis in dairy cattle is investigated empirically and is discussed as an example of problems that must be addressed prior to and during quantitative analyses of such data. The aim of the study is to explore qualitative aspects and potential mutual influences of collecting metritis score data and metritis treatment data, and how the relationship between these two types of data is influenced by human percep-

tions and decisions. The study also considers potential consequences for the quality and subsequent analysis of field data on herd and national levels. The research tool is qualitative analysis of observations in veterinary practice and statements from semi-structured interviews.

Methods

The context

Legislation for a new type of voluntary dairy herd health programme was introduced in Denmark in 2006 [3]. The programme aims at improving the detection and registration of the most important health disorders to allow accurate monitoring of the development of disease incidence over time, hence using these data for disease control measures. The veterinarian and the farmer join the programme by signing a 'herd agreement' specifying a set of rules for mandatory systematic data collection. This agreement gives the farmer a more liberal access to antibiotics. The intention behind this legislation probably was to motivate the farmer to enhance disease prevention through dialogue with and the advice given by the veterinarian. By the end of 2008, approximately 100,000 cows, or approximately 20% of the total Danish dairy cattle population, were enrolled in the program. In these herds, all treatments and scores related to metritis must be recorded systematically, according to a common manual (consult table 1 to see the scorings of metritis) and entered into the Danish Central Cattle Data Base.

The programme is based on systematic weekly/fortnightly clinical screening of all cows in a herd at specific expected high disease risk periods, i.e., at drying off and at calving (5-21 days *post partum*). The mandatory screenings focus on general condition, metritis/vaginitis, mastitis and body condition. Optional screenings focus on ketosis and limb disorders [3]. No official treatment threshold was linked to the metritis scale, but leading Danish veterinarians in the field recommend using a grade of 5 on the scale as a cut-off value for initiating medical treatment, and statements from veterinarians at meetings indicate that this criterion seems to have been generally accepted as a rule of thumb.

Selection of participants

A list of veterinarians with two or more 'herd agreements' within 3 geographical regions was obtained from a central registry of veterinarians. Veterinarians were phoned, starting at the top of the list. Twelve veterinarians, with between 2 and 15 herd agreements per veterinarian; (median: 4 herds) and with from 3 to 30 years of experience in cattle practice agreed to participate after a short introduction. Only one veterinarian from each practice was included. Anonymity was guaranteed to promote openness and confidentiality.

Table 1: Table of metritis score definitions and examples of present usage in practice.

Scores	Clinical signs - vaginal examination	Cases	
		Practical scoring	Decision making on treatment
0	None or very small amount of clean mucous discharge - no odour	L elaborates on the use of score 0: "Well, some should maybe have been 1 or 2. The score 1 I have never used." L scores all cows with a normal puerperal discharge 0.	
1	A very small amount of bloody mucous discharge - no odour		
2	Small amount of bloody mucous/grey discharge - no odour		
3	Large amounts of bloody seromucous/grey-yellow discharge - scabs on tail - no odour	J: "I use 2 - which means I will not treat, but I would like to see the cow again for control [...] I could use 3-4. But I just use 2, and the farmer knows what it means". J uses 0 for cows that are immediately characterized as non metritic.	
4	Large amounts of grey/yellow seromucous discharge - no abnormal odour	K: "My metritis score 4. It is when there is plenty of discharge, that smells and there is no temperature". J: "I can not differentiate as sharp as it is suggested by the system, so I only use 5-7-9".	A uses 4 and rectal temperature as a minimum threshold for metritis treatment.
5	Little to medium amounts of purulent discharge - difference in consistency and colour - smell abnormal		L uses the combination of score 4 and a flaccid uterus by rectal examination to initiate treatment with prostaglandin.
6	Medium amounts of discharge - difference in texture and colour - smell abnormal		K, I, E, J & B are explicitly using 5 as a minimum threshold for treatment.
7	Medium to large amounts of discharge - beginning to look red-brownish - stinks	I: "I have never given a cow score 9 if she was not very ill. We saw a cow I gave 8 [...] If she had had sunken eyes I had probably given her 9 with the same vaginal findings"	D, C, L, & H using a variable threshold for treatment and makes individual decision on individual cows based on multiple clinical criteria (incl. metritis score).
8	Large amounts of greyish discharge - stinks	K's scoring is influenced by rectal temperature: the higher temperature, the higher metritis score.	H attempts to exclude score 8-9 from the scale: "If they have a cow there is as sick as 8-9 they should call in advance."
9	Large amounts of brown-yellow/brown discharge- typically a retained placenta - "smells like h...!"		

The table explains the metritis scores with definitions. Cases from the interviews are given to demonstrate how the scores are used in a practice context, and how they are used during decision making for determining treatment threshold for metritis. Capital letters refer to specific veterinarians.

Participant observation

Observations of veterinary work on farms and interviews were made by the first author [DBL] from January to March 2008. The veterinarians were observed during 1-4 herd visits when the veterinarian did practical scoring and medical treatments. Observations and discussion notes from the herd visits were used later to initiate and guide the interviews of the veterinarians.

Qualitative semi-structured research interviews

All veterinarians were interviewed about their decisions related to metritis using a semi-structured research methodology [4]. The duration was 1/2 hour to 1 1/4 hour per interview. Based on the observations, cases, herd documents and interview themes (table 2), the veterinarians were encouraged to tell about their own personal experiences, perceptions and practical observations regarding

diagnosis (including scoring) and treatment of metritis. DBL directed the conversation through the themes and followed-up on the statements given by the interviewed veterinarian. Most interviews were initiated by either a general opening: 'Could you comment on your thoughts on metritis treatments in the scheme' or more specific: 'This morning I [DBL] observed the following situations in a herd (e.g. scoring a cow and initiating a metritis treatment), would you please elaborate on that specific situation?'

Data Analysis

The qualitative analysis is based on a phenomenographic approach; that is a qualitative method to use empiric data (e.g., interview) to describe the variation in and logical relations between human perceptions of a phenomenon [5,6]. All interviews were recorded with a digital voice recorder and transcribed in full length. Different forms of interaction between practical metritis scoring and treatment decisions were identified. Statements or parts of the interview with a coherent meaning were condensed into short, descriptive headings in a process called 'meaning condensation' [4] Headings were categorized, as we identified differences in the way veterinarians experience the phenomenon of generating score data and decision making in relation to treatment of metritis and their motivation to produce data. This information was condensed into a 'model of understanding' that demonstrates the relationship between perceptions and data quality. The veterinarians' perceptions of the reasoning behind their own decisions were explored. Citations are typically used to demonstrate typical views and meanings.

Results

The use of metritis scores for decision making

All veterinarians initially stated that they used the metritis score as a means to identify a need for treatment. In Table 1, cases of the practical use of metritis scores and decision making on treatment are described. These cases exemplify that the practical usage involves implicit adjustments of treatment criteria to a given situation, i.e., explicit criteria of treatment are not necessarily used by the individual veterinarian. Three types of interactions between scoring and decisions of treatment were identified (Figure 1).

Table 2: Interview themes

Clinical registration
Diagnostic criteria
Treatment strategies
Treatment effect in relation to production parameters
Control of clinical effect
Herd status
Farmer's influence
Influence of strategy in veterinary practice
Ideology
Legislation

As illustrated in Figure 1, one category of veterinarians based their treatment decisions entirely on the metritis score (case 1). Another category of veterinarians included other observations in the treatment decision (case 2). One example also demonstrates how the metritis score was manipulated in order to fit the decision already taken by the veterinarian concerned, but was based on other implicit (not recorded) observations (case 3).

Case 1. In the interview we touch upon organic farmers' explicit wish to minimise the use of medicine, either because of ideology, association between treatments and longer withdrawal period of milk in organic herds, or for other reasons. As an aid to understanding the quote, note that the veterinarian equates 'smell' and metritis score 5 or higher, and that legislation requires that follow-up treatments are done by veterinarians in organic herds.

DBL: "I was wondering if you are running this programme in an organic herd - and the farmer argues for minimal medicine usage - for both economic and ideological reasons. Would you change your treatment threshold?"

VETERINARIAN: "Not voluntarily! I will always treat the ones that smell. Perhaps I could reduce the length of treatment, if the farmer is cranky about it; also because we have to do the follow-up treatment ourselves. Otherwise I always treat a minimum of two days after first treatment."

Case 2. The case is based on an observation in a herd, where DBL had observed the veterinarian examining a cow and recorded a metritis score of 7. The veterinarian decided not to treat the cow. He was asked to elaborate on the case:

VETERINARIAN: "It's a question about looking at the cow. It did not have fever, and it looked 'nice'. No reaction on ketosis sticks. So a score 7 - I believe that the cow can manage the disease without treatment, because she has a good general condition. Treatment might be an issue later - perhaps only because of sequels for reproduction. But my immediate appraisal is that the cow requires no treatment."

Case 3. The treatment criteria were discussed with the veterinarian in case 3. The veterinarian that had selected a treatment criterion at score value 5 had told DBL during the morning's herd visits that 'a cow scored 5 could smell more in one herd than in another'. He is asked to elaborate on the statement during the interview.

VETERINARIAN: "When you stand with your hand in the cow without knowing whether you should treat or not, then I look at the cow; body condition score, milk yield, rectal temperature - and which herd she is in. The herd management means a lot. In some herds she may be left in a corner, and maybe ... what

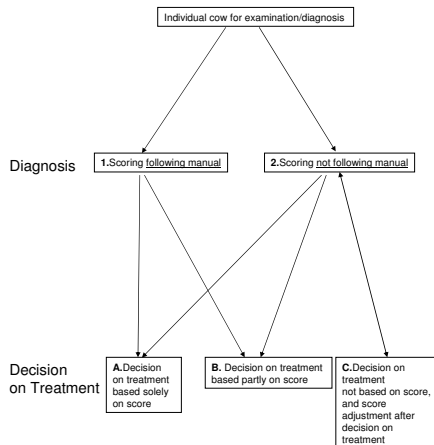


Figure 1
The interactions between diagnostics (incl. metritis score) and decisions on treatment of metritis. The diagram shows that for individual cows diagnosed with metritis, several different pathways of decision related to the metritis score are taken by the interviewed veterinarians.

if her metritic condition worsens? In these herds I treat the cow. In other herds she will never be overlooked. In other herds it is absolutely certain that they'll call me in two days if the metritis condition develops."

DBL: "Do you then score 4 in 'herds where you do not treat'?"

VETERINARIAN: "Yes - because a 5 is treated. The score 5 will vary between herds, but only a little bit."

Model of understanding with regard to decision levels

Based on analysis of the veterinarians' perceptions of how they wished to use the metritis score in their practice and on dialogue with the farmer and surroundings in general, a model of understanding was developed (Figure 2). Three levels of decision were revealed: cow level (individual cows), farm level (multiple cows in a specific farm) and population level (multiple cows in multiple farms). None of the veterinarians took decisions exclusively on one level or were motivated solely through one category of motivation, but they might have been more or less focussed on each of the three levels/categories of motivation.

At the level of the individual cow, the veterinarians seemed to base their treatment decisions on the cow's characteristics. They focussed generally on the practical

use of the score to support treatment of each individual cow, indicating that decisions can differ both within and between herds.

At the farm level, the veterinarians seemed to integrate farm-related information into the decision as to how to treat an individual cow for metritis. When taking decisions on this level, a veterinarian often used predefined herd-specific standard treatments, sometimes with considerable variation between herds (e.g., milk withdrawal period due to individual farmers' wishes). To various degrees, the veterinarians included practical conditions and perceptions such as farmers' inability to manage follow-up treatments or restrain cow properly for intravenous injection. This can give a pattern of treatments which is strongly influenced by the veterinarian's perception of the specific farm and by his or her evaluation of the local context. That is, treatment data as an indicator of a certain disease manifestation may only be valid within the herd.

When veterinarians used standard treatment decisions and included population level considerations and general evidence into the criteria (e.g. using the same cut-off value on metritis scale in all herds), they were generally focussed on the importance of generating data for valid epidemiological analyses across herds. They would therefore both score metritis and make decisions on treatments in a more uniform way across herds, attempting to produce data of both high accuracy within-herd and between-herd.

Categories of motivation for generating data

Four different categories of motivation among the veterinarians for collection and usage of the metritis data were derived from the analysis and given the headings: 1) epidemiological, 2a) advisory, 2b) autonomous advisory, 3) law-abiding and 4) clinical. In Figure 2, the order of these categories is based on the authors' suggestion concerning how these motivations may link to the decision levels and, consequently, data quality. Each veterinarian could be influenced by different motivational factors as described above.

1) Epidemiological

Veterinarians motivated by epidemiological considerations would follow the guidelines for the scoring and would treat based on certain criteria which vary little between cows and herds, so as to be able to create meaningful data valid in large scale analyses (across herds and veterinary practices). Such veterinarians would generally want to focus on possibilities for across-herd data analyses and, with time, be able to formulate meaningful disease control strategies based on empirical data at the herd level. Veterinarians in this category are aware of the possibility of actually basing their decisions on epidemiologi-

cal analyses in the future, and they are highly motivated to use, for instance, multi-factorial analysis on the herd level or higher levels in their daily work (Figure 2).

2a) Advisory

Veterinarians could be motivated by the capacity of scores to function as an entrance to advisory services on the farm level. Such veterinarians are motivated to collect valid data at the herd level. They perceive the collection of the data in and of itself as the basis for taking relevant action at the farm. They may skip the process of systematic analysis of data and give advice based on their immediate evaluation of the results compared to previously collected data ('qualitative monitoring'). Consequently, they are typically focused on internal validity within each farm context, which may make them less concerned with the problems of adjusting treatment criteria and types of treatments between herds. However, 2 subgroups of advisors are identified, 2a) that follow score definition - making data both valid within herd and potentially valid between herds--and 2b) that act autonomously as described below.

2b) Autonomous advisors

These are veterinarians who primarily followed their own definitions of different scoring values, such as excluding certain scores (see examples in table 1). They find the definitions incorrect. If the veterinarian strictly follows his/her own scoring guidelines, the data will be internally valid, but clearly cannot be used between herds.

Autonomous veterinarians are, in general, motivated by the combination of analysis- and experience-based decisions; they act autonomously in the sense that they appreciate the results of analysis, but only if it becomes integrated into the local herd context.

4) Law-abiding

Veterinarians stated that metritis scoring is enforced by law. This was the primary motivating factor for running the herd health programme, rather than, for example, creating possibilities to perform epidemiological analyses or base advice on systematically collected data. This motivation could potentially lead to 'justifying,' i.e., adjusting of the score to fit to the treatment decision. This category of veterinarians based the treatment decision on an overall evaluation of the case, irrespective of the existence of scores.

5) Clinical

These veterinarians clearly spoke of the scores as a 'diagnostic tool' related to each individual treatment decision rather than being part of a collaborative data collection. For example, they could add rectal temperature and other parameters into the scoring (see Table 1 for examples),

which might also lead to lack of data validity, though seen from a clinical point of view, highly relevant. Veterinarians who claimed to be motivated by the use of scoring and data collection for their immediate clinical decisions also included their perceptions of treatment prognoses and experiences from relatively few cases. Veterinarians in this category primarily base decisions about treatment (and/or advice in general) on their personal experience (Figure 2), and not on the basis of analysis, as their 'epidemiological counterparts'.

External factors influencing treatment decisions

Based on the interviews, we identified four types of influencing factors related to treatment decisions:

1. *Production/economy*: Some veterinarians emphasised the positive influence of timely treatments on production in terms of increased milk yield and improved fertility. This also includes considerations on withdrawal time of milk.
2. *Animal health/welfare*: Some veterinarians claimed to consider this as a driving factor when treating as early as possible. Some interviewed veterinarians also referred to experiences with reduced risk of left displaced abomasum and early cullings due to metritis as result of following this programme.
3. *Common strategies in groups of veterinarians*: Some veterinary group practices had developed common 'good practice treatment strategies' (e.g., application of corticosteroids in addition to the antibiotic treatment), which influenced all decisions of each individual veterinarian, and yet still left room for context specific evaluations and decisions.
4. *Public health/antibiotic resistance*: Concerns related to spread of antimicrobial resistance could lead to the non-use of broad-spectrum antibiotics and intrauterine treatments.

Discussion

Considerations on validity of qualitative analysis

Qualitative research methodologies are often used to understand aspects of human perception of life in general and have earlier been used in veterinary sciences for similar purposes [1,7,8]. In this particular interview study, the aim was to reveal perceptions and reasoning behind generation of data and to describe the interaction and relations between the recording of metritis scores and veterinarians' decision making connected to metritis treatment and potential links to data quality. This understanding provides insight into potential errors (bias and random error) related to data based on clinical examina-

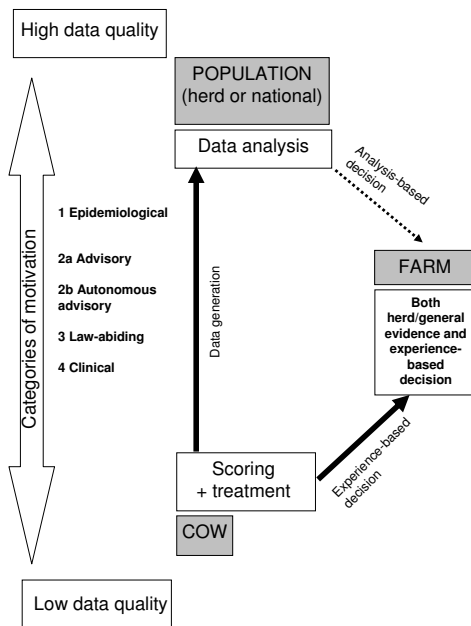


Figure 2
Model of decision levels and categories for motivation. The model shows that veterinarians work on the cow, farm or population level. They generate data between the cow level (scoring and treating metritis) and the population level (data analysis), and potentially use observation or data through either experience- or evidence-based decisions at the farm level. Quality of data (e.g., intra and inter observer agreement) is affected by the 'categories of motivation'. Consequently, the data are more or less suited for subsequent analysis-based decision making on farm and population level. The dotted arrow between population level and farm level indicate that few veterinarians use data analysis in their daily practice and advice.

tions and human decision making, although it may not cover all possible sources of bias in the whole population of veterinarians.

