Invited review: Systematic review of diagnostic tests for reproductive-tract infection and inflammation in dairy cows[1]

M. W. de Boer,*†[2] S. J. LeBlanc,‡ J. Dubuc,§ S. Meier,# W. Heuwieser,‖ S. Arlt,‖ R. O. Gilbert,¶ and S. McDougall*

*Cognosco, Anexa Animal Health, Morrinsville 3300, New Zealand
†Epicentre, Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North 4442, New Zealand
‡Department of Population Medicine, Ontario Veterinary College, University of Guelph, Guelph, Ontario N1G 2W1, Canada
§Département de Sciences Cliniques, Faculté de Médecine Vétérinaire, Université de Montréal, Saint-Hyacinthe, Québec J2S 7C6, Canada
#DairyNZ Limited, Hamilton 3240, New Zealand
‖Clinic for Animal Reproduction, Faculty of Veterinary Medicine, Freie Universität Berlin, 14163 Berlin, Germany
¶Department of Clinical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853

ABSTRACT

The objective of this study was to conduct a systematic and critical appraisal of the quality of previous publications and describe diagnostic methods, diagnostic criteria and definitions, repeatability, and agreement among methods for diagnosis of vaginitis, cervicitis, endometritis, salpingitis, and oophoritis in dairy cows. Publications (n = 1,600) that included the words "dairy," "cows," and at least one disease of interest were located with online search engines. In total, 51 papers were selected for comprehensive review by pairs of the authors. Only 61% (n = 31) of the 51 reviewed papers provided a definition or citation for the disease or diagnostic methods studied, and only 49% (n = 25) of the papers provided the data or a citation to support the test cut point used for diagnosing disease. Furthermore, a large proportion of the papers did not provide sufficient detail to allow critical assessment of the quality of design or reporting. Of 11 described diagnostic methods, only one complete methodology, i.e., vaginoscopy, was assessed for both within- and between-operator repeatability (κ = 0.55–0.60 and 0.44, respectively). In the absence of a gold standard, comparisons between different tests have been undertaken. Agreement between the various diagnostic methods is at a low level. These discrepancies may indicate that these diagnostic methods assess different aspects of reproductive health and underline the importance of tying diagnostic criteria to objective measures of reproductive performance. Those studies that used a reproductive outcome to select cut points and tests have the greatest clinical utility. This approach has demonstrated, for example, that presence of (muco)purulent discharge in the vagina and an increased proportion of leukocytes in cytological preparations following uterine lavage or cytobrush sampling are associated with poorer reproductive outcomes. The lack of validated, consistent definitions and outcome variables makes comparisons of the different tests difficult. The quality of design and reporting in future publications could be improved by using checklists as a guideline. Further high-quality research based on published standards to improve study design and reporting should improve cow-side diagnostic tests. Specifically, more data on intra- and interobserver agreement are needed to evaluate test variability. Also, more studies are necessary to determine optimal cut points and time postpartum of examination.
Key words: vaginitis, purulent vaginal discharge, cervicitis, endometritis

INTRODUCTION

Systematic reviews use a predefined methodology for the selection of studies and then evaluate those studies based on a series of criteria designed to assess the experimental design, the sample size, the sampling approach, the statistical approach, and the strength of the inferences (Tranfield et al., 2003). Systematic reviews, together with meta-analyses, are regarded as the highest source of scientific evidence (Arlt et al., 2010). This methodology has been more commonly used in human medicine than in veterinary medicine and animal science, but is relevant in the latter as well (Sargeant et al., 2006; Grindlay et al., 2012).

The prevalence of endometritis in dairy cows is reported to be between 5 and 68% (Barlund et al., 2008; Gautam et al., 2009; Cheong et al., 2011). These large variations are at least partially due to inconsistencies of timing of examination relative to calving, diagnostic method, and definition of endometritis as well as true differences in prevalence between populations. Anaero-

bic and aerobic, gram-positive and gram-negative bacteria can be isolated from the uterus of more than 90% of cows in the first 2 wk postpartum, with the prevalence of infection declining with time (Földi et al., 2006). The time required for normal uterine and cervical involution varies among cows from 25 to 47 d after calving (LeBlanc, 2008). To generate more consistency, definitions have been proposed recently to define purulent vaginal discharge (clinical endometritis; **PVD**) and (cytological or subclinical) endometritis (Sheldon et al., 2006; Runciman et al., 2009; Dubuc et al., 2010a). Reporting the definition of disease and other critical information in papers on diagnosis of acute postpartum metritis in dairy cows is inconsistent (Sannmann et al., 2012).

