J. Dairy Sci. 97:3983–3999
http://dx.doi.org/10.3168/jds.2013-7450

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-

Received September 1, 2013.
Accepted March 29, 2014.
[1]The authors have no conflicts of interest to disclose.
[2]Corresponding author: mdeboer@anexa.co.nz

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-

phoritis were defined as inflammation (measured as an increase in inflammatory cells, generally associated with an undesirable outcome or impaired reproductive performance) within the vagina, cervix, uterus, oviduct (uterine tube), or ovaries, respectively.

Studies conducted on dairy cattle that included these conditions, and where comparisons were made between diagnostic tests of any type or between a single test and a reproductive outcome, were selected for critical appraisal. Studies with interventions (treatments) for the reproductive-tract disease were included. These were only included when the study design (i.e., a negative control group was used in which cows received no interventions, or an assessment of diagnostic criteria before treatment) or analytical processes (i.e., stratification to consider the negative control group, or covariate adjustment of the effect of the intervention) dealt appropriately with confounding such that the test characteristics of the test itself could be assessed. Studies that did not control for these interventions, but were included for other reasons, were excluded from test validation assessed using reproductive outcomes.

Data from controlled trials, cohort studies, and quantitative study designs evaluating diagnostic tests were included. Primary papers reporting original data were included; reviews and meta-analyses were excluded. Studies reporting on in vitro and postmortem effects were excluded, as well as case reports and case series, or studies that were described as preliminary results, personal experiences, and unpublished data other than conference proceedings. Only papers in English and published in peer-reviewed journals and conference proceedings that were available online were considered. No date limitations were applied. After the selection process, only reported data were used—authors were not contacted to provide any additional information.

The literature search was performed by the first author on 6 February 2013 using CAB Abstracts, MEDLINE, and Web of Science simultaneously within the search engine Web of Knowledge using the search terms "dairy AND (cow* OR cattle OR bovine) AND (vaginitis OR purulent vaginal discharge OR cervicitis OR endometritis OR subclinical endometritis OR clinical endometritis OR cytological endometritis OR salpingitis OR oophoritis)." The selected search terms were kept broad to increase the search result. For example, search terms around diagnosis, such as "diagnostic tests," were not included to minimize the risk of failing to detect papers.

Selection and assessment of diagnostic papers were performed in 2 stages (Figure 1). In stage 1, all titles and abstracts of the identified studies were assessed by the first author using the eligibility criteria above. Only papers available at the libraries of Massey University, Palmerston North, New Zealand; Freie Universität Berlin, Germany; or University of Guelph, Canada, or available on the Internet were included for stage 2. This stage involved screening of full manuscripts. Each was comprehensively evaluated for inclusion by 2 assessors. The first author evaluated all manuscripts for inclusion. A second evaluation for inclusion was provided by one of the coauthors (22 to 24 papers per second assessor). To prevent bias, none of the authors evaluated manuscripts (co)authored by themselves. A manuscript was included when both assessors concluded that the inclusion criteria were met and a median of 7 papers per second assessor were included. Following agreement that a paper would be included in the review during the full evaluation at stage 2, 2 structured assessments were performed. Data on the contents and methods of each paper was entered into a spreadsheet (Excel; Microsoft) developed a priori, modified from the Cochrane Handbook (Supplementary Table S1: http://dx.doi.org/10.3168/jds.2013-7450; Higgins and Green, 2011). Also, the checklist for diagnostic methods for paratuberculosis in ruminants, STRADAS-paraTB (Standards for Reporting of Animal Diagnostic Accuracy Studies for paratuberculosis; Gardner et al., 2011), was modified to include criteria relevant to diagnostic methods for reproductive-tract diseases (Supplementary Table S2: http://dx.doi.org/10.3168/jds.2013-7450). The scoring of the individual items of the checklist was done on a 6-point scale (strongly agree to strongly disagree, or not determined; Arlt et al., 2010). Spreadsheets were collated in a purpose-built SQL database, and data were analyzed using Stata 12.1 (StataCorp, College Station, TX).