The study makes use of an inductive research methodology called phenomenography [6]. We aim at identifying categories of perception of the phenomena; 'scoring and recording data on metritis' that relate to the quality of the data that are produced. We analyse and build 'a model of understanding' based on DBL's observations and the individual veterinarians' perceptions expressed in their local context. Our basis for the model is thus the empirical data

and not an initiating general theory or hypothesis. From these data we wanted to identify a limited number of ways to understand the phenomenon. It was therefore essential to extract as much information as possible from each context of interest, allowing in this case for a long interaction period between each interviewed cattle veterinarian working with a herd health programme and the researcher. In qualitative research, the data collected from each interviewee should be regarded as the sum of words, tone of voice and body expressions observed during the interaction period, as well as the observer's immediate feelings, experiences, and thoughts on the subjects and the observed [9]. However, we acknowledge the risk of influential interaction between the interviewer and the interviewed during the interviews that could influence the statements of the interviewed e.g., the use of leading questions.

In the phase of analysis it is important to determine when no additional information can be extracted from the interviews and field observations or from additional interviews [9]. 'Information redundancy' or 'data saturation' is a measure of the power and validity of the qualitative studies [9]. Information redundancy or data saturation is reached when we are able to build a model that describes the phenomenon coherently with no internal contradictions. There are no exact criteria to determine when that state is attained. The number of participants (12) was chosen in this study and is in accordance with recommendations for this type of research [9]. Detailed discussion on the methodologies including issues of representativeness and validity, and hence the usefulness of data for quantitative and qualitative research, can be found elsewhere [9,10]. However, it is important to emphasize that the methodology and study design do not enable us to make inferences on the number of veterinarians in each identified categories of motivation. That is, we cannot estimate the quantitative distribution of various ways of reasoning or to give quantitative estimates of bias and random error. This will require another study design. The results of the present study could potentially provide the basis for such a study.

Considerations on data quality and different quantitative analysis

The epidemiological issue of variation and bias are linked tightly with the terms accuracy and precision. Accuracy and precision of disease detection and classification methods at the cow level over time are central to minimizing variation and bias, regardless of the later use of the data for quantitative analyses. Definitions of accuracy and precision here are defined in accordance with Dohoo *et al.* [11]. Accuracy means the average similarity between the observation/classification and the 'true disease state/class'. Because no gold standard for metritis scoring and

few validated criteria for metritis treatment exist at present, the accuracy of observations (scores) and classification (treatment or not) cannot be evaluated against a 'gold standard'. However, under the assumption that the 'true disease state/class' exists, observers' ability to score or classify accurately within and between observer and within and between herd will influence the validity of data, and hence the subsequent analytical use either within the herd ('herd analysis') or between herds ('national analysis'). Accuracy within observer is a prerequisite for valid 'herd analysis' (assuming one observer per herd), and accuracy between observers is a prerequisite for valid 'national analysis'. Precision means the similarity between multiple scorings or classifications of the same condition, either within or between observers. In practice the same cow will very rarely be evaluated twice by the same or another observer at the same time point. In any case, the importance of precision seen from an analytical point of view relates to number of observations required to reveal non-random differences between groups ('significance testing'). Hence sources of variation and bias (poor accuracy and precision) in centrally collected data files--including unstructured human influence--must be revealed, evaluated and discussed in depth prior to a quantitative analysis. This may allow subsequent analytical control of bias.

Sources of bias and variation in veterinary records

Records of metritis scores, ideal for monitoring of disease incidence, should not be influenced by metritis treatment data, because the scores should be given on the basis of strictly defined criteria and should be calibrated within and between observers. Neither should the metritis score be influenced by factors which could potentially influence a treatment decision (e.g., recorded daily milk yield). The treatment data, ideal for epidemiological analysis, should be a result of validated known (explicit) treatment criteria to ensure comparability between cases/non cases, while registrations of additional explicit factors should provide a basis for analytical control of interactions and confounding. However, central data bases are based on field data from multiple observers, which create non-ideal data. In practice, treatment decisions often involve a complex set of observations based on previous experience, local context and external evidence, a situation similar to the concept of evidence based medicine [12].

We have shown in accordance with Kristensen *et al.* [7] that lack of uniformity of scores (e.g., different scores within the same clinical entity and adjustment of scores to suit decisions) leading to reduced intra- and inter-observer agreement are likely to occur in medical records of field data. The sources of misclassification bias (e.g., differences in treatment criteria for metritis scores within and between herds) can represent both the lack of clear

case definitions in field data and the use of different opinions on when to treat, also in cases where different observers might agree on the metritis score they use (case 1 versus case 2 - fixed versus varying criteria for treatment). Further, we have identified interaction and feedback mechanisms between diagnostic observations (scores) and decisions (criteria to treat) which implicate that errors are not independent. Some veterinarians regard the two records as totally correlated, others regard them as entirely independent, and still others regard them as correlated, but adjust the score to suit a decision taken (justification).

This study indicates that some veterinarians working within the herd health programme are primarily focused on case-related problems (at the level of the individual cow), hence lack focus on potential subsequent use and validity of their clinical records in a broader perspective. On basis of this, we suggest that the importance of the epidemiological aspects on data quality of field data should be articulated and emphasised in the education of veterinarians, both at student and post-graduate level.

Potential consequences of bias and variation in veterinary records

Veterinary medical records can be applied in the dairy sector in many ways and for many reasons. In the following we will discuss the consequences of variation and bias in relation to monitoring of animal disease incidence on herd and national level, causal analysis on national level, as well as estimation of validated treatment criteria.

Monitoring of disease incidence (metritis score) over time can be used on the herd level to evaluate, for instance, effects of preventive interventions. Observers within the same herd should be able to obtain unbiased data. Accuracy between herds is irrelevant for evaluating data on herd level e.g. over time. Improved precision of the scores (less variation) will reduce the number of observations needed to obtain an acceptable level of certainty. If metritis is monitored as part of a national programme, accuracy between veterinarians is required. The large variation in the use of the metritis scores and treatment criteria between veterinarians revealed in this study indicate that there is a huge variation between observers. This should clearly be improved before analysing the data on national level.

Causal analysis of cow-level and herd-level risk factors for metritis at the national level based on Danish central data base files was performed by Bruun *et al.* [13] using treatment data as measure of disease. Our study shows that it is very difficult to give a valid biological interpretation of results from across-herd estimates of quantitative associations between clinical conditions (e.g., metritis scores) and disease treatments. The statement from case 2, above

'I believe that the cow due to a good general condition can manage the disease without treatment' - demonstrates that such associations are influenced by multiple factors, both explicit (e.g., acceptable milk yield) and implicit (e.g., perception of good prognosis). This particular veterinarian in case 2 chose to not treat a cow despite a metritis score of 7 (stinking discharge - see table 1 for detailed description). This veterinarian's perception of 'good condition' (true or not) might lead to a lower probability of treatment in average to high yielding cows.

Treatment criteria can be discussed and to some extent calibrated between veterinarians. This would improve comparability between cases and non-cases from different settings, and enable researchers to take into account additional variables in subsequent analyses.

Our study shows that variation and bias in field data (records of metritis scores and metritis treatment) within the herd health scheme are very likely and that the origin is complex, sometimes including feedback. When regularly trained and calibrated, the group of epidemiologically oriented veterinarians might provide data on the metritis scores that are valid for subsequent across-herd analyses of, for instance, quantitative relations between metritis and risk factors or effects of metritis on production. The problem will be to identify the veterinarians belonging to this category in a large file with routinely collected data.

The association between (true) disease state and treatment probably cannot be detected and recorded systematically in all herds, especially not when treatment criteria are based on a combination of factors and rarely made explicit. Consequently, analytical control is probably not possible. If the implicit and explicit treatment criteria are applied on a larger scale, underestimation of effects may occur in some herds, overestimation in others. Unfortunately, there seems to be little evidence in across-herd studies that this problem is even recognized in depth and dealt with. The feedback mechanisms between outcome and risk factor, as well as the interaction between risk factor and herd/veterinarian revealed in this study suggest that observational studies, including meta-analysis, should be interpreted with caution. Including 'random effects' of herd or veterinarian in the analyses will not solve all the problems revealed in this study (e.g. feedback and interaction).

Results of randomised clinical trials can supplement studies involving observational data by creating an understanding of connections between clinical signs and treatment criteria. Only a few controlled clinical trials on early metritis diagnostics and treatments are published. Consequently, little 'external evidence' can be found in

the literature concerning diagnosis and treatment of 'early metritis' [14-17]. This means that very little guidance based on epidemiological analyses or systematically collected veterinary experience can be used as 'validated treatment criteria of metritis'. A possibility to circumvent this gap of herd specific knowledge is to perform within-herd clinical trials as proposed by Kristensen [18].

Has the veterinary paradigm shifted in the minds of veterinarians in practice?

Herd health programmes often aim at close monitoring of disease incidence to allow timely diagnosis, subsequent intervention and evaluation of effects indicating the paradigm shift in veterinary dairy medicine from cows to herds and from treatment to prevention [19]. The results of the present study illustrate how difficult it can be to integrate a systematic approach to clinical examinations and provide useful data - even within the framework of a herd health programme. Some of the veterinarians involved in this study seemed to base both cow-level decisions and, to some extent, farm advice on personal judgements and tacit knowledge, despite their proclaimed intentions to base their daily practice to a higher degree on epidemiological considerations. The results of this study indicate that it is difficult to obtain valid data across herds and between veterinarians when their decision making procedures and motivation to collect data are so different.

Conclusion

Variation and bias in data based on clinical examinations can be linked to veterinarians' individual perception of the purpose of, and their motivations for, data collection. Some veterinarians conduct clinical examinations to support their treatment decision, while others see it as either as a data collection scheme for use at herd level or national level. A model of understanding is developed based on veterinarians' considerations and procedures involving both individual cow characteristics and factors at farm and population level. The study demonstrates that treatment decisions often are likely to be based on both implicit and explicit types of information. Factors identified in the study were the individual cow's general clinical condition and anamnesis, herd and farm related factors, common treatment strategies developed in groups of veterinarians, as well as the veterinarian's perception of the prognosis for treatment(s) with regard to production, economy, animal health and welfare. Acknowledgement of the interaction between human decisions, motivations for disease recording and data quality can potentially lead to improved data quality and/or improved interpretations of the results of quantitative data analyses if the knowledge is communicated to both practicing veterinarians and educational systems. The identified sources of variation and bias should be taken into consideration by

researchers and decision makers (e.g. in organisations and governmental institutions).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DBL has conceptualized and conducted the interviews, transcribed and performed the major parts of the analysis and writing process. MV has contributed substantially with regard to the methodology, analysis and writing. CE has revised the manuscript critically for important intellectual content, in addition to his contribution to the general concepts of the study. All authors has read and approved on the contents of the final manuscript

Acknowledgements

We thank the interviewed veterinarians for willingly sharing time, thoughts and perception during observations and interviews. Language editor Peter Gordy is gratefully acknowledged for multiple revisions.

References

- 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.
- Baadsgaard NP, Jorgensen E: **A Bayesian approach to the accuracy of clinical observations.** *Prev Vet Med* 2003, **59**:189-206.
- Ministry of Food, Agriculture and Fisheries: **Act of New Health Management in cattle herds. Order No 1045 of 20 November 2006 [in Danish].**
- Kvale S: *Interview - an introduction to the qualitative research interview* Copenhagen, Denmark: Hans Reitzels Forlag; 1994. [in Danish]
- Åkerlind GS: **Variation and communality in phenomenographic research methods.** *High Edu Res Devel* 2005, **24**:321-334.
- Barnard A, McCosker H, Gerber R: **Phenomonography: A qualitative research approach for exploring understanding in health care.** *Qual Heal Res* 1999, **9**:212-226.
- Kristensen E, Nielsen DB, Jensen L, Vaarst M, Enevoldsen C: **A mixed methods inquiry into the validity of data.** *Acta Vet Scand* 2008, **50**:30.
- Vaarst M, Bennedsgaard TW, Klaas I, Nissen TB, Thamsborg SM, Østergaard S: **Development and daily management of explicit strategy of nonuse of antimicrobial drugs in twelve Danish dairy herds.** *J Dairy Sci* 2006, **89**:1842-1853.
- Onwuegbuzie AJ, Leech N: **A call for qualitative power analysis.** *Qual & Quan* 2007, **41**:105-121.
- Aagaard-Hansen J: **The challenges of cross-disciplinary research.** *Soc Epi* 2007, **21**:425-438.
- Dohoo IR, Martin W, Stryhn H: *Veterinary epidemiologic research* AVC Inc., Charlottetown, Prince Edward Island, Canada; 2003.
- Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS: **Evidence based medicine: what it is and what it isn't.** *BMJ* 1996, **312**:71-72.
- Bruun J, Ersboll AK, Alban L: **Risk factors for metritis in Danish dairy cows.** *Prev Vet Med* 2002, **54**:179-190.
- Goshen T, Shpigel NY: **Evaluation of intrauterine antibiotic treatment of clinical metritis and retained fetal membranes in dairy cows.** *Theriogenol* 2006, **66**:2210-2218.
- LeBlanc SJ, Duffield TF, Leslie KE, Bateman KG, Keefe GP, Walton JS, Johnson WH: **Defining and diagnosing postpartum clinical endometritis and its impact on reproductive performance in dairy cows.** *J Dairy Sci* 2002, **85**:2223-2236.
- LeBlanc SJ, Duffield TF, Leslie KE, Bateman KG, Keefe GP, Walton JS, Johnson WH: **The effect of treatment of clinical endometritis on reproductive performance in dairy cows.** *J Dairy Sci* 2002, **85**:2237-2249.
- LeBlanc SJ: **Postpartum uterine disease and dairy herd reproductive performance: A review.** *Vet J* 2008, **176**:102-114.
- Kristensen EL: **Valuation of dairy herd health management.** In *PhD Thesis Faculty of Life Sciences, University of Copenhagen, Denmark*; 2008.
- LeBlanc SJ, Lissimore KD, Kelton DF, Duffield TF, Leslie KE: **Major advances in disease prevention in dairy cattle.** *J Dairy Sci* 2006, **89**:1267-1279.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp



3.6 Clinical field trials in a dairy herd health management program: treatment effectiveness on milk production in case of early postpartum vaginal discharge

Manuscript V

D. B. Lastein & C. Enevoldsen
Department of Large Animal Sciences
Faculty of Health and Medical Sciences
University of Copenhagen
Grønnegårdsvej 2, DK-1870 Frederiksberg C
Denmark

Clinical field trials in a dairy herd health management program: treatment effectiveness on milk production in case of early postpartum vaginal discharge

D. B. Lastein ^{a*} & C. Enevoldsen ^a

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

*Corresponding author: dorte.bay@gmail.com (Lastein, D.B.): phone 0045-20641151

Abstract

Background

Effect evaluation of therapeutic intervention in veterinary practice and herd health management programs (HHMP) could potentially be improved by adding a 'locally customized trial approach' to the veterinarian tool box. We explored the practical potential and limitations of a customized trial approach aimed at estimating the difference in treatment effect and the (disease) effect of metritis based on vaginal discharge despite treatment on financially relevant performance measurements: predicted energy-corrected milk yield at 60 days postpartum and 305 days total yield.

Results

Pragmatic, 'within practice', multi-herd, ear tag-allocated, non-blinded active controlled clinical field trials in four private Danish dairy herds were integrated into the HHMP for one year. We allocated 136 cows with vaginal signs of metritis before 21 days postpartum to two treatment protocols (penicillin or tetracycline) and included 744 non-metritic cows in the analysis. We experienced some analytical problems related to small, unbalanced group size because of low disease incidence and the 'pseudo-random' allocation procedure. In addition, variance heterogeneity was problematic for the model of total milk yield. We found no statistically significant systematic treatment effects of the two protocols on short- or long-term milk yield. The disease effect despite treatment was inconsistent and differed in both magnitude and direction depending on herd. Adjustment for parity and retained placenta was required but did not interact with treatment.

Conclusion

A 'herd-specific trial approach' can be used as a practical and feasible supplement to a highly structured dairy HHMP for improving evaluation of the effects of interventions like therapeutic treatments. Estimates of different effects can be obtained through a relatively pragmatic and simple data collection and

corresponding statistical analysis. No evidence of differences in treatment effect on milk yield between the antibiotic protocols was found in these trials; however, heterogeneity of disease effect despite treatment was evident across herds. Despite the motivation of veterinarians and farmers and professional supervision, the obtained data quality and non-adherence to the protocol emphasize the importance of interactions between humans, practicalities and data even in these HHMPs.

Keywords

Clinical field trial, herd health management, dairy, metritis, milk yield, effectiveness, treatment, pragmatic, herd-specific

Introduction

Veterinary medicine has developed by continuously implementing discoveries from human and veterinary medicine research into general rules and techniques for daily ‘best practices’. Currently, as veterinarians organize into larger groups, consensus about best practices ought to be ensured in the practice unit. Therefore, the veterinary practice unit needs to become more involved in systematic evaluation of the discoveries in scientific research and the scientific evidence for the practices they currently apply. Clients and veterinary authorities also focus more on documentation for the applied interventions, including prudent use of medical treatments, especially antibiotics. Computerized data recording has facilitated analysis of data in practice. For these reasons, it is both relevant and possible for a professionally working group of veterinarians to set up their own system for providing scientific evidence about the effects of current and new interventions in affiliated herds. Because of the large population size and increased automation of data collection, this capacity is particularly relevant and feasible in dairy herds.