High intra- and interobserver agreement are required for good quality tests (Greiner and Gardner, 2000a). Agreement can be statistically analyzed by 2 different methods: kappa statistics (value between -1 and 1; κ), which calculates agreement beyond chance (Dohoo et al., 2009), and the correlation between tests (value between -1 and 1; r; Greiner and Gardner, 2000a). The performance of diagnostic tests should ideally be validated against a test producing only correct results, i.e., a gold standard (Greiner and Gardner, 2000b). Some diagnostic tests produce a dichotomous test result (diseased or not diseased). Other tests will produce an ordinal or a continuous outcome (Greiner and Gardner, 2000b), such as a gross vaginal discharge score from 0 to 5 (McDougall et al., 2007) or the proportion of polymorphonuclear leukocytes (**PMN**) in a uterine cytology smear (Gilbert et al., 2005). For tests with ordinal or continuous outcomes, cut points need to be established to determine whether a test result is categorized as positive or negative (Greiner and Gardner, 2000b). Cut points can be established using receiver-operating characteristic analysis, which provides an assessment of sensitivity (**Se**) and specificity (**Sp**) over the range of test scores (Gardner and Greiner, 2006). Tests are described (test characteristics) using Se and Sp, which are the probability of a positive test result in a disease-positive animal and the probability of a negative test result in a nondiseased animal, respectively (Greiner and Gardner, 2000b). Used in conjunction with the prevalence of the condition, predictive values for test results can then be calculated to provide interpretive guidance.

Often a gold standard is not available (Gardner and Greiner, 2006). In these circumstances, tests are validated against a nonperfect test or a biological outcome, e.g., calving-to-pregnancy interval or pregnancy by a given interval postpartum (LeBlanc et al., 2002; Barlund et al., 2008). Statistical methods have also been developed for tests in absence of a gold standard (**TAGS**); these assume that neither test is perfect and adjust the estimates of Se and Sp accordingly (Pouillot et al., 2002). Finally, Bayesian methods can be used to develop receiver-operating characteristic curves to determine cut points when a gold standard is not available (Choi et al., 2006).

Traditional literature reviews may be biased if authors use criteria for inclusion or exclusion of specific papers that are not robust. For this reason, a more evidence-based approach, such as a systematic review, is required to reduce the potential lack of critical assessment (Tranfield et al., 2003). A systematic review uses a transparent and repeatable process to first select the papers to be included in a review and then second to use a consistent approach to assess the quality of the study design, case inclusion, clinical or laboratory procedures, analysis, and reporting. Instead of a traditional literature review, the aim of this study was to conduct a systematic review on diagnostic methods for reproductive-tract diseases in cows. No data are currently available on the quality of design and reporting of papers describing diagnostic methods for these diseases other than for metritis (Sannmann et al., 2012). The first objective was to critically appraise the quality of design and reporting of papers selected using an evidence-based method. A systematic review has not been performed on these diagnostic methods; therefore, other objectives were to assess diagnostic methods, diagnostic criteria and definitions, repeatability, and agreement among methods for diagnosis of reproductive-tract diseases in dairy cows (i.e., vaginitis, cervicitis, endometritis, salpingitis, and oophoritis). This appraisal was conducted using selection criteria, a data extraction template, and a quality checklist, which were developed a priori with the involvement of each of the authors of this manuscript.

METHODS

A protocol was developed a priori, which included a detailed description of the review process, the inclusion criteria, and the reporting process using guidelines from the Cochrane Collaboration (Higgins and Green, 2011) and the Centre for Reviews and Dissemination, University of York (Centre for Reviews and Dissemination, 2009). The populations of interest were postpartum dairy cows tested for vaginitis, cervicitis, endometritis, salpingitis, or oophoritis, irrespective of breed, type of housing, geographic location, or calving distribution. For this review, pathological definitions of the reproductive-tract diseases were used, that is, including both clinical (grossly evident) and subclinical (i.e., absence of clinically evident disease, hence relying on ancillary laboratory tests for diagnosis) disease. Vaginitis, cervicitis, endometritis, salpingitis, and oo-