RESULTS AND DISCUSSION

Search Results and Paper Selection

In this review, 51 papers were critically appraised to assess the currently available diagnostic methods for vaginitis, cervicitis, endometritis, salpingitis, and oophoritis in dairy cows (Supplementary Table S3: http://dx.doi.org/10.3168/jds.2013-7450). Initially 689, 40, and 871 publications were identified by the databases CAB Abstracts, MEDLINE, and Web of Science, respectively. The combination of these databases covers the vast majority of veterinary and animal science journals (Grindlay et al., 2012). Therefore, no other methods (e.g., manual searches) beyond the initial database search were used to retrieve additional papers, and manuscripts published after the search date (6 February 2013) were not included.

Included Papers. Selection for inclusion was made on the basis of criteria based on relevance developed

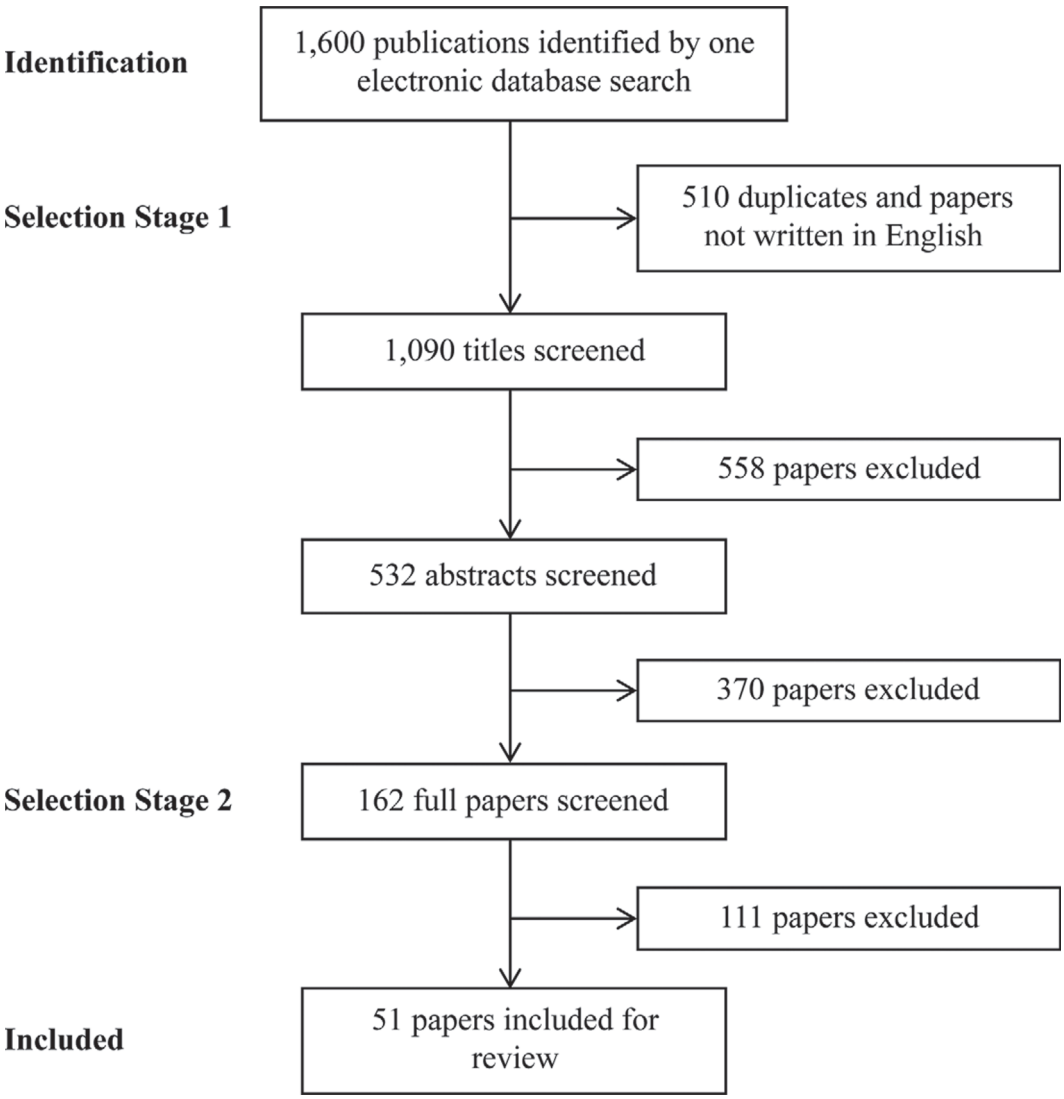

Figure 1. Flowchart of selection process of papers identified on 6 February 2013 by using the search terms "dairy AND (cow* OR cattle OR bovine) AND (vaginitis OR purulent vaginal discharge OR cervicitis OR endometritis OR subclinical endometritis OR clinical endometritis OR cytological endometritis OR salpingitis OR oophoritis)" in the 3 databases CAB Abstracts, MEDLINE, and Web of Science simultaneously within the search engine Web of Knowledge for quality appraisal and data synthesis.