The overall objective of this study was to demonstrate the implementation of herd specific randomized clinical field trials in a dairy herd health management program (HHMP) using systematic clinical examination in the early postpartum (pp) period, as proposed in related work (Lastein, 2012). We suggest that a relative pragmatic trial approach aiming at evaluating effectiveness instead of efficacy (Zwarenstein et al., 2009), will allow well-qualified veterinarians in a highly structured veterinary practice to continuously generate new analysis-based knowledge and evaluation techniques of relevance to veterinary practice. This particular implementation of a multi-herd clinical field trial exemplifies an approach to evaluate the effectiveness of medical treatment of early postpartum metritis diagnosed by means of vaginal discharge. The specific objectives of the study were as follows:

1. to describe and evaluate the conceptual and practical experiences, both the potential and limitations, of clinical field trials integrated into the HHMP context;

2. to estimate the effectiveness of different metritis treatment protocols on milk yield;
3. to estimate the disease effect of metritis despite treatment on milk yield; and
4. to evaluate the influence of prognostic factors [herd, parity (PAR), retained placenta (RP)] on difference in metritis treatment effectiveness and the disease effect of metritis despite treatment

Material and methods

The trial context

A clinical field trial was conducted in 4 herds from June 2008 to May 2009 to evaluate the integration of trials into a HHMP being run in Danish commercial dairy herds. The final dataset was extracted from the 'Danish Cattle Database' (DCDB) in August 2010. Representatives of one veterinary practice had volunteered to participate because of their interest in increasing the effectiveness of metritis treatments. The herds were selected due to their inclusion in a highly structured HHMP offered by the practice and the herd owners' motivation and willingness to participate. In other words, the herds were not selected because of perceived problems with insufficient effectiveness of metritis treatments. The experiments were not financially subsidized in any way. Consequently, the final sample size was determined by the incidence of metritis during the enrolment period (1 year), not by a priori estimation. Trial objectives, design, and protocol were developed in close cooperation with the veterinary practice and herd owners through several meetings between the first authors, the veterinarians, and the farmers (Lastein, 2012). The treatment protocol is described in Table 1. Thus, the design framed what was perceived as meaningful and practically feasible in the given context of one Danish veterinary practice in these 4 herds (e.g., this is referred to as 'customized' or 'herd-specific'). The metritis diagnosis was based mainly on a vaginal discharge score (VDS; described in detail later). Two medical treatments for metritis were compared to estimate differences in treatment effectiveness: the old standard treatment as the active control and the potentially new standard treatment as the experimental treatment. Cows classified as non-metritic were included in the analysis to allow estimation of a potential disease effect despite treatment. The medication principles were as follows (details in Table 1):

- 1) Experimental treatment (PENICILLIN): 3 consecutive days of benzylpenicillin procaine intramuscular (IM) + intrauterine (IU) application of penicillin/streptomycin/sulfadimidin pessaries
- 2) Control treatment (the standard regime of the participating herds)(TETRACYCLIN): 3 consecutive days of oxytetracyclin IM and IU application of oxytetracyclin pessary

Parallel group trials with identical trial protocols were implemented in four private herds within one veterinary practice. The trials were conducted as an integrated part of an extended HHMP where all cows were planned to be vaginally examined by the veterinarian ('gloved hand') 5–21 days after calving at the veterinarian's weekly/fortnightly herd visits. VDS, vaginal wall lacerations, ketosis test on urine or milk, and

general appearance were evaluated, scored, and recorded on paper. The veterinarian was intended to examine all cows between 5-21 days postpartum at a 'planned herd visit' (preferably as soon as possible

Table 1. Inclusion criteria and protocols for a multi-herd clinical field trial conducted to estimate effectiveness of medical treatment of metritis based on vaginal discharge in 4 Danish dairy herds.

Inclusion criteria	Methods/label	Threshold for inclusion/treatment
Vaginal discharge score (VDS) Vaginal exploration – 'gloved arm'	0–9	≥ 4 (equal to or worse than considerable volume of mucopurulent discharge and no smell)
		If VDS=4, then a minimum of 1 positive of the following 3 criteria
Ketosis	Cow-side test: Urine (Ketostix®, Bayer Diagnostics Europe Ltd. Dublin, Ireland) Milk (KetoLac® BHB, Sanwa Kagaku Kenkyusho Co. Ltd., Nagoya, Japan)	>4 mmol/l (moderate) >200 µmol (≥ 3)
Vaginal wall lesions	Present/not present	Any lesions of the vaginal wall
General condition	0 (not affected)/1 (affected: rectal temperature >39.5°C or dull appearance)	1
Exclusion criteria		Caesarean Escape therapy: if more than 2 treatments (of 3 days duration) were required, any antibiotic treatment could be initiated by the veterinarians
Treatment group (Tx)	Treatment initiated before 21 days pp/re-treatment until 30 days pp	
1 PENICILLIN	3 days intramuscular injection with 50–60 ml Penovet® Vet (Boehringer Ingelheim, Copenhagen, Denmark) (300,000 IE/ml) + intrauterine application of 3 pessaries Sulfa-streptocillin® (Boehringer Ingelheim, Copenhagen, Denmark)	
2 TETRACYCLIN	3 days intramuscular injection with 50–60 ml Aquacycline (10%)® Vet (Ceva Animal Health, Vejle, Denmark) + intrauterine application of 1 pessary (500 mg) Terramycin®(Orion Pharma, Nivå, Denmark)	
3 NON-METRITIC	Non-metritic: cows with no metritis treatments	

from 5 days postpartum) and allocate those cows that met the inclusion criteria (see below) according to a within-herd 'pseudo-random' allocation of the cows by ear-tag (even/uneven numbers) to one of two different treatment groups within each herd ('stratified randomization') (Dohoo et al., 2003). Consequently, the veterinarians were in charge of both the allocation and the treatment procedures and were not blinded to these procedures. Examinations were predominantly performed by the same two veterinarians. However, four other veterinarians from the same practice were intermittently involved in data collection, as well. Agreement among the veterinarians about the VDS scale was evaluated in a preceding pilot study (weighted kappa=0.648 [0.62;0.67]) (Dohoo et al., 2003; Lastein and Enevoldsen, 2010). All data on clinical scores, medical treatments (including the trial treatments), and milk production (11 annual test days in a national scheme) were recorded in the DCDB. The veterinarians were responsible for recording both the clinical scores and the initial medical treatments of genital diseases in the DCDB. The herd managers were

responsible for recording medical follow-up treatments, dates of calvings, complications at calving (e.g., dystocia), and culling dates in the DCDB.

As a practical issue, we accepted that cows could be gynaecological examined and treated on days other than the 'herd visit', being 0-21 days postpartum). On such occasions, the defined inclusion criteria were to be followed. However, VDS was not recorded if the date of the herd visits with clinical examination differed from the treatment date. Such irregularities can be considered as 'non-adherence' in less pragmatic designed trials. Pragmatic and financially relevant results measurements (outcomes) related to milk yield were chosen for the trial: predicted milk yield (energy corrected) at 60 days pp as a short-term result, and 305-day total milk yield as a long-term result.

Several information meetings for veterinarians and farmers were held before and during the trial with emphasis on understanding of the trial design, clinical examination, and VDS, the principles and advantages of randomization (e.g., prevention of preferential treatment), adherence to protocol, data quality of treatment recording, and reduction of loss to follow-up. All farmers gave informed consent.

Trial protocol

The VDS was an ordinal 0 to 9 scale with increments of 1 where 0 indicated a minimum of transparent discharge and 9 indicated large quantities of fetid discharge. A VDS of 5 and above indicates an abnormal fetid odour. For a more detailed description, we refer to our related work (Lastein, 2012). Table 1 shows diagnostic criteria for inclusion and the protocol for medical treatment. The inclusion criteria are largely comparable to the definitions of clinical and puerperal metritis proposed by Sheldon and co-workers (Sheldon et al., 2006): a VDS between 5 and 6 represents clinical metritis and a VDS between 7 and 9 represents puerperal metritis. We therefore adhere to the use of the classification term 'metritic' in the present article despite the fact that the specified inclusion criteria were not validated in a trial setting with a negative control group and that some cows with VDS=4 can be included in the metritic group.

We defined the following variables for the data analysis: Herd represents identification numbers 1 to 4. PAR represents first, second, and later lactation cows. RP represents presence or absence of veterinary registration of treatment of RP \leq 5 days after calving. According to the protocol, RP was treated with the same regime as metritis. Additional medical therapy for other disease entities (e.g., antibiotic for mastitis and/or steroids for ketosis) was allowed within the protocol. The only restrictions were on the antibiotics for metritis in the two treatment groups.

Data management

During trial conduct and prior to analysis, the first author performed a manual check of the records of medical treatments to detect errors and irregularities. This procedure showed that errors were common but also that the reasons for errors differed from herd to herd. Errors related to missing recording of 2nd

and 3rd treatments (farmer's procedures) and divergence between initial and subsequent disease code recordings were dominant (e.g., follow-up treatment was recorded with another disease code). The veterinarians and farmers were faced with these errors and had reasonable explanations of why the patterns of error occurred. The importance of this type of 'recording non-adherence' was evaluated. As the errors were mainly indicative of 'deliberate qualitative data manipulation to obtain practical recording procedures' and common error in practical coding of different diseases (metritis vs. retained placenta), the data was edited according to the principles in Table 2. The editing process led to the following case definitions used in the final dataset:

- The first registration of RP should be made before or at 5 days pp. If RP was recorded at >day 5, the code was changed to metritis.
- The initial metritis treatment had to occur in the 0–21-day period pp. Initial recordings after 21 days postpartum were not considered in the study.
- Follow-up treatments for RP and metritis were edited to correspond to the initial treatment code.
- Re-treatment of metritis could occur at any time before 30 days pp. The definition of 'retreatment' was coded as follows: the difference in two initial metritis recording dates was equal to or greater than 4 days (the protocol prescribes 3 days of treatment).
- Discontinuation: Cows requiring more than two treatments (of 3 days duration – initial treatment and first re-treatment) were excluded from the trial because escape therapy was permitted for these cows. We have no information about whether or not alternative protocols were used for such cows.

Table 2 shows examples of initial coding in DCDB and subsequent editing to illustrate the applied editing principles. The final coding of RP, metritis (given as MET), and re-treatment (given as RT) are shown in Table 2.

Methods of analysis

In our data, we theoretically have access to complete and uniform information from all cows and all herds, which allowed us to use a multivariable analysis that provides estimates of possible treatment and disease heterogeneity between herd and other relevant subgroups/prognostic factors. By 'heterogeneity of effect', we mean whether the magnitude and/or direction of treatment and/or disease effects depend on the level of prognostic factor(s). Heterogeneity of effect is evaluated by a statistical interaction between treatment groups and the prognostic factors. If the treatment effects are homogeneous and consistent (same direction) in all herds and subgroups, we have solid evidence that we can issue the same recommendation concerning the treatment protocol in all herds and all subgroups (the use of a 'practice standard protocol/treatment'). Essential subgroups, beyond herd, in this study are cows with RP and cows in

different PAR. The reason for choosing to include these subgroups are that the pathogenesis of metritis is expected to be different in cows with or without previous RP, and the difference in treatment effect has been demonstrated for an Israeli context (Goshen and Shpigel, 2006). Similarly, metritis is known to be related to dystocia and probably other conditions occurring at parturition. Because these conditions are strongly associated with PAR, we need to evaluate all combinations of factors. Finally, despite the attempt to randomly allocate cows, the number of cows may have come out differently in the intervention groups (unbalanced). The adequacy of the randomization procedure applied is assessed in the results section and discussed. For these reasons, we applied a multivariable analysis to account for unbalances of important known predisposing factors if they occur in the data despite randomization (Dohoo et al., 2003).

Table 2. Errors in disease coding that were detected in the trial data from the Danish Cattle Database. Missing registrations and incorrect recordings of follow-up treatment were re-coded. Examples of the definition of retreatment (RT) and discontinuation are shown. Metritis (MET). Retained placenta (RP).

Days postpartum														Final dataset corrections		
	0	1	2	3	4	5	6	7	8	9	10	≤30	RP	MET	RT	
Missing registrations																
						MET		MET					0	1	0	
		RP					MET						1	1	0	
Wrong 'follow-up' registration																
						RP	MET	MET					1	0	0	
				MET	RP	RP							0	1	0	
				RP	MET	MET							1	0	0	
Combination of missing registration and retreatment																
							MET		MET			MET	0	1	1	
Example of discontinuation																
			MET	MET	MET		MET	MET	MET			MET	0	1	1	

Because the trial was designed as a component of the HHMP, the trial had a pragmatic objective (Lastein, 2012), which dictated an analysis by the 'intention to treat' (ITT) principle (Thorpe et al., 2009). The ITT principle implies that all allocated cows were to be analysed according to randomization and not according to 'per protocol' or 'actually received treatment' (that is, we included cows that were non-adherent to the protocol). In the ITT analysis, we used a superiority test, which tests whether one treatment is different from another (Habicht, 2011). The ITT analysis was expected to give a conservative estimate of difference because 'random non-adherence' to protocol would dilute any true difference in the observed data and give a result that can be used as a 'guide for practical decision making in real world situations' (Zwarenstein et al., 2009). Because we included two treated groups and one group of 'non-metritic' cows in our analysis,

we could evaluate both difference in treatment effect between the treated groups and disease effect despite treatment (difference between metritic treated cows and non-metritic cows) accounting for other influential prognostic factors. Cows were the unit of concern.

Lactation curve models – estimation of the outcome variables ECM60 and ECM305total

Test-day lactation curve models of repeated records within lactation were used to estimate both the short- and long-term milk yield for each cow (the results measurements or the Y-variables). Milk yield (kg) and fat and protein concentrations in milk were recorded 11 times a year for each cow in all herds. These records were transformed into energy-corrected milk at test day (ECMT) with the following formula:

$$\text{ECMT} = \text{milk yield}[\text{kg}] * ((0.383 * \text{fat}\% + 0.242 * \text{protein}\% + 0.7832) / 3.140)$$

Model 1 was used to estimate parameters in a lactation curve for each cow. The model is a random coefficient (mixed) model, which defines a piecewise linear function with a 'break' at 60 days pp. Each test day observation thus consists of ECMT and a variable representing the number of days since calving pp or days in milk (dim). The lactation model is based on the following variables: 1) $\text{dimun60} = (\text{dim} - 60) / 60$ if dim is less than 60, and otherwise $\text{dimun60} = 0$; and 2) $\text{dim60} = (\text{DIM} - 60) / 245$ when $\text{dim} > 60$ and other $\text{dim60} = 0$. Dimun60 describes the change (slope) in ECMT before 60 days pp and dim60 describes the change (slope) in ECM after 60 days pp. The intercept of the model describes ECM at 60 days pp. Model 1 is specified in detail by Krogh (2012) as follows below.

A maximum likelihood analysis was conducted separately for each treatment group, within each PAR, and within each herd (Proc Mixed, SAS 9.2) (SAS Institute Inc, 2003). Model 1 allows for completely different shapes of the lactation curve for each individual cow in the group (random intercept and random slopes for each cow). The parameter estimates for each cow were used to estimate the (predicted) daily ECM yield at 60 days pp (ECM60) and the total ECM yield during the first 305 days pp (ECM305total). The major advantage of a model like model 1 compared to simply calculating average milk yield for each cow is that the information from other cows in the analysis allowed us to predict milk yield after early culling (extrapolation of the slope). Prediction was made with as little as one test day. This approach was expected to minimize selection bias on effects measurements resulting from possible premature culling due to metritis or one particular treatment. Because the analysis was conducted within herd, within PAR, and within treatment group, we maintained the characteristics of the treatment and disease (metritis), herd and parity within the predicted estimates. The use of predicted values could lead to a conservative estimate of milk loss if culling because of metritis in general or either of the protocols was frequent before first milk yield recording. We excluded milk yield recordings after 400 days pp and the last milk yield record

in each lactation due to factors like an expected risk of increased variation near drying off, late pregnancy,

$$ECMT_{ij} = \beta_{0j}DIMun60_{ij} + \beta_{1j} + \beta_{2j}DIM60_{ij} + \varepsilon_{ij}$$

where

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

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

$$\beta_{2j} \sim \beta_2 + \mu_{2j}$$

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

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

where i = test date and j = cow.

(1)

Model 1 - description: $ECMT_{ij}$ (outcome) represents the milk yield in kilograms/day on test day i of the j^{th} cow. β_{0j} represents the slope of the lactation curve from 0 dim ('days in milk') until 60 dim for the j^{th} cow. β_{1j} (intercept) represents the estimated milk yield in kilograms/day at 60 dim for the j^{th} cow ($dimun60=0$ and $dim60=0$). β_{2j} represents the slope of the lactation curve from 60 dim until 305 the j^{th} cow. β_0 , β_1 , and β_2 represent the fixed effects or the average coefficients for all cows in the analysis. The random estimates μ_{0j} , μ_{1j} , and μ_{2j} represent the individual j cows' deviations from the corresponding fixed effects. The random variation parameters μ_{0j} , μ_{1j} , and μ_{2j} , are assumed to have normal distribution with zero mean. The parameter estimates $\sigma_{\mu 0}^2$, $\sigma_{\mu 1}^2$, and $\sigma_{\mu 2}^2$ represent the variance of the fixed effects of $dimun60$, intercept, and $dim60$, respectively ('the variation around the average effect'). The parameter estimates $\sigma_{\mu 01}^2$, $\sigma_{\mu 02}^2$, and $\sigma_{\mu 12}^2$ represent the covariance between $dimun60$ -intercept, $dimun60$ - $dim60$, and intercept- $dim60$ ('determines the correlation between slopes and peak'). The residuals (ε_{ij}) represent the variation between test days and are assumed to follow a normal distribution with zero mean, having a variance of σ_{ε}^2 .

or milking frequency. Test-day results that had missing or zero values for fat percentage, protein percentage, or kilograms of milk were also excluded from the data file. We chose to include all cows that had at least one test day, as mentioned. If more than two lactations from the same cow occurred, only the first was retained. The models were run with the between-within option for degrees of freedom and an unstructured variance structure (no assumptions of variance structure). If converge problems emerged, then a variance component structure was used that works under the assumption that all variances are equal. Similar models were used in analogous analyses (Bennedsgaard et al., 2003; Krogh, 2012) and also implemented in other parts of the Danish HHMP.

The point estimate at day 60 was used directly as outcome for an effect model (ECM60). Based on the point estimates at days 0, 60, and 305, the outcome $ECM305_{total}$ was calculated as the area under the curve from day 0 to day 305 for each cow and used as the outcome for another effect model ($ECM305_{total}$).

Effect models

Both outcome variables (ECM60 and ECM305total) in the trial were analysed with a multivariable analysis of variance (ANOVA) (Proc GLM, SAS 9.2.) (SAS Institute Inc, 2003). We specified the following initial (full) least squares ANOVA model 2:

Y (ECM60 or ECM305total) =

Tx Herd PAR RP (main effects)

Tx*Herd Tx*PAR Tx*RP Herd*PAR Herd*RP PAR*RP (two-way interactions)

Tx*Herd*PAR Tx*PAR*RP Herd*PAR*RP (three-way interactions)

Tx*Herd*PAR*RP (four-way interaction)

(2)

Because treatment group (Tx – including the two intervention groups and the non-metritic group) was the factor of primary interest, it was forced in as a fixed variable in both models. Attempts to reduce model 2 were made by manual backward selection according to the hierarchical principle, which means that the most complex terms were removed first while retaining all less-complicated terms. Significance level is set to $p=0.05$. The principle also implies that only p -values (F-test) for the most complex term can be used for evaluation. One term was removed at a time. Prognostic terms stayed in models based on a qualitative judgment of the statistical significance, the number of observations and their mean values within each comparison group, and the mean square error (MSE). This model-building strategy was chosen to avoid overlooking any potentially relevant interactions related to the treatment/disease effect. There are three assumptions for ANOVA: independent observations on the dependent variables, normal distribution of residuals, and homogeneous variances across groups. The independence of observations was ensured by including herd as a fixed effect in the analysis and excluding observations exceeding one lactation per cow. Standardized residuals were used for graphical evaluation of the assumption of normal distribution and homoscedasticity after model reduction. In addition, variance homogeneity was evaluated by recoding all combinations of variables into one variable and testing whether the variances between each combination were different (Levene's test). A further model check was performed by examining individual observations with extreme residuals (above 3 or below -3), leverage (threshold 0.9), and deviating CookD-values (Dohoo et al., 2003). Variables were kept in the model if no obvious reason to exclude them (e.g., typing error) was present. In case of significant interaction terms these were evaluated by sliced analysis (Least square [LS] means statement – slice option) was performed to compare the overall difference between treatments (Tx) within herd (F-test). Statistical significance of the differences in estimates of LS means (means, adjusted for

the covariates) was tested with the F-test. If the F-test of the slice was statistically significant, we tested all possible differences between the treated metritic cows (Tx1 and Tx2) and the non-metritic cows (that is, differences in treatment and disease effect despite treatment on ECM60 and ECM305total) within herd.

We performed a sample size estimation running different scenarios for milk yield at 60 dim (Proc power, SAS 9.2) under the assumption of a balanced design, a two tailed superiority test and a 95% confidence level for one herd. Seven scenarios with combinations of difference in treatment effect at 1, 2, 3 and 5 kg ECM and SD at 5 and 8 kg ECM, respectively. The simulation indicated that sample sizes between 50-200 cows with power at 80-90 % would detect a difference of at least 2-3 kg milk. The simulation did not take into account clustering effect, interactions terms, and non-adherence and may thus underestimate the required sample size.

Results

Descriptive analysis

A total of 942 cows (15.4% with a metritis treatment) calved in the study period and six cows were excluded due to caesarean section or discontinuation/escape therapy. After editing according to the exclusion criteria for milk records, the lactation curve models produced a total of 880 lactation curves with estimates of ECM60 and ECM305total available for analysis with the effect models. The milk yield estimates were based on 6754 test day records (average 7.7 test day records per lactation [min. 1; max. 12]). In total, 136 of 880 (15.5%) of the 880 cows were metritic according to the trial protocol (Table 3).

Table 3. The distribution of herds, associated herd veterinarian (Vet), breed, number of calvings in study period, % of calvings with metritis, number of cows included and excluded, and % included with metritis in a multi-herd clinical field trial integrated into a Danish herd health management program.

Herd	Vet	Breed	# Cows calved	# Metritic (%)	# Cows excluded*	# Cows included	# Metritic in study (% of included)
1	1	Holstein	467	77 (16.5%)	1/3/25	438	70 (16.0%)
2	1	Holstein/DR**	196	18 (9.2%)	1/0/11	184	18 (9.8%)
3	1	Holstein/DR**	159	20 (12.3%)	1/0/8	150	20 (13.3%)
4	2	Holstein	120	30 (25%)	0/0/12	108	28 (25.9%)
Total			942	15.4%	62 (6.7%)	880	136 (15.5%) Non-metritic: 744

* Exclusion due to caesarean/discontinuation (potential escape therapy)/ exclusion due to data management of missing or zero values for fat percentage, protein percentage, or milk yield or less than 2 test day records as last record was not used for analysis.

**DR = Danish Red

The distribution of cows in treatment groups within herds and their prognostic attributes are shown in Table 4 (no statistical tests). Initially, an unbalanced distribution of cows to the two treatment groups of metritic cows in herd 4 was noted (21 versus 7) (Table 4, grey shading, 3rd column). The column values for

Ntotal and %cows with VDS and days pp for herd visits and treatment, respectively, demonstrate the complexity of 'real world data'. Of the total 880 calvings, 89.9% had VDS records and treatment records on the days of the planned herd visit, 6.8% had missing data on VDS either because of no registration of any clinical examination on a planned herd visit or no VDS recorded despite a date of an examination. Furthermore, 1.5% cows and 3.3% cows had VDS records that were recorded before or after the treatment record. The distribution of valid, missing, and irregular VDS recordings differed among herds [e.g., herd 4 had the lowest level (82%) of concurrent planned examinations and treatments and highest level of missing records (16%)] (data not shown). However, the distribution of VDS recordings between treatment groups (Tx1 versus Tx2) was comparable: approximately 75% VDS recordings on herd visits and 15% missing VDS records and approximately 10% VDS records before and 10% after treatment (data not shown). Removal of cows from the herds (culling) are due to both selling live cows or death (slaughter or euthanized or dying by itself). Additional analysis of culling was performed that is the percentage of calvings followed by death (not selling live animals) before 400 dim varied between 11% and 57%. However, within herd, the distributions of death were comparable between treatment groups and between metritic and non-metritic groups of cows. Therefore, we have no substantial evidence that unbalanced removal caused selection bias. In addition, our modelling procedures aimed at correcting for this problem by using predicted milk yield. The distribution of PAR and percentage of cows with RP between Tx1 and Tx2 within herd indicated some unbalance (grey shading in Table 4, last columns). We note that a higher proportion of cows classified as metritic had a diagnosis of RP than cows classified as non-metritic (Tx1: 28.4%, Tx2:41.9%, non-metritic: 8.1%).

Lactation curve model

A total of 36 analyses of lactation curves were conducted (4 herds, 3 PAR groups, and 3 levels of Tx), so that individual yield estimates for each cow were produced. Of these models, 15 did initially not converge with an unstructured covariance pattern, and a variance component structure was used instead, which allowed convergence.

ECM60 model

The reduction of the initial full ECM60 model revealed a borderline statistically significant three-way interaction term involving treatment group (Tx*HERD*PAR) (Type III, F test, $p=0.055$). However, the evaluation of 36 levels in the interaction term showed that the contrasts causing the interaction were very extreme and due to five cows. A removal of the interaction reduced the R^2 -value only from 0.568 to 0.558 and increased the Root MSE from 4.82 kg to 4.84 kg. For these reasons and the borderline level of statistical

Table 4. Distribution of prognostic factors for the two intervention groups (Tx1 and Tx2) and the non-metritic cows in a Danish clinical field trial integrated into a Danish herd health management programme (HHMP).

HERD	Treatment group	No. cows, N _{total}	VDS* (% cows with VDS)	Days pp for HHMP visit	Days pp for treatment	Culling + p10** (% cows culled within 400 days pp)	Parity (1/2/>2) (%)	RP (%) ***
1	Tx1	36	2-5-9 (100)	2-7-11	5-7-14	58-274-469-144 (78)	55.6/16.7/27.8	33.3
	Tx2	34	2-6-9 (100)	4-7-16	4-8-16	53-315-473-122 (82)	41.2/29.4/29.4	47.1
	Non-metritic	368	0-2-5 (95)	-1^8-20	-	62-266-467-148 (84)	35.9/36.1/27.99	8.4
2	Tx1	7	2-4-9 (100)	3-6-10	3-9-11	366-410-707-366 (71)	42.9/28.6/28.6	28.6
	Tx2	11	1-6-8 (100)	5-9-14	6-7-14	298-493-692-298 (64)	27.3/9.1/63.6	27.3
	Non-metritic	166	0-2-5 (96)	1-8-20	-	55-354-735-116 (84)	39.2/27.7/33.1	7.2
3	Tx1	10	1-5-7 (80)	5-8-12	3-9-15	113-352-615-113 (70)	60/20/20	0
	Tx2	10	4-6-9 (90)	4-8-13	1-7-13	143-194-536-143 (90)	50/10/40	30
	Non-metritic	130	0-2-9 (90)	2-9-20	-	50-390-762-96 (73)	35/40/55	11.5
4	Tx1	21	0-8-9 (95)	0-7-12	0-7-12	118-333-540-151 (76)	52.4/28.6/19.1	33.3
	Tx2	7	5-8-9 (86)	3-8-10	7-9-11	136-267-409-136 (86)	28.6/57.1/14.3	57.1
	Non-metritic	80	0-1-6 (81)	-1^7-18	-	51-273-697-113 (80)	37.5/40/22.5	2.5
Total	Tx1	74	0-5-9 (95)	0-7-12	0-7-15	58-314-707-150 (76)	54.1/21.6/24.3	28.4
	Tx2	62	1-6-9 (97)	3-8-16	1-8-16	53-319-692-136 (80)	38.7/25.8/35.5	41.9
	Non-metritic	744	0-2-9 (93)	-1^8-20	-	50-292-762-140 (82)	35.2/33.7/31.1	8.1

The description in the table presents the reduced dataset after implementation of exclusion criteria and data management preparing for effect model building. All variables are presented as min-median-max unless otherwise noted.

*All VDS recordings included, also recordings before and after the planned HHMP visit.

**Culling represent both removal from herd (live animals), slaughter and death, p10 ~10% percentile

*** RP=retained placenta (farmer's observation/veterinary treatment)

^ 3 cows are identified with examination dates before calving.

Grey shading of cells indicates problems related to randomization (unbalance between Tx1 and Tx2).

significance ($P=0.055$), we removed the three-way interaction from the model. Another statistically significant three-way interaction term was found (HERD*PAR*RP). This interaction was kept in the model according to the modelling strategy described in the methods section. The graphical and manual model validation procedure gave no indication of concern (overall homoscedastic standardized residuals, though some departure from normality within a minor subset of groups; no extreme leverage values; extreme CookD values checked for errors in raw data). Levene's test for variance homogeneity was statistically non-significant ($p=0.065$). However, the 'low' p-value warrants some concern regarding this assumption.

All observations remained in the dataset. Both Tx*HERD and HERD*PAR*RP were statistically significant in the final model ($p=0.002$ and $p=0.001$, respectively). Table 5 shows the LS means (LSM) of ECM60 within treatment group within herd (adjusted for the RP and PAR in the final model), and Figure 1 illustrates the differences graphically. Because of the Tx*HERD interaction, the effect of Tx was estimated separately for

each herd by a sliced analysis (Table 5) that showed that ECM60 was overall affected by Tx in herd 3 (F-test, $p < 0.001$) while the effects were statistically non-significant in the remaining herds (F-tests, $p > 0.15$).

By pairwise comparison of Tx1/Tx2 and non-metritic within herd while adjusting for RP and PAR, we found no statistical evidence of difference in treatment effect (differences between Tx1 and Tx2) in any herds. We found no difference between treated-metritic cows (Tx1/Tx2) and non-metritic cows (no 'disease effect despite treatment') in herd 2 ($p > 0.15$) and herd 4 ($p > 0.33$). In herd 1, we found a borderline statistical significance between Tx1 and the non-metritic group ($p = 0.054$), indicating a difference in ECM60 of 2 kg in favour of the treated metritic cows. In herd 3, we found a statistically significant difference between Tx1 vs non-metritic and Tx2 vs non-metritic ($p = 0.020$ and $p = 0.002$, respectively), indicating a difference in ECM60 of 4–5 kg in favour of the non-metritic cows.

Table 5. The overall average ECM60 across treatment group (Tx1, Tx2, and non-metritic) and herd in the final model to evaluate treatment effect and disease effects despite treatment of metritis in a multi-herd clinical field trial.

	Tx	N	ECM60 (LSM) [kg ECM]	Std error	p Sliced analysis	ECM305total (LSM) [kg ECM]	Std error	P Sliced analysis
HERD 1	1	36	37.5 ^a	0.82	0.153	9510	217	0.373
	2	34	35.9	0.84		9122	220	
	Non-metritic	368	35.8 ^a	0.40		9217	104	
HERD 2	1	7	38.4	1.87	0.151	10521 ^d	492	0.015
	2	11	38.0	1.58		10737 ^e	415	
	Non-metritic	166	35.7	0.68		9695 ^{de}	178	
HERD 3	1	10	30.5 ^b	1.68	<0.001	7975 ^f	441	0.010
	2	10	29.1 ^c	1.57		8202 ^g	413	
	Non-metritic	130	34.2 ^{bc}	0.65		9036 ^{fg}	170	
HERD 4	1	21	27.6	1.23	0.551	6732	323	0.535
	2	7	28.7	1.86		6818	489	
	Non-metritic	80	26.6	1.02		6385	269	

* The predicted ECM60 and ECM305total averages (Least Square Means, LSM) are adjusted for retained placenta and parity. ECM=energy corrected milk.

^{a,b,c,d,e,f,g} Represent pairwise statistically significant differences (p level 0.05 or borderline $p < 0.1$) in ECM60 and ECM305total between combinations of Tx and the non-metritic group (not adjusted for multiple comparison). All other combinations of Tx and non-metritic group within herd are statistically non-significant.

The sliced analysis of all three combinations of the interaction HERD*PAR*RP showed that HERD was statistically significantly associated with ECM60 at all combinations of PAR and RP. PAR was statistically and significantly associated with ECM60 in almost all combinations of HERD and RP (seven of eight were clearly significant, one of eight was borderline significant). However, RP was statistically significantly associated only with some combinations of HERD and PAR (data not shown). The directions of the statistically significant differences were all in favour of higher milk yield in the cows without RP (Figure 2). This result indicates that the disease effect of RP despite treatment differs from herd to herd and from PAR to PAR and is not associated with subsequent treatment for metritis (no interactions among PAR, RP, and Tx).

ECM305total model

The full four-way ECM305total model was reduced according to the described principles. Similar to the ECM60 model, a statistical significance ($p=0.046$) of the three-way interaction term (Tx*HERD*PAR) was considered unreliable because of small group sizes and minor changes to the overall model fit when the interaction term was removed (R^2 reduced from 0.47 to 0.46 and Root MSE increased from 1269 kg to 1276 kg). Model assumptions were checked, and we found some problems related to normality of residuals (minor) and variance heterogeneity (significant Levene's test). Despite these findings, we proceeded our planned analysis, as transformation or exclusion of influential observations would complicate interpretability of interaction terms and appear in-consistent with the pragmatic aim of the trial (e.g., reduce applicability of results to 'all metritic cows'). No data were omitted from analysis. In the final model, both Tx*HERD and HERD*PAR*RP were statistically significant ($p=0.008$ and $p<0.001$, respectively). A sliced analysis of Tx within HERD showed that Tx was statistically significantly associated with ECM305total in herd 2 and herd 3 (F-test, $p=0.015$ and $p=0.010$, respectively). The association was not evident for herd 1 and herd 4 (p values >0.37). The differences in LSM of ECM305total are illustrated in Figure 1, and the LSM within Tx within herd are presented in Table 5. When pairwise comparisons of Tx1, Tx2, and non-metritic within herd were performed, we found statistically significant differences between Tx1 and Tx2 versus non-metritic in herds 2 and 3 (herd 2: $p=0.10$ and 0.01 ; herd 3: $p=0.01$ and $p=0.05$, respectively). In herd 2, the effect of disease despite treatment at 305 days was a predicted ECM approximately 350 kg in favour of the treated metritic cows. In herd 3, the results indicated an effect of disease despite treatment at 305 days with a predicted ECM of between 800–1000 kg in favour of the non-metritic cows.

Overall the results indicate that no difference in treatment effect could be detected in the herds in this trial but that the disease effect despite treatment of metritis differed in quantity and direction between herds.