a priori and approved by all authors, similar to other systematic reviews (Sannmann et al., 2012). To reduce bias, each full evaluation of the papers was undertaken by 2 assessors. Fair agreement ($\kappa = 0.33$) existed between pairs of assessors for including papers. The majority of the discrepancies (n = 27) between the assessors were on studies not accounting for treatments of reproductive-tract disease by study design or statistical analyses. Prior to the inclusion process, the inclusion criteria were not tested on papers by the assessors. This may be a possible reason for these disagreements. Therefore, training sessions undertaken before start of a systematic review may decrease heterogeneity of interpretation, even when all are involved preparing the inclusion criteria.

Excluded Papers. Of the excluded manuscripts (n = 1,549; Figure 1), 55 were the same data set published twice, 458 were written in a language other than English, 30 were in vitro or postmortem studies, 45 were studies on animals other than dairy cows, 78 were not evaluating 2 or more diagnostic tests or one test with reproductive outcomes, 614 did not report on a disease relevant to this review or the disease of interest was an outcome variable for other conditions, 71 did not report original data, 59 were not published in journals or issues of conference proceedings (e.g., books and theses),

and 90 publications were not full manuscripts (e.g., short communications, letters, or abstract only). In 39 cases treatment was not accounted for in the study design or statistical analyses, and 10 were classified as case-control studies or case reports.

Although broad search terms were used, this resulted in exclusion of an unknown number of potential relevant papers, e.g., a publication by Bonnett et al. (1993). Besides "dairy cows," that publication did not include any of the other search terms used in the title or abstract, and no key words were described. It is unclear how many publications were missed because of this, and this emphasizes the importance of the use of appropriate key words in papers. Similar to other systematic reviews, papers written in a language other than English and those not published in peer-reviewed journals were excluded (Roy and Keefe, 2012; Sannmann et al., 2012). Papers that appear relevant may not have been included due to the selection of search terms or may have been selected but were subsequently excluded because they failed to meet the specific inclusion criteria.

Quality Assessment of Included Papers

Disease Definition, Test Characteristics, and Cut Points. A full description of the disease, the diagnostic method, and the rationale of the diagnostic cut point is important information that needs to be provided to make an informed judgment on the validity of the diagnostic test, the test performance, and application of the results. In total, 31 (61%) papers referenced the disease of interest or the diagnostic method for disease detection, whereas the remaining 20 (39%) papers described the disease or diagnostic method without citation. Sannmann et al. (2012) reported an even higher proportion of papers on diagnostic methods for acute postpartum metritis (64%) that did not state or cite the disease definition. In total, 25 (49%) papers included in this review referenced or derived from the study data the described diagnostic cut points. In 15 (60%) of these papers, cut points were justified by citing references, 8 (16%) used statistical techniques to analyze cut points, and 2 (4%) provided both for different tests used (Table 1).

The test characteristics of diagnostic methods were discussed in 29 (57%) of the included papers. In 21 (41%) and 23 (45%) of the papers diagnostic cut points and possible sources of error were discussed, respectively. Only 20% of the reviewed papers discussed all 3 criteria. Test characteristics, cut points, and sources of error are important criteria that can influence selection and application of tests and interpretation of test results. To be able to assess and compare the diagnostic methods or results, this information should be discussed (Sannmann et al., 2012) so that the reader is aware of these factors influencing the study outcome.

Checklist for Quality of Design and Reporting. A large proportion of the papers did not provide sufficient detail to allow critical assessment of the quality of design and reporting (Supplementary Table S2: http://dx.doi.org/10.3168/jds.2013-7450). For only one criterion (item 2: stating the research question or study aims) a large majority of the papers (82%) scored "strongly agree" or "agree." For 5 criteria (items 1, 5, 16, 17, and 22) including the description of diagnostic test