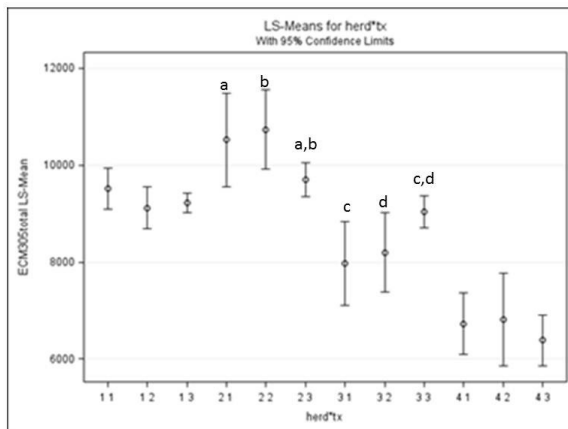
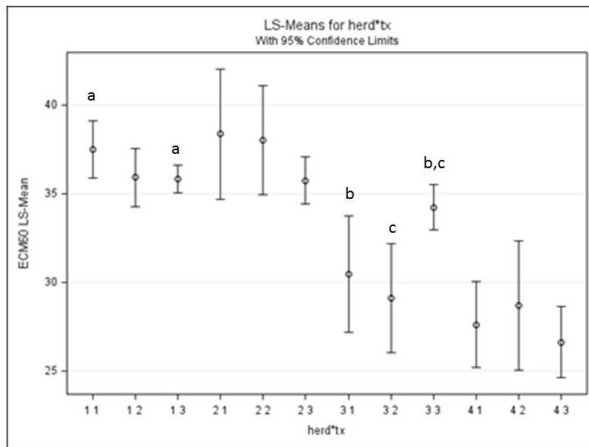


Figure 1. Within-herd (1–4) distribution of predicted energy-corrected milk yield at 60 days pp (ECM60 Least Squaremeans (a) and total yield across 305 days (ECM305total Least Squaremeans) (b) for treated, metritic cows on penicillin protocol (Tx1) or tetracycline protocol (Tx2) and non-metritic cows (Tx3) in a multi-herd trial.*

*For instance; on the x-axis: herd*tx =1.1 equals Tx1 in herd 1 and so forth. Milk yields are adjusted for parity (1., 2.,>2.) and retained placenta (treatment) status. Statistically significant differences ($p \leq 0.05$) are marked with letters above the pairwise differences.

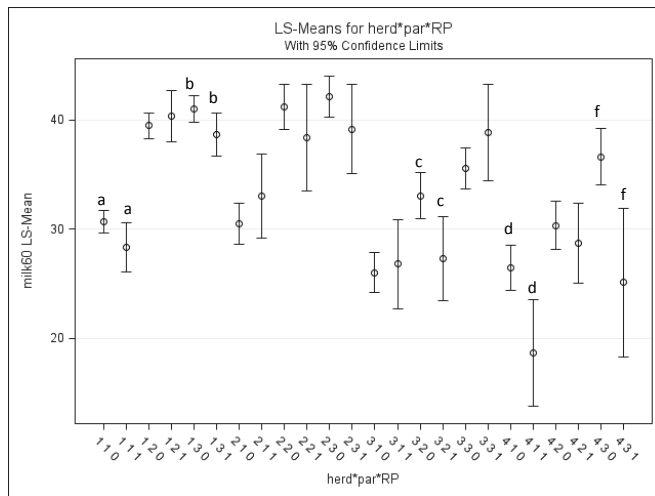


Figure 2. Within-herd (1–4) distribution of predicted energy-corrected milk yield at 60 days postpartum for parity 1., 2., and >2. for cows with and without a diagnosis and treatment for retained placenta (+RP=1, -RP=0). *

*For instance; on the x-axis: herd*par*RP =1.1.0 equals herd 1, parity 1 and no RP and so forth. Milk yields are adjusted for metritis treatment. Significant differences ($p < 0.05$) are marked with letters above the pairwise differences.

Discussion

Our study demonstrated that it is possible to implement and conduct herd-specific, ‘pseudo’-randomized, controlled trial in multiple herds as part of a HHMP in a veterinary practice unit under the pragmatic principles of ITT. We demonstrated no differences in treatment effect on milk yield. From this result follows a practical recommendation in the specific practice context: The choice of antibiotic protocol (penicillin versus tetracycline) should be based on arguments other than difference in milk yield in these 4 herds. Furthermore, we obtained estimates of metritis effect (despite treatment) on milk yield that differed from herd to herd without any consistent patterns of causal effects. From this result follows a practical recommendation in the specific practice context: No practice-specific recommendations on the effect of medication to reduce milk loss caused by metritis can be given. The choice of whether to treat or not to treat metritis and the estimates of the subsequent effect should be given entirely at the herd level. However, both future potential and limitations have become apparent during the implementation process and the statistical analyses. In the following section, we discuss premises for randomized clinical field trials in the HHMP context in addition to a discussion of the results of the statistical analyses.

Potentials of clinical field trials

The trial approach in the HHMP setting allows valid inferences about cause-and-effect relations and can give estimates of the effectiveness of many different types of interventions (being therapeutic or management related) within each herd. Inferences beyond herd or potentially practice level are not

intended. Our 'bottom-up' approach supports design of herd-specific trials that can be altered to suit specific veterinary practice organizations or herd contexts. Similar context-specific inferences on cause and effect relations can be difficult to obtain through observational analysis of local herd-specific data or large multi-herd data files and impossible to obtain from evidence presented in the traditional scientific literature. Therefore, we argue that clinical field trials are valuable supplements to the veterinary HHMP toolbox already consisting of local observational evidence, general evidence, and individual knowledge (possibly tacit), experience, and personal preferences. These elements form a valuable platform for applying best practices according to one definition of 'evidence based (veterinary) medicine' [2].

We found that the veterinarians and farmers showed enthusiasm during the design phase when protocol and practical design should be decided upon. During the trial, discussion groups with farmers, local veterinarians, and the first author were formed. Here, we experienced growing understanding of the principles behind trials, randomization, and the inferential problems related to preferential treatment (non-random non-adherence). This 'ownership' of the trial is described by Farrell et al. as one of many important factors that promote successful conduct of trials (Farrell et al., 2010). Also, these authors recommend that "to overcome barriers to participation, a trial should address an **important** research question, and the protocol and data collection should be as **straightforward** as possible, with demands on clinicians and participants kept to a minimum". We have tried to accomplish these recommendations of importance and simplicity by using the 'bottom-up' strategy in which we gave support to veterinarians' development of their own practice (or herd)-specific protocols. This approach is expected to limit the complexity of a trial. For instance, in this trial, the veterinarians rejected systematic clinical control of the metric treated animals to assess clinical cure, which led to a very pragmatic design and analysis (few exclusion criteria, financially relevant endpoint, pure ITT analysis) (Thorpe et al., 2009; Zwarenstein et al., 2009).

Full acceptance of treatment allocation was obtained for the involved herds. However, we have experienced problems related to preferential treatment in other herds than those described in this study. This drop-out issue at the herd level could be important for future implementation in a broader spectrum of herds with less-motivated veterinarians and herd owners. For instance, one issue could be whether the concept is applicable to herds to guide the therapeutic use of antibiotics prior to a possible drug liberalization process, which might not be of major interest to the veterinarians (lack of ownership).

Limitations of clinical field trials

Data quality concerning dropout at the cow level (e.g., missing data on test day records) and non-adherence to protocol are major issues for trial design and conduct. The qualitative observations during the supervision of the trial at visits in the participating herds and the veterinary practice and the data description above gave us no indications that the randomization or the exclusion of cows from analysis was

influenced by systematic errors (e.g., non-random non-adherence such as personal preferences for certain treatments). During this study, 62 cows (6.7% of all cows that calved) were excluded because of prior knowledge about problems related to prediction of yield of cows with caesarean section, discontinued cows (escape therapy), and cows with missing (dropout at cow level) or unreliable data in test day records. Compared to another multi-herd study (8 herds) of the effects of the dry period, which was controlled by skilled research technicians at weekly visits and where 16% of animals were excluded (Sørensen and Enevoldsen, 1991), we find that the 7% dropout probability in our trial is acceptable and anticipate only a very low risk of bias due to non-random drop-out. We had no support from trained research technicians in our study; the responsibility of inclusion and treatment was left to the local veterinarians under supervision by the first author and milk data collection to the national milk test scheme. However, beyond the 7% excluded because of specified exclusion criteria, we acknowledge that some degree of non-adherence in our data exists (e.g., in Table 2, we have corrected and interpreted errors in recordings after talking to all farmers about their errors in registration patterns instead of excluding them). We also deliberately chose not to fully evaluate all inclusion criteria on every cow and exclude on the basis of 'erroneous inclusion' (e.g., treatments of metritis before or after the planned herd visit). The reason for accepting this 'type of non-adherence' was that to 'evaluate effectiveness in real-world settings to guide practical decision making', such dropout and random non-adherence is a premise that must be minimized but also accepted (Thorpe et al., 2009; Zwarenstein et al., 2009). To allow valid inferences for our HHMP context, a pragmatic trial design and corresponding analytical methodology based on the ITT principles must be coherently implemented. If biological explanations of treatment efficacy (as for registration of drugs) are the aim of a trial, we should use other trial designs, including thorough strict control of protocol adherence and per protocol analysis of data. Consequently, despite the limitation in data quality, we believe that the combined result of the initial introduction of the trial for the farmers and the veterinarians, the training activities for the veterinarians, the practical allocation procedure, and subsequent recordings is very satisfactory and allows us to conclude that a simple trial design like ours is a practically feasible component of HHMPs like the HHMP applied by the participating Danish veterinarians.

During the trial, we experienced some scepticism regarding the study period consisting of the enrolment period (1 year) and the subsequent observation period (collection of milk test day records for approximately 1 year). However, the veterinarians agreed on these long periods to allow estimation of the effectiveness of treatment and disease effects based on the most pragmatic and financially important key performance indicator (milk yield for a full lactation) and to obtain sample sizes that would (hopefully) yield sufficient power for statistical tests (e.g., ability to detect difference if true difference is present). However, the study showed that duration should be carefully balanced with motivation over time and sample size

considerations (especially in cases of low incidence of disease). In future trials, an estimate of effectiveness on peak yield and a herd-specific correlation between peaks and 305-day yield might be used to predict 305-day yield and thus shorten the study period. We acknowledge that the 'within-practice' approach (e.g., clustering effect) and testing of interactions has stressed the sample size to a maximum (e.g., reduced power), as the sample size cannot be adjusted accordingly due to the practical circumstances. The problem with heterogeneity of variances of the predicted ECM₃₀₅total in our study (reduces precision and validity of effect estimates) also indicates that more work is needed for defining and refining a long-term milk yield results measurement and/or the use of alternative analytical methods. A shorter study period with frequent adjustments of research questions and corresponding trials probably would help maintain motivation. Such a continuous cyclic process is similar to the evolutionary operations (EvOp) principle applied in other manufacturing industries (Box et al., 1978; Schwabe et al., 1977).

In general and as mentioned above, the HHMP context imposes practical limitations on sample size with corresponding limitations on the statistical inferences that are possible. The maximum sample size (if enrolling all calved cows to examination and subsequent allocation to treatment) will be totally governed by the inclusion criteria in question, choice of the clinically relevant difference (presumably larger than in more explanatory-type trials), the disease occurrence, and the length of the study period. Especially, the clinically relevant difference deserves attention because it is crucial for the evaluation (prediction) of the possible benefits of new knowledge of effect estimates. In the future we also need to evaluate the potentials of Bayesian statistics for the analysis. The advantage of Bayesian statistics is that we can include prior knowledge in the analysis in a systematic fashion. This inclusion makes good sense because we rarely start from scratch when we examine a herd-problem.

A discussion of misclassification into the metritic and non-metritic groups in our study is highly relevant. The situation of misclassification could be a consequence of poor precision of VDS scoring or bias in clinical judgment. Also, as part of the pragmatic design, we did not control the inclusion criteria after enrolment and did not exclude cows treated for metritis that had VDS and other clinical scores non-adherent with the protocol. We further acknowledge that some cases were treated before herd visits and systematic clinical examination because of 'call on demand' visits (but were still intended to be treated according to protocol). We argue that these cows very well could be the fraction of metritic cows that are severely affected. By including these cows in the analysis, we hoped to demonstrate the practical applicability of the design (e.g., no necessity of authority approval). Whatever the reason for keeping non-adherent cases of any type in the analysis, the probability of finding a true difference in treatment or disease effect despite treatment is expected to decrease when they are retained (e.g., we potentially underestimate the difference in effect of

treatment and the effect of disease despite treatment). This dilution influence could be one reason for some of our findings in both the ECM60 and the ECM305total models. The most fundamental problem related to misclassification still remains being the validity of the inclusion criteria in the Danish HHMP: Which diagnostic criteria for metritis are predictive of milk loss in a given herd (if any)? This problem could be solved by introducing an untreated control group into a clinical field trial like ours. The participating veterinarians and farmers were not willing to conduct negative controlled trials thus the inclusion criteria could not be validated in this particular trial. However, if negative controls were included, the protocol related to discontinuation (escape therapy) and exclusion criteria should be altered because systemically affected diseased cows cannot legally be withheld treatment in Denmark. In our study, an untreated control group was not chosen by the participating veterinarians and farmers for financial, ethical, or legal reasons (e.g., metritis with toxæmia is a potentially life-threatening disease).

The issue of the overall validity of the inclusion criteria is central both in the herd/practice trial context and as seen from a national/industry level perspective (e.g., efforts at reduction of antibiotic usage). That is, is the chosen threshold of treatment appropriate in some or all of these herds? In this trial, the inclusion criteria (equal the treatment threshold) as chosen by the participating veterinarians were: treatment in case of mucopurulent non-odorous discharge ($VDS \geq 4$) and related clinical signs of disease (see Table 1). We regard this threshold as the lowest meaningful threshold if treatment should not be considered as preventive medication with antibiotics, which is prohibited in Denmark. This 'low' treatment threshold might increase the likelihood of misclassification. Such a misclassification would have twofold consequences: from a national perspective, the unnecessary use of antibiotic, and from a trial perspective, the further dilution of the possibility of estimating the 'true disease effect' (despite treatment). Although a recent Danish observational study found that a similar criterion ($VDS \geq 4$) was associated with an impairment of reproductive performance (Elkjær, 2012), we could argue that the predictive threshold for milk yield could very well be different and herd dependent because of different disease effects of metritis and RP. It is very probable that milk yield primarily is reduced among cows that are systemically affected by their pp disease. We argue that these cows are represented by either increased temperature and/or VDS worse than 4 (4 defined as mucopurulent, non-smelling discharge) and/or other clinical sign of systemic affection (reduced feed intake, dullness). A validation procedure for the inclusion criteria for treatment based on vaginal discharge and the status of RP is an obvious extension of this clinical field trial that indicates inconclusive disease effects. Such a validation process should involve both improvement of the application of the VDS score and other clinical signs and comparison of different threshold strategies. Also, our trial data could have been evaluated with multiple endpoints (milk yield and reproduction) to evaluate potential differences in validity of inclusion criteria for these different goals.

The 'non-blinded randomization and allocation procedure' dividing the metritic cows into the two treatment groups (Tx1 and Tx2) was based on ear-tag allocation (even/uneven numbers). The procedure was a very practical solution applicable in all herds using consequent ear tagging. The method appeared valid in most herds (balance in total number of cows in groups and prognostic factors). However, some problems related to the procedure became evident in herd 4, where it resulted in severely unbalanced groups. The reason for this skewness is unknown because the farmer claimed that he had ear-tagged the cows chronologically as new-born calves. Alternative methods could be tested in future trials [e.g., allocating according to calving dates, alternating treatments, and random number charts]. Also, stratification based on important prognostic factors (e.g., PAR and RP) and randomization within these blocks are alternative practically applicable methods to ensure balanced allocation to treatment groups.

Effectiveness of treatment and disease effect despite treatment

This study demonstrates how a pragmatic approach and ITT analysis of a multi-herd, active controlled trial including non-diseased cows can be used to evaluate effects of medical intervention scenarios. However, caution is warranted with respect to the interpretation of the estimates of effect due to some signs of variance heterogeneity (an important ANOVA assumption).. We have deliberately omitted transformation and elimination of outliers to correct for the variance problems. We decided to do so to stay in line with our pragmatic trial strategy (the real world is heterogenic) and to ease interpretation. We chose to use predicted values to minimize selection bias (mainly a problem for 305 days yield) and we modelled the predicted yield in subgroups to maintain the characteristics of the important independent variables in the estimates that were later used in model 2. We acknowledge that this choice potentially could increase the estimates of effectiveness and the heterogeneity of effect due to HERD and PAR compared to using, for example, yield at first test day. This alternative would require some adjustment for stage of lactation, which is achieved effectively with our model. The opposite directions of effects in our study could not be due to some pre-adjustment of the results measurements causing merely increased precision. .

In our analytical set-up a p-value for the fixed effect of Tx can mean multiple things. Initially, as main effect Tx only tells us whether there is a statistically significant difference in milk yield of between any combinations of the two intervention groups and the non-metritic cows or not. In our case, as we also tested for every possible interaction with the prognostic factors (HERD, PAR, RP) and found some of these statistically significant, evidence of 'general effect of the Tx variable' was quickly rejected. We had to explore the 'heterogeneity of effects' – the dependency of the prognostic variables on the effect of treatment of metritis on milk yield. Especially the 'local or context-specific' effect of herd is of special interest in the HHMP context. By our model, we would be able to conclude on the following scenarios;

general or local (herd) differences in treatment effect and heterogeneity across prognostic factor(s) and general/local (herd) disease effect despite treatment and heterogeneity across prognostic factor(s).

Below we discuss the herds separately (or in groups) to illustrate the inconsistency in the statistically significant estimates obtained:

- In herd 2, metritic cows with treatment tended to be higher yielding than non-metritic cows. An explanation for this result could be a higher risk of metritis among high-yielding cows [a tendency also seen in Goshen et al (2006)] (perhaps mediated through breed differences); an indirect positive effect on milk yield of prolonged non-pregnancy status caused by uterine infection (Dohoo and Martin, 1984); a higher risk for other disease (resulting in milk loss) among non-metritic cows; a reduction in milk loss from early antibiotic treatment due to other diseases (e.g., mastitis) in metritic cows; extreme misclassification error (too many ‘metritic’ cows experience no milk yield loss, so treatment threshold was too low); or some other unknown (to us) non-adherence to the randomization procedure (preferential allocation). Our results imply that it could be necessary to account for yield in previous lactation (for multiparous cows) and/or genetic potential, breed, additional treatments and diseases in the pp period, and pregnancy status, etc., if disease effect despite treatment is to be evaluated in the clinical trial set-up. The reason for considering these factors is that this part of the analysis is ‘observational of nature’ (e.g., randomization does not control for these factors between metritic and non-metritic groups). Another explanation of the counterintuitive finding could be that herds 2 and 3 were related because of their shared ownership and an on-going cooperation between the herds during the trial: movements of cows between these herds, based on their milk yield potential (e.g., high-yielding cows moved to herd 2 from herd 3) could have induced stress-related metritis in these cows, but still their milk yield could be higher than the ‘original herd 2 cows’. This hypothesis illustrates the complications that human interaction with data within a dairy management system potentially can introduce into interpretation of results (e.g., qualitative interaction (Ducrot et al., 1998)). By asking the herd-owner, this hypothesis was qualitatively rejected. But, the considerations show us the importance of knowing the origin of the data. Unfortunately, the sample sizes of data from an individual herd are insufficient for statistical evaluation of all the hypotheses presented here. However, work should be done to handle these potential biases in future trial designs also including the evaluation of disease effect despite treatment.
- In herd 3, the metritic cows tended to produce less milk than the non-metritic cows. Reasons for such a finding could include ineffective treatment protocols in case of a ‘true’ disease effect. Our findings of a milk loss of 4–5 kg ECM at 60 days pp adjusted for PAR and RP are somewhat larger

than the results from two out of ten observational studies showing 0.4 kg/day milk loss during the entire lactation and 2.3 kg/day up to 119 days, described in a review (Fourichon et al., 1999). The seemingly large milk loss in herd 3 could indicate that this herd has problems related to a treatment effect of the applied protocols because a smaller disease effect despite treatment apparently could be achieved in the other herds.

- In herds 1 and 4, no tendencies to differences in disease effect despite treatment of metritis in milk yield were evident. The most obvious reasons for such findings are effective treatment protocols (e.g., elimination of disease effect), no 'true' disease effects, or misclassification bias.

Another implication of this study relates to the proportion of cows diagnosed and treated for RP that required a subsequent metritis treatment (see Table 4); the statistically significant interaction among RP, PAR, and HERD; and the heterogeneity of the obtained estimates. These findings indicate a direct disease effect of RP on milk yield despite treatment, an effect that differed between herds and PAR within herd. These findings are in concordance with the results of Goshen et al. (2006), who discussed that RP and metritis are two separate disease entities and that they should be studied independently [6]. To test our findings of the importance of RP, clinical field trials with special focus on RP and treatment effectiveness are needed.

In general, our analysis gave us **no evidence of any consistent general or local difference in treatment effect** on milk yield (short or long term) of metritic cows treated with a penicillin protocol versus metritic cows treated with a tetracycline protocol (when adjusted for herd, RP, and PAR). These results might be anticipated because we conducted an active controlled trial with two treated groups, where the difference in milk yield between groups was expected to be smaller than the difference in milk yield in a negative controlled trial to detect 'treatment effect' (diseased untreated versus diseased treated cows). Also, the differences between the protocols were minor and related only to antibiotic group and not administration route or dosage. Similarly, another trial found no difference in treatment effect when comparing active treatments with different groups of antibiotics (penicillin IM, tetracycline IU, and ceftiofur IM) with respect to their effects on clinical cure and milk yield until 12 days post treatment (Smith, 1998). No long term effects on milk yield were evaluated by Smith (1998). The reason for our inconclusive findings can thus be as follows: 1) There is no true difference between our protocols, or 2) the true difference is too small to be detected with the accomplished sample size, level of adherence, and analytical method. For the practical implication of the results, one could argue that a non-inferiority or equivalence analysis of the data would have been more useful to the veterinarians because such an analysis can be interpreted in both directions. This analytical aspect could be reviewed in depth in the future but will require an additional data analysis (per protocol analysis and definition of effect margins) (Habicht, 2011). As a consequence of the statistically

non-significant findings, the two treatment groups could have been aggregated to give higher precision about the estimates of disease effect despite treatment, but we prefer to present the results according to the planned trial design.

We found that **both magnitude and direction of the disease effect of metritis despite treatment varied from herd to herd**. The non-systematic variation in direction of the effect estimates (e.g., inconsistency) is an important finding in these results, and the potential reasons are explained above. However, the inconsistencies emphasize that in practice, general advice and recommendations on the treatment protocol of metritis, RP, and perhaps production-related disease in general should be given with great caution. We recommend that local evidence of causal effects be used at all times. We also learned that clinical field trials could, with greater advantage, be performed purely at the herd level as opposed to practice level.

Conclusion

We have implemented a herd-specific, randomized and non-blinded multi-herd clinical field trial in a HHMP to evaluate the effectiveness of therapeutic interventions against metritis (diagnosis based on vaginal discharge). The trial was designed in close cooperation with and conducted by the veterinarians in practice in four private dairy herds. The trial aimed at 'supporting decisions to choose treatment protocol on herd or practice level'. We included 136 metritic and 744 non-metritic cows. We found no evidence of differences in effect of the applied metritis treatment protocols (penicillin >< tetracycline) on short- and long-term milk yield (predicted ECM at 60 days pp and total over 305 days lactation) when adjusting for herd, PAR, and RP. We found no evidence of difference in treatment effect between protocols. We found diverse disease effects on milk yield in the herds despite treatment. Consequently, the study demonstrates the importance of addressing the heterogeneity of treatment and disease influences when evaluating effectiveness. We found that clinical field trials and pragmatic principles for randomization and analysis can be implemented meaningfully in a dairy practice context. The study shows that data collection processes involving human interference are prone to error even in a highly structured HHMP. These data quality issues are relevant for all types of quantitative analyses.

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(10), 3174-3183.
- Box, G., Hunter, S., Hunter, W., 1978. *Statistics for Experimenters*. Wiley, USA. ISBN: 0471093157
- Dohoo, I.R., Martin, W., Stryhn, H., 2003. *Veterinary epidemiologic research*. AVC Inc., Charlottetown, Prince Edward Island, Canada. ISBN: 0 919013 41 4
- Dohoo, I., Martin, W., 1984. Disease, production and culling in Holstein-Friesian cows: IV. Effects of disease on production. *Preventive Veterinary Medicine* 2(6), 755-770.
- Ducrot, C., Calavas, D., Sabatier, P., Faye, B., 1998. Qualitative interaction between the observer and the observed in veterinary epidemiology. *Preventive Veterinary Medicine* 34(2-3), 107-113.
- Elkjær, K., 2012. Reproduction in the post partum dairy cow - influence of vaginal discharge and other possible riskfactors. Ph.D. Thesis. Science and Technology, Aarhus University, Denmark.
- Farrell, B., Kenyon, S., Shakur, H., 2010. Managing clinical trials. *Trials* 11(78).
- Fourichon, C., Seegers, H., Bareille, N., Beaudeau, F., 1999. Effects of disease on milk production in the dairy cow: a review. *Preventive Veterinary Medicine* 41(1), 1-35.
- Goshen, T., Shpigel, N.Y., 2006. Evaluation of intrauterine antibiotic treatment of clinical metritis and retained fetal membranes in dairy cows. *Theriogenology* 66(9), 2210-2218.
- Habicht, A., 2011. *Vurder selv evidens* (in Danish). Munksgaard, Denmark.
- Krogh, M.A., 2012. Management of data for herd health performance measurements in the dairy herd. Ph.D. Thesis. Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.
- Lastein, D.B., 2012. Herd-specific randomized trial - an approach for effect evaluation in a dairy herd health management program. Ph.D. Thesis. Faculty of Health and Medical Sciences, University of Copenhagen, Denmark
- Lastein, D.B., Enevoldsen, C., 2010. Visual assessment of within and between observers' agreement on vaginal discharge scores in a cattle practice context. *Proc. WBC XXVI in Santiago, Chile, 2010*.
- SAS Institute Inc, 2003. *SAS 9.2*. Cary, NC, USA.
- Schwabe, C., Riemann, H., Franti, C., 1977. Herd health programs. In: *Epidemiology in veterinary practice*. Lea & Febiger, Philadelphia, USA, 246-248. ISBN: 0-8121-0573-7
- Sheldon, I.M., Lewis, G.S., LeBlanc, S., Gilbert, R.O., 2006. Defining postpartum uterine disease in cattle. *Theriogenology* 65(8), 1516-1530.
- Smith, B.I., 1998. Comparison of various antibiotic treatments for cows diagnosed with toxic puerperal metritis. *J. Dairy Sci.* 81(6), 1555-1562.

Sørensen, JT., Enevoldsen, C., 1991. Effect of dry period length on milk production in subsequent lactation. *J. Dairy Sci.* 74(64), 1277-1283.

Thorpe, K.E., Zwarenstein, M., Oxman, A.D., Treweek, S., Furberg, C.D., Altman, D.G., Tunis, S., Bergel, E., Harvey, I., Magid, D.J., Chalkidou, K., 2009. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *Journal of Clinical Epidemiology* 62(5), 464-475.

Zwarenstein, M., Treweek, S., Gagnier, J.J., Altman, D.G., Tunis, S., Haynes, B., Oxman, A.D., Moher, D., 2009. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *BMJ* 337.

4 Discussion

The overall aim of the project described in this thesis was *to develop, implement, and conceptually validate randomized controlled trial designs that can be used by practicing cattle veterinarians for continuous development and evaluation of current and new diagnostic criteria and medical interventions in a dairy herd health management context.*

I have exemplified this overall aim with a case: evaluation of effectiveness of treatments for bovine metritis diagnosed as vaginal discharge in the early postpartum period. Medical treatment of metritis is one example of many possible interventions in a HHMP. I have synthesized the results of my studies in the tutorial section in this thesis and in cooperated some of them in a flowchart below (Figure 3). Details about the development, implementation, and conceptual validation of the components of the ‘trial cycle’ are in sections 3.2–3.6. The ‘trial cycle’ can be used for continuous formulation and re-formulation of herd-specific hypotheses in a highly structured HHMP based on a mixed methods approach. That is, the ‘trial cycle’ enables cattle veterinarians in private practices to evaluate the effect of interventions in a herd or practice context. Subsequently, the effect evaluation can be used to support a formalized and transparent ‘best decision making’ as an integrated part of ‘evidence-based veterinary practice’.

The flowchart (figure 3) illustrates some of the key aspects I will address in this discussion. The discussion follows the chart and describes some required actions for a veterinarian in practice (marked with italics) and the corresponding discussion.

A veterinarian encounters a problem in a herd that induces increased variation in the productivity of the cows within the herd (pre-trial period in flowchart). She must find evidence to support her next decision on a strategy (prevention or intervention) for reducing this variation (‘exceptional variation’) so that only the random variation remains. She wants to know if her choice of strategy is effective.

The hierarchy of evidence of effect [16] illustrates the view that a randomized trial approach is superior for evaluation and estimation of effects of interventions. The simple answer would be to use the randomized trial approach for effect evaluation in the herd.

However, the veterinarian must consider several issues before deciding whether to follow a trial approach to find and evaluate evidence of effect of interventions in her practical HHMP context.

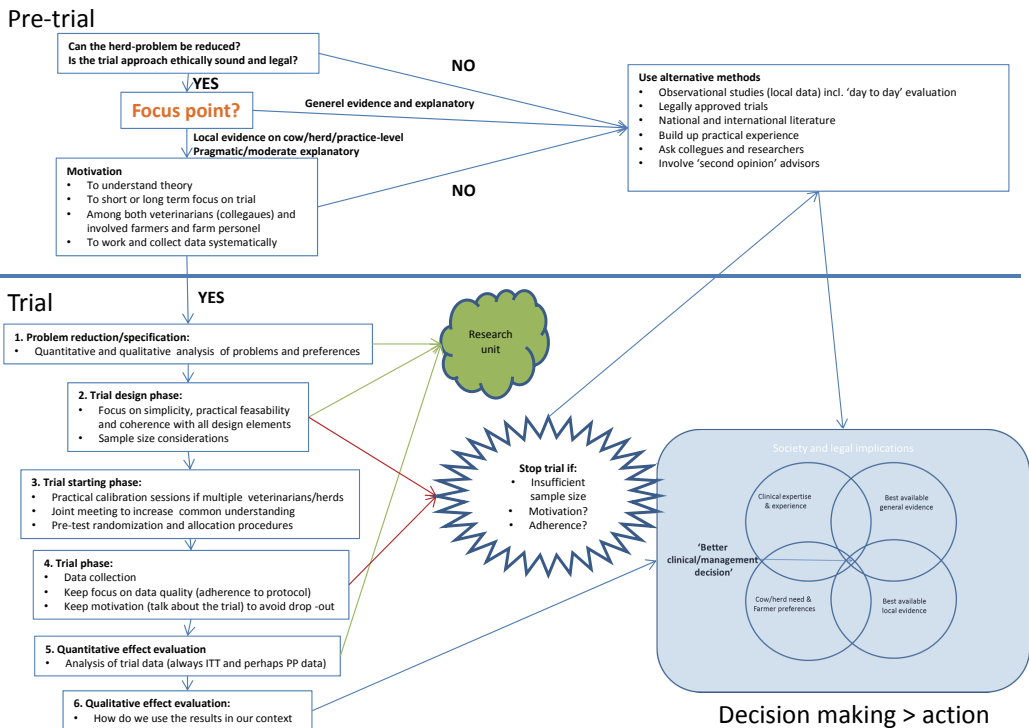


Figure 3. Flow chart to guide the discussion of this Ph.D. thesis on the implementation of randomized clinical field trials in dairy herd health management programs.

The complexity of a farm system with human involvement in decision making and data collection and with feedback mechanisms among animals, humans, land, and legal restraints, etc., makes the system difficult or maybe impossible to fit into the structured frames of trial theory (a 'wild problem') [12]. For the trial approach to be applicable, the evaluation task must be reduced to a 'tame problem' of relatively simple and/or technical character. Also the interventions tested must be ethically justifiable and legal without requirements for approval from the authorities.

I have illustrated how a 'wild problem' such as metritis treatment in a veterinary practice can be reduced through a 'bottom-up' approach by implementing retreatment and escape therapy and calibration of data collection as an integrated part of the trial design (manuscript V). The description of trial theory that takes ethical and practical clinical decision making into account (manuscript I and manuscript III) exemplifies numerous possibilities for adapting trial design to most practical contexts.

The veterinarian must also look into her/his own perception of the problem (or focus point) and decide whether general explanatory evidence of a biological association is the best evidence to solve the herd problem. If so, alternatives to the trial approach in HHMP should be considered. The obvious first choice would be a thorough review of the scientific literature.

The diversity of definitions of clinical or production-relevant bovine genital diseases and scarcity of medical treatments validated against a key performance indicator as for instance lactational milk yield in the literature (manuscript II), however, illustrate how difficult it could be for veterinarians in practice to obtain valid general evidence to support daily decisions in the specific herd-context. Whether the situation is similar for other clinical and management-related herd problems is beyond this discussion.

The trial approach could be suitable if the focus point is more pragmatic on issues regarding cow, herd, or practice level – and the veterinarian seeks support for practical decision making.

I found that veterinarians in practice differ very much, both in their practical actions and in their background for decision making even within the frame of a structured HHMP (manuscripts III and IV). These findings are in accordance with other research findings that veterinarians base their decisions on both experience and analysis [25]. My results also imply that a pragmatic approach to trial design [26,27] is more suitable for trials in the HHMP setting than explanatory trials, so the design can encounter more varying attitudes towards protocol and trial procedures and focus on endpoints relevant for the end-user (in the HHMP context, that is the farmer). I have found little evidence of profound acknowledgement of the pragmatic–explanatory continuum in the veterinary literature and veterinary epidemiological textbooks, although in conceptual terms, it also relates to internal and external validity. In human medicine, the concept seems to be more discussed and to some degree incorporated into trial theory and trial practice [28,29]. I am aware that the concept of a very pragmatic trial design can appear rather controversial to veterinarians and other academics educated in the natural science community and primarily focusing on ‘a positivistic approach’ to evidence (e.g., general evidence). The concepts of pragmatic approaches probably introduce a higher level of individuality than usual in veterinary science and are less stringent in definitions, etc., which is counterintuitive to the ‘positivist perception’ of valid evidence. It is important to note the discussion for and against pragmatic trials. Kent and Kitsios (2009) argue, that pragmatic trial design cannot claim more accurate measures of a ‘true treatment effect’ than an explanatory trial design. This is because this ‘true treatment effect’ is illusionary, as the effect will always vary with the context. But, Kent and Kitsios (2009) also argue that an efficacy trial giving no evidence of treatment effect might be more valuable than a pragmatic trial (e.g., indicate no effect under the ‘most favourable conditions’ compared to ‘no effect under real-world conditions’) [28].

The veterinarian must keep looking inward to explore the motivation for the trial approach. The motivation consists of the combined relationship among understanding of trial theory, epidemiological and statistical principles, ability to keep short- or long-term focus on the trial and to work and collect data systematically, and last but not least, the perception of the motivation among colleagues in practice and the involved farmers and farm personnel. The trial approach should not be pursued unless motivation is high. Alternative methods for effect evaluation (lower in the evidence hierarchy than randomized trials) should be used to reach a higher level of certainty in the decision making.

In performing the work during my study period, I have confirmed the findings of motivation factors—clinical, advising, law-abiding, and epidemiological (manuscript IV)—through additional information about the perceptions of motivation to evaluate gathered among different groups of veterinarians (focus groups, meetings, informal dialogue, etc.). I find that factors that promote motivation (incentives) for evaluation are both personal and voluntary (e.g., academic doubt, the joy of ‘proving’, ethics, and understanding of epidemiological principles) and non-personal and involuntary (e.g., legislation, penalties). De-motivating factors such as a lack of understanding and time and complicated access to data were also identified among this extended group of Danish veterinarians. An important conclusion from this study is that any veterinarian in charge of a field trial must understand and accept the difference between aims and procedures of both pragmatic and explanatory trials and the continuum in between. I foresee that if the differences are not appreciated, the development and implementation of any specific trial design will become inconsistent and the trial team (veterinarians and farmers) will become de-motivated, uncertain of the results, and more hesitant to initiate another ‘trial cycle’. Farrell and co-workers suggest that personal interest in evidence-based practice, academic relations, plenty of staff (support and help), and understanding of basic principles (defined as being comfortable when explaining trial concepts) will encourage clinicians in the recruitment phase [30]. These requirements to facilitate implementation of randomized controlled trials in the HHMP could be met if larger veterinary practices had a professional clinical research unit to support the planning and analytical phases.

The limitations of the trial approach identified in this study have led me to stress the importance of meeting the veterinarians and farmers on whatever ‘level in the hierarchy of evidence’ they are at a given point in time. ‘Best evidence’ could be changed into ‘better evidence’, indicating that moving one step up in the hierarchy is better than staying on the current level. For instance, a veterinarian who works in a HHMP relying entirely on experience-based evidence could consult colleagues and ‘second opinion’ advisors to seek ‘better evidence’ and not initiate clinical field trials. Alternatively, a veterinarian who already uses locally collected data for retrospective analysis of observational data could initiate trials to seek ‘better

evidence'. In this way, the entire veterinary profession would reach higher levels of evidence for their actions. It could be argued that this is how 'things already work'. However, I have experienced that veterinarians in all ages and educational and experience levels have asked where to find their best evidence. This leads me to conclude that efforts should be made to make the hierarchy of evidence and the paths to obtaining evidence more explicitly known to the veterinarian in practice.

If the veterinarian decides to follow the trial approach (Trial period in flowchart), she could read the tutorial (manuscript I) and start following the six phases. During the whole process, she could need assistance from a 'research unit'. Initially, problem reduction and specification of the trial hypothesis are planned. Both quantitative and qualitative enquiries are required at this stage. She should investigate local data related to the herd problem by means of the available tools in the Danish HHMP to specify and reduce the problem.

A quantitative observational approach to HHMP data is presented in detail by Krogh (2012) [15]. Krogh describes different approaches to data analysis in a HHMP where a continuous evaluation of time series of measurements of key performance indicators is a critical component (e.g., example 1 in the Tutorial, manuscript I). The time-series analysis should separate **exceptional variation** from the random variation in the time sequence of the performance measurements from the production process. If a subsequent search for causes of this *exceptional variation* (e.g., disease incidents or management failure) is successful and the causes can be removed (either by prevention or intervention), the remaining variation in the production process should be the **random variation**. A process that produces only random variation (often called noise) is predictable within limits determined by the degree of random variation, and well-defined predictability (basically a prediction model) is essential for planning in herd management. Reduction of noise will improve precision of the predictions, which will improve the effectiveness of management. In the HHMP context, the aim is to reduce both the exceptional and random variation by implementing effective management or intervention strategies on the herd level. The trial approach can thus be helpful to evaluate whether a new strategy is more or less effective than the current strategy.

During the planning phase, the veterinarian should also ask herself (and her colleagues) about diagnostic tests and data collection routines and make every detail explicit. Farmers should be asked for their preferences. Are there work routines that are unacceptable? Practical details must be planned and kept simple. During the trial conduct, expert support from a research and development unit could also facilitate improved common understanding ('communication') on any aspect within the trial concept. During the process, it is important to involve all parts of the project, including the farmer and his personnel. This will encourage everyone to stay motivated, to adhere to the protocol, and to successfully run the trial.

Leblanc et al. [14] suggested turning “dairy herd health management or production medicine into an integrated, holistic, proactive, data-based, and economically framed approach to prevention and enhancement of performance”. This statement seems to represent a rather quantitative and rational view. My qualitative studies among farmers and veterinarians demonstrate the importance of establishing real communication (that is, to obtain a common understanding of a topic) with the people involved at all levels of farm systems management. We, as herd health management consultants, must learn to ask (and listen to!) the farmers to construct systems that can produce ‘tailor-made or herd-specific answers’ that suit personal preferences and not just economic theories. Preferences and perceptions of veterinarians and farmers about HHMPs have been shown to differ [31].

During the trial planning, starting, and conduct phases, the veterinarian must focus on some criteria for ‘success’, defined as the possibility of obtaining a reliable result that is still meaningful to the end-users at the end of the trial. Trials in a HHMP should be stopped if the chance of success is considered unacceptable by the veterinarians and farmers. Perhaps the decision to initiate the trial process was wrong? Perhaps sufficient sample size cannot be accomplished within a reasonable timeframe? Perhaps motivation and meaning are lost along the way (e.g., change in legislation)? Perhaps non-adherence is identified, and the reasons and levels for the non-adherence cannot be explained or justified? Consider stopping the trial and look for alternative evidence to solve the herd problem.

If it is likely that internal validity of the results of a trial cannot be obtained to an acceptable degree within a certain veterinarian/practice/herd context (e.g., very low sample sizes, low power, high level of non-adherence, and non-random allocation and drop-out), then the inferences drawn from the trial might not be better than the individual perception of effect already present. In such a situation, the trial might as well be stopped. As an example, I can justify decisions to continue the trial conducted in this project (see section on Study context) although I found a ‘high level of potential non-adherence to the protocol’. For the 11 herds, the ITT dataset included 206 cows. When the dataset was reduced to a PP dataset, only 129 cows were included, corresponding to a 40% reduction. However, I explored the raw data thoroughly and found that this reduction primarily was due to cows that were not examined and treated on the regular herd visit. Thus, I did not have the chance to verify whether the inclusion criteria were ‘correct’. However, a lack of ability to verify does not necessarily imply that the inclusion criteria were not followed. In a pragmatic trial setting, such verification of inclusion criteria and potential exclusion of the observations would be considered non-coherent with the design [26]. In this case, I was able to justify the inclusion of the cows because raw data were available. In the HHMP setting with good data recording in general, such qualitative validation of data is useful to document the data collection process. Consequently, I accepted the ‘low’

level of adherence in the ITT dataset and potential variation in inclusion criteria and proceeded with a pragmatic trial conduct and analysis of the ITT dataset. Also, the recording errors in the database (as described in manuscript V) were evaluated by asking the farmers (again) about their recording routines, and these findings indicated that 'errors in recording' was a side effect of the legal enforcement, not a side effect of the trial situation (e.g., the cows were treated according to protocol, but recording was according to minimum requirements of legislation). If the opposite had been the case, the farmers should have agreed to correct the procedures or the trials should have been stopped. An important side effect of my work with these data is that I have demonstrated a series of problems with data collection in the field, even in a highly structured HHMP. The logical conclusion is that data quality due to human qualitative interactions must be worse in files collected from herds without such structured programs.

In the analytical phase, some problems with data analysis could arise for the veterinarian in practice. Depending on her personal skill with quantitative analysis, help from the 'research unit' could be warranted. However, if the problem is reduced to a minimum of complexity (within herd, two parallel groups), one continuous or dichotomous results measurement is chosen, and randomization/allocation has been performed well (balanced numbers and known prognostic factors), then simple comparisons of averages (t-tests) or proportions (Chi-square) could be a pragmatic and sufficient choice of analytical principle.

Unfortunately, even reduced problems can be more complex beneath the surface. To illustrate my point, let me revisit the analysis of randomized clinical field trials presented in this thesis to estimate effects of medical treatment of metritis on milk yield. In our ANOVA model, the aim was to compare milk yield averages between intervention groups treated for metritis and non-metritic cows adjusting for the level of other prognostic factors. However, the ANOVA requires that the interval scale measurements can be modelled under the assumptions of a normal distribution. Yet we found some violation of assumptions of variance homogeneity (under the assumption of a normal distribution of the residuals) during the model check of the ANOVA model (model 2 in manuscript V). That is, we experienced some *exceptional variation* in milk yield (especially 305-day milk yield). In an analysis of a trial in a traditional (and explanatory) research setting, we probably would have considered the following remedies to obtain variance homogeneity across the combinations of explanatory variables in the ANOVA model:

Transform the response variable (e.g., square root, $1/Y$, or logarithmic depending on the distribution; e.g., depending on the 'fanning patterns' of residuals). A major disadvantage of this approach seen in our pragmatic and practical context is that the interpretation of the effect estimates and their potential multiplicative associations may become more complicated [32].

Remove certain extreme individual measurements or categories of measurements: in statistical jargon, often called outliers (either via a more narrow sampling frame or post-modelling exclusion). This process would have excluded the 'exceptional variation' leaving only 'random' variation to model.

With these tools, we could make the study population fit to some convenient statistical theory. However, more and more radical transformation of variables and restrictions on data (inclusion criteria and exclusion criteria) will detach the study population (the metric cows) more and more from the reference population (the herd) and, consequently, from the general population (Danish dairy cows). That is, we reduce both internal (and external validity, if generalizability was the goal). I acknowledge that the estimates derived from a model that do not meet the theoretical assumptions should be interpreted with caution. However, in this HHMP context, I chose to neglect variance heterogeneity, outliers, or other violations of theoretical constraints to be better in line with the concept of pragmatic trial theory and the evaluation of effectiveness. My aim was to obtain maximal internal validity for all cows within a herd or practice context. However, given that my results are internally valid, I could argue that the results of a pragmatic trial could be considered more externally valid than an explanatory trial because of the low degree of constraints on included cows, as discussed in detail in human medicine [29]. In line of this argument, the results could be of some evidential value beyond the study population to support decisions by other veterinarians in practice that are not motivated for conducting their own trials. However, this would only be appropriate in trials that show evidence of a treatment effect, if the argumentation of Kent and Kitsios (2009) above should be followed [28]. So to be fully consistent also with the herd-specific aim and pragmatic design, I can only recommend using the results within the herd and practice contexts they are developed in.

In the end, a decision must be taken and put into action in the real world HHMP, perhaps based on the result of the trial. Results can be applied in the context of the herd or veterinary practice in accordance with the principles of 'evidence-based veterinary medicine/practice'. Local evidence obtained through the trial approach will never stand alone as support for veterinarians in practice. Veterinarians must rely on all of their contextual knowledge of the problem to take 'the better clinical or management decision'.

5 Conclusions

The following overall conclusion will describe the possibilities for integration of herd-specific randomized trials in a dairy herd health management program as the initial phases of a concept development. The individual subprojects in this thesis aimed at contributing to this overall conclusion. The conclusion are divided into two parts, as follows: (1) methodological and conceptual ideas of trials in the HHMP mainly supported by literature and personal experience and insight gathered during the entire project, and (2) practical issues related to planning and implementation of trials supported by concrete findings and problems during the conduct and analysis of both the empiric material and the trial data.

Conclusions related to overall methodological and conceptual issues are as follows:

- Herd-specific clinical field trials can be implemented together with retrospective performance measurements, general scientific evidence, and clinical experience and personal preferences in a specific context as part of an 'evidence-based veterinary practice' with the purpose of finding meaningful scientific evidence of effects of interventions in dairy HHMPs. An example is the development and implementation of the trials in manuscript V in the search for answers to research questions regarding applied procedures that are difficult to justify via the existing literature (manuscript II and III). Special prerequisites related to the ethical position of the herd problem, the focus point (pragmatic at cow, herd, or practice level), and the motivation of the individual veterinarians/farmers to work systematically is clearly not always met. The trial approach should be considered as a possibility only for some herd problems, some veterinarians, some practices, some farmers, and some herds. It would be an advantage if a structured HHMP were in place before introduction of the concept.
- A mixed methods approach, involving both quantitative and qualitative research methods as used in this thesis, is useful for developing and evaluating trial designs that can suit many kinds of herd problems in the dairy HHMP context. The reason for this usefulness is that both measurable entities and human perceptions can be integrated into the same context.
- Increased knowledge of basic ontological principles of scientific research in general (e.g., emphasis on the theory of science) will improve a trial team's (researcher, veterinarians, farmers) understanding of any contexts that involve humans, such as farming systems. In this case, the data collection processes such as clinical examination and scoring in the HHMP are affected by the 'life-worlds' of each data collector (manuscript IV).

- Acknowledgement and acceptance of context-specific action patterns and motivation factors, as illustrated in manuscript III and IV, illustrate the need for a ‘tailor-made’ herd-specific clinical field trial design. One design does not fit all! Lack of motivation and understanding by the trial team to pursue the aim of the trial could lead to poor data quality (reduced precision in diagnostic measurements, and non-adherence and bias).
- A minimum understanding of basic trial theory is mandatory to appreciate and implement the trial approach. The work presented in the tutorial (manuscript I) can contribute to such understanding. Any veterinarian who initiates trials and interprets the results in a HHMP must be familiar with the following issues and their consequences:
 - 1) Randomization (that is equal chance to be allocated to either intervention group) and blinding to control bias
 - 2) The difference between an efficacy and effectiveness trial (and the continuum in between); that is, the continuum of trial designs between ‘biological effect of an intervention’ (**explanatory** trials and per protocol analysis) and ‘effect of a policy/strategy or decision’ (**pragmatic** trial and intention to treat analysis)
 - 3) Adherence to protocol and the related ‘dilution effect’ of the results

Conclusions about practical issues related to planning and implementation of trials:

- I propose three **conceptual** types of trial designs based on the findings of a wide range of applied actions patterns and rationales in Danish veterinarians work (manuscript III):
 - 1) Moderately explanatory within-herd clinical field trial with clinical focus
 - 2) Pragmatic within-herd clinical field trial with production focus
 - 3) Pragmatic multi-herd clinical field trial with production focus
- I propose the following overall framework for **practical** trial designs:

Study population: within-herd/within practice

Non-blinding and simple allocation procedures that fit into normal work procedures Control group: active or negative control

Design: parallel group, modified cross-over, or factorial design

Possibility of stopping the trial early: escape-therapy or group-sequential design
- As an example of the trial approach, a pragmatic clinical field trial design with production focus was practically feasible in 10 private non-compensated dairy herds within two veterinary practices, as presented in manuscript V. The trials were conducted as non-blinded, ear tag–allocated, active controlled 2-parallel group designs with four different protocols comparing ‘current’ antibiotic

treatment with 'new' antibiotic treatment of metritis for evaluating effectiveness on milk production (two within-herd trials, two multi-herd trials). A method for analysis of differences in treatment effect in estimated peak milk yield at 60 days in milk (ECM60) and estimated 305-day total yield (ECMtotal305) between the intervention groups is demonstrated (ANOVA). Problems related to model assumptions were identified. No evidence for treatment effect is found. Heterogeneity of disease effect despite treatment between groups of cows with different prognostic factors (herd, parity, retained placenta) is a finding of major importance for future trial design and interpretation of the scientific literature.

- Several issues related to statistical implications was revealed; 1) Low power of analysis can be a problem with small herd sizes and low disease incidences in some herd(s) despite long trial periods, 2) Non-adherence is frequent even in the relatively controlled settings such as the Danish HHMP and among motivated veterinarians, 3) The use of negative control groups and validation of treatment criteria are possible under Danish legislation if escape therapy for systemically ill cows is implemented. 4) A relatively simple principle for analysis of trial data can be used if the sources of bias are reduced; however, statistical support could be warranted for some veterinarians in practice who should analyse data from clinical field trials.

The result of an effect evaluation in a herd-specific trial can potentially give support to different changes in practice routines and potentially guide 'the veterinarian decisions towards a more evidence-based practice' within the herd-context (specific herd problems). My final remarks concerning the prerequisites for successful implementation of randomized clinical field trials in the Danish HHMP context can be condensed into the following keywords:

Veterinarians and farmers should ensure context-specific understanding, acceptance, and motivation.

Cows should be explicitly included in the trial in the highest possible numbers.

Data should be collected and analysed, and the results applied in their specific context.

6 Perspectives

This section focuses on topics that warrant further research and development. The efforts could be two-fold: 1) to improve interventions for uterine diseases and 2) to provide better understanding and evaluations of the potential for implementing randomized trials in the HHMP context.

Future aspects of interventions for uterine disease in the Danish HHMP:

- Validation of the herd-specific diagnostic criteria for early postpartum metritis on both milk yield and reproduction could be performed in the proposed trial design if negative control groups are included.
- An evaluation of effects of non-antibiotic interventions is very relevant and might be possible and applicable in a setting with organic rules aiming specifically at reduced use of medication. In this context, the trial should provide estimates of the effect of the difference between use and non-use of antibiotics or between use of antibiotics and alternative interventions.

Future prospects for randomized trials in the HHMP context:

- Identification of other clinical entities or management procedures to be tested in the proposed trial design is a natural extension of the work in this thesis. Specific recordings and their context-based data collection procedures should be described in detail. Mastitis is an obvious possibility but could carry even more complicated diagnostic challenges than the described 'metritis case'. Management decisions (such as dry cow management) are also obvious possibilities.
- Automation of trial support procedures including analysis of group sequential testing and by means of some data management platform might solve some of the adherence problems identified in this project.
- An exploration of the character and extent of data manipulation as a consequence of the legislation of the Danish HHMP could reveal new insights into the nature of qualitative interaction between humans and restraints like legislation in the creation of data in national databases.
- In depth-identification and prioritization of barriers against and motivation towards introducing the trial approach into the HHMP on either a voluntary or mandatory basis would be obvious as the first step in a further development of the herd-specific randomized trial approach.

7 References – Thesis

1. Schwabe C, Riemann H, Franti C (1977): Herd health programs. In: Epidemiology in veterinary practice. Lea & Febiger Philadelphia, USA: 246-248. ISBN: 0-8121-0573-7
2. Enevoldsen, C.(1993): Dairy Herd Health Management [In Danish with english summary]. Ph.D. Thesis . The Royal Veterinary and Agricultural University, Copenhagen, Denmark.
3. Nir Markusfeld, O (2003): What are production diseases, and how do we manage them? Acta Veterinaria Scandinavica , 44 (Suppl. 1): 21-32.
4. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS (1996): Evidence based medicine: what it is and what it isn't. BMJ, 312(7023):71-72.
5. Schmidt PL (2007): Evidence-based veterinary medicine: evolution, revolution, or repackaging of veterinary practice? Veterinary Clinics of North America: Small Animal Practice, 37(3):409-417.
6. Vaarst M, Paarup-Laursen B, Houe H, Fossing C, Andersen HJ (2002): Farmers' choice of medical treatment of mastitis in Danish dairy herds based on qualitative research interviews. J Dairy Sci, 85 (4):992-1001.
7. Andersen HJ (2004): Rådgivning, bevægelse mellem data og dialog [In Danish]. PhD Thesis. Mejeriforeningen, Århus. ISBN: 87-89795-81-4
8. Hegelund A (2004): Veterinary paradigms and practices. Ph.D. Thesis. The Royal Veterinary and Agricultural University, Copenhagen.
9. Aagaard-Hansen J (2007): The challenges of cross-disciplinary research. Soc Epi, 21(4):425-438.
10. Kristensen E, Bay Nielsen D, Vaarst M, Enevoldsen C (2008): A mixed methods inquiry into the validity of data. Acta Vet Scand, 50:30.
11. Åkerlind GS (2005): Variation and communality in phenomenographic research methods. High Edu Res Devel 2005, 24(4):321-334.
12. Krogstrup H (2011): Kampen om evidens; Resultatmåling, effekt-evaluering og evidens [in Danish]. Hans Reitzels Forlag. Copenhagen, Denmark. ISBN: 9788741255163
13. Bekendtgørelse om sundhedsrådgivningsaftaler for kvægbesætninger [Law in Danish] (2010) and related changes, [<https://www.retsinformation.dk/Forms/R0710.aspx?id=132648>] Fødevarerministeriet, Copenhagen, Denmark. Assessed 19-6-2012.
14. LeBlanc SJ, Lissemore KD, Kelton DF, Duffield TF, Leslie KE (2006): Major advances in disease prevention in dairy cattle. J Dairy Sci, 89(4):1267-1279.
15. Krogh, MA (2012): Management of data for herd health performance measurements in the dairy herd. Ph.D.Thesis. Faculty of Health and Medical Sciences, University of Copenhagen, Denmark. ISBN: 978-87.7611-505-0
16. Habicht A(2011): Vurder selv evidens (in Danish). Munksgaard, Denmark. ISBN: 978 87 628 1111 9

17. Kristensen EL (2008): Valuation of dairy herd health management. Ph.D. Thesis. Faculty of Life Sciences, University of Copenhagen, Denmark.
18. Enevoldsen C (2006): Epidemiological tool for herd diagnosis. Proc.WBC XXIV abstracts, Nice , 376-383.
19. Sargent RG (1982): Verification and validation of simulation models. In: Progress in modelling and simulation. Edited by Edited by Cellier FE. London: Academic, 159-169.
20. Kvale S (1994): Interview - en introduktion til det kvalitative forskningsinterview. [in Danish]. Copenhagen, Denmark: Hans Reitzels Forlag. ISBN: 87-412-2816-2
21. Barnard A, McCosker H, Gerber R (1999): Phenomonography: A qualitative research approach for exploring understanding in health care. Qual heal res, 9:212-226.
22. Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, Tunis S, Bergel E, Harvey I, Magid DJ et al. (2009): A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. CMAJ, 180:47-57.
23. Lastein DB, Enevoldsen C (2010): Visual assessment of within and between observers' agreement on vaginal discharge scores in a cattle practice context. In: XXVI World Buiatrics Congress 2010: WBC 2010 Congress Abstracts. Santiago, Chile
24. Schulz K, Altman D, Moher D (2010): CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. Trials, 11(32).
25. Mee JF (2010): Treatment of fertility problem cows: what do veterinary practioners actually do? In XXVI World Buiatrics Congress 2010: WBC 2010 Congress Abstracts. Santiago, Chile
26. Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, Tunis S, Bergel E, Harvey I, Magid DJ et al. (2009): A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. Journal of Clinical Epidemiology, 62(5):464-475.
27. Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, Oxman AD, Moher D (2008): Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ, 337, a2390.
28. Kent D, Kitsios G (2009): Against pragmatism: on efficacy, effectiveness and the real world. Trials, 10(1):48.
29. Treweek S, Zwarenstein M (2009): Making trials matter: pragmatic and explanatory trials and the problem of applicability. Trials, 10(1):37.
30. Farrell B, Kenyon S, Shakur H (2010): Managing clinical trials. Trials, 11(78).
31. Kristensen E, Enevoldsen C (2008): A mixed methods inquiry: How dairy farmers perceive the value(s) of their involvement in an intensive dairy herd health management program. Acta Veterinaria Scandinavica 2008, 50(1):50.
32. Dohoo IR, Martin W, Stryhn H (2003): Veterinary epidemiologic research. AVC Inc., Charlottetown, Prince Edward Island, Canada. ISBN: 0 091013 41 4

8 Appendix

Appendix A – Description of vaginal discharge score

Appendix B – Method for calibration of vaginal discharge score

Appendix C – Trial protocols

Appendix A

Vaginal discharge score (VDS) 5-21 days post-partum is used as part of inclusion criteria in the clinical field trials evaluating treatment effect of metritis.

Modified from Danish version of scale (lr.dk, 2012 - lr.dk, 2012. Vejledning til kliniske registreringer [in Danish].)

VDS	Odour	Discharge description (volume, contents/colour)
0	No	No or minimal volume of clean mucous discharge
1	No	Minimal volume of bloody mucous discharge
2	No	Small volume of bloody mucous discharge
3	No	Considerable volume of bloody sero-mucous or mucopurulent discharge
4	No	Considerable volume of mucopurulent discharge (yellow)
5	Abnormal	Minimal to plenty amounts of purulent discharge (yellow)
6	Abnormal	Considerable volume of purulent discharge (yellow)
7	Fetid	Considerable volume of purulent/haemorrhagic discharge (yellow, red, brown)
8	Fetid	Plenty volume of watery haemorrhagic discharge (red, brown, grey)
9	Fetid	Large amounts of watery discharge and debris (red, black)

Appendix B

Calibration of veterinarians working with a clinical field trial for evaluating effectiveness of metritis treatments in a Danish dairy herd health management program. Presented as an abstract (and poster) at XXVI World Buiatrics Congress 2010: WBC 2010 Congress Abstracts. Santiago, Chile

DEPARTMENT OF LARGE ANIMAL SCIENCES
FACULTY OF LIFE SCIENCES
UNIVERSITY OF COPENHAGEN



Visual assessment of observer agreement on vaginal discharge scores in a cattle practice context

Dorte Bay Lastein, DVM, Ph.D. student
Carsten Enevoldsen, DVM, Ph.D., DipECBHM
University of Copenhagen, Denmark
Faculty of Life Sciences
Grønnegårdsvej 2 DK-1870 Frederiksberg C
Contact: bay@life.ku.dk

Introduction

Should cows be treated for uterine disorders early post partum? Which clinical diagnostic indicators should be used and are valid data available?

Can vaginal discharge withdrawn from vaginas with the gloved hand and visually inspected until 21 days post partum (pp) be a precise method for diagnosis of puerperal and clinical metritis in veterinary practice as defined by Sheldon et al 2006?

In Denmark a Herd Health Management (HHM) program with systematic screening of all fresh cows is implemented in some herds (www.liv.dk). An ordinal vaginal discharge score (VDS) is given to all cows in these herds. VDS is measured on an ordinal scale from one to nine (table 1) and registered in the national cattle data base.

Multiple studies use scoring of vaginal discharge to categorize puerperal uterine status. If within and between observer agreement is assessed, it is often done by means of kappa values. However this method does not give an intuitively simple overview of structural agreement within a practice – who agrees with who? If additionally quality assessment of observations should be used in clinical practice in the future practice development such alternative methods might be useful.

In this study we have combined a traditional description of within and between observers agreement on VDS (weighted kappa, as described by Watson and Petrie, 2010) in a practical context with the use of an alternative method for analyzing and visualizing the agreement within a group of veterinarians – PCA analysis.

The objectives of this study were 1) to test a practice-based setup to assess agreement on vaginal discharge scores (VDS) and 2) to demonstrate a visual evaluation method of the agreement within and between observers (PCA PLTOS).



Figure 1
Various types of vaginal discharge in 'the gloved hand'.

Material & Methods

Seven veterinarians participated; six veterinarians working within same practice and the first author. A 'scoring chart' with definitions was presented before the study (table 1). The veterinarians had varying degree of experience with examining cows post partum, but had all used the same scoring chart before. The veterinarians were informed not to comment cows and discharge during the examinations.

The vaginas of 21 cows on days 5-21 pp were explored twice by the same veterinarian who refracted uterine discharge. The examining veterinarian showed the hand to the other participants so that volume, consistency and odour could be evaluated (figure 1). The cows were mixed and rearranged in the headlocks before the second examination to ensure blinding. The veterinarians recorded their scores on separate sheets for each scoring. Each cow had a number written on the thigh for identification.

Table 1
'Scoring chart' showing the 10 point ordinal scale of vaginal discharge scores (VDS) used in a Danish Herd Health Management program for evaluating uterine status 5-21 dpp.

VDS	Smell	Discharge description (volume, consistency/colour)
0	No	No or minimal volumes of transparent mucoid discharge
1	No	Mainimal volumes of homogenous and/or mucoid discharge
2	No	Small volumes of homogenous and/or mucoid discharge
3	No	Consistible volumes of homogenous, mucoid or mucopurulent discharge
4	No	Consistible volumes of mucopurulent discharge
5	Abnormal	Mainimal to small volumes of purulent discharge (yellow)
6	Abnormal	Consistible volumes of purulent discharge (yellow)
7	Foal	Consistible volumes of purulent/mucopurulent discharge (yellow, red, brown)
8	Foal	Large volumes of watery homogenous discharge (red, brown, grey)
9	Foal	Large amounts of watery discharge and/or debris (red, black)

The data were arranged with cows as columns and veterinarians (objects) as rows. Columns were mean centred to reduce influence of cows' variation in VDS, hence only variation due to observer variation is considered. Principal component analysis (PCA) (e.g. score and loading plots) was conducted with Analytica[®].

Weighted kappa values within and between veterinarians were estimated for comparison (SAS proc freq).

PCA Results

The first 4 components in the PCA model explained 34%, 22%, 14% and 8% of the variation in data, respectively. Figure 2 demonstrates that the 1st component, measures variation due to within observer agreement. The first scoring is represented by red squares and the second scoring by grey squares. It is evident that red and grey are separated in first PCA component.

The equal distances between the data points representing each pair of observations (or veterinarian) could indicate that clinical experience does not seem affect the within observer agreement.

The 2nd PCA component measures the variation due to veterinarians because the veterinarians (represented by different capital letters) are spread along the vertical axis.

The PC1/PC2 plot also demonstrates groupings of veterinarians with high levels of agreement. For instance, veterinarians L/T and A/Ka seem to agree on both first and second scoring.

The loading plot (figure 3) shows how 3 cows (#6, #15, #19) contribute with the most variation.

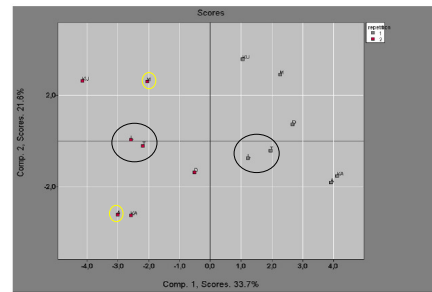


Figure 2: PCA Scoreplot

Scoreplots of PC1/PC2 showing the distribution of 7 veterinarians' ability to score vaginal discharge from 21 cows twice. Capital letters identify each veterinarian. Red indicates first observation and grey second observation. More than 50% of the variation in data are explained by the two primary PCA components. Large black circles indicate a pair of veterinarians that agree on scores in both first and second examination and form a structural relation. Small yellow circles indicate selected veterinarians in a non-structural relation with L.

Kappa for comparison

The overall weighted kappa within veterinarian (1st to 2nd observation) is 0.57 [0.49-0.65] and the overall weighted kappa between veterinarians (1st observation) is 0.64 [0.62-0.67]. This resembles the PCA results in that within veterinarian variation varies more than between veterinarians.

The weighted kappa between selected veterinarians illustrates the differences in kappa values between structural and non-structural groups as identified by PCA score plot (figure 2).

A structural group is formed by L and T (wkappa=0.75 [0.60-0.87]). The distance between L and A (wkappa =0.67 [0.54-0.80]) in the score plot is larger than the distance L-T, which is also reflected in the kappa values. A non-structural relation exists between L and M. Hence, the kappa values are similarly lower (0.55 [0.43-0.67]).

Discussion

The study was practically feasible in a 300 cow dairy herd and could be conducted in 1½ hours on a single farm, hence applicable for calibration purposes in a veterinary cattle practice. Setups with more than 2 subsequent vaginal examinations were dismissed for ethical reasons. Larger studies with more cows with better representation of high scores could be useful. Further, the related 'latent class models' of VDS scores could be used to identify a latent true status of uterine affection, hence addressing accuracy of VDS.

This study shows that within observer disagreement in this setup are substantial. This could be due to both differences in volume of the vaginal discharge between first and second examination (e.g. poor study design) or a poor ability to score uniformly. We conclude that the present setup is most suitable for assessing between observer agreement. More time between examinations are necessary to provide better insight into within veterinarian agreement.

The setup and the PCA analysis demonstrate a potential to monitor and calibrate observations and data quality in practice and research. The visual assessment of structural groups can be used to form practice teams of agreeing veterinarians to improved data quality on herd level. The methods could be used in other areas, for instance body condition or lameness scoring. Increased focus on observer agreement and the potential improved validity of within and between herd analysis of health and production problems could become a integrated part of quality assessment in modern dairy practice.

However, the PCA method does not quantify the agreement, merely illustrates potential groups of agreeing observers by showing relative variation. Using a combination of PCA, kappa or other measures of agreement could reduce this disadvantage.

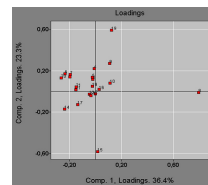


Figure 3: PCA loading plot

The loading plot shows the 21 cows examined in the study of vaginal discharge as data points. The variation within observations for each cow (two subsequent observations by 7 veterinarians) indicates that some cows contribute with large variation (positioned a long distance from the other cows).

References

- Sheldon IM, Lewis GS, LeBlanc S, Gilbert RO: Defining postpartum uterine disease in cattle. *Theriogenology* 2006, 66:1516-1530.
- www.lanrbugetinfo.dk/Kvaag/Filer/nyr_vet_klinik.pdf (in Danish)
- Watson, P.F. and Petrie, A. Method agreement analysis: A review of correct methodology *Theriogenology*, 2010, 73: 1167-1179

Appendix C

Treatment protocol for 10 trials (herd 10 dropped out). Cows in each herd were ear-tag allocated into treatment groups T1 and Tx2. Intramuscular injection=IM. Intrauterine pessaries=IU

Practice A	Protocol	Administration route/medication	Dosage	Duration of treatment
Practice A	A Herd 1-4 (inclusion criteria described in manuscript V)	Tx1: IM: Benzylpenicillin procaine*	50-60 ml (300.000 IE/ml)	3 days
		IU: Dihydrostreptomycin sulfat/Benzylpenicillinprokaine/Sulfadimidin pessaries**	3 x (500 mg/ 250.000 IE/4g)	1st day
		Tx2: IM: Oxytetracyclin-hydrochlorid /hydrat***	50-60 ml	3 days
		IU: Oxytetracyclin pessary****	1 x 500 mg	1st day
	B Herd 4-8 (same inclusion criteria as herd 1-4)	Tx1: IM: Benzylpenicillin procaine*	50-60 ml (300.000 IE/ml)	3 days
		IU: Dihydrostreptomycin sulfat/Benzylpenicillinprokaine/Sulfadimidin pessaries**	3 x (500 mg/ 250.000 IE/4 g)	1st day
	Tx2: IM: Sulfadiazin + trimethoprim (240 mg/ml)*****	50-60 ml	3 days	
	C Herd 9 (same inclusion criteria as herd 1-4)	Tx1: IM: Benzylpenicillin procaine*	50-60 ml (300.000 IE/ml)	1-2 days
IU: Dihydrostreptomycin sulfat/Benzylpenicillinprokaine/Sulfadimidin pessaries**		3 x (500 mg/ 250.000 IE/4 g)	1st day	
	Tx2: IM: Sulfadiazin + trimethoprim (240 mg/ml)*****	50-60 ml	1-2 days	
	Practice B Herd 10 (Inclusion criteria – vaginal discharge score ≥6)	Tx1: IM: Benzylpenicillin procaine*	50-60 ml (300.000 IE/ml)	3days
IU: Dihydrostreptomycin sulfat/ Benzylpenicillinprokaine/Sulfadimidin **		3 x (500 mg/ 250.000 IE/4g)	1st day	
Tx1: IM: Benzylpenicillin procaine*		50-60 ml 300.000 IE/ml)	3days	

*Penovet®Vet. Boehringer Ingelheim DK A/S, Strødamvej 52, 2100 Copenhagen, Denmark

**Sulfa-streptocillin®Vet. Boehringer Ingelheim DK A/S, Strødamvej 52, 2100 Copenhagen, Denmark

***Aquacycline®Vet 10%. Ceva Animal Health A/S, Ladegårdsvej 2, 7100 Vejle, Denmark

****Terramycin®Vet.pessaries. Orion Pharma Animal health, ,Møllevej 9A, 2990 Nivå, Denmark

***** Norodine Vet. Scanvet (Norbrook), ScanVet Animal Health, Kongevejen 66, 3480 Fredensborg, Denmark

“Mess is difficult to describe. Especially in detail”

Nørretranders (1991)

Mærk verden – en beretning om bevidsthed.

Gyldendal

DEPARTMENT OF LARGE ANIMAL SCIENCES
FACULTY OF HEALTH AND MEDICAL SCIENCES
UNIVERSITY OF COPENHAGEN
PH.D. THESIS · 2012

ISBN 978-87-7611-538-8

DORTE BAY LASTEIN

Herd-specific Randomized Trials

– an approach for Effect Evaluation in a Dairy Herd Health Management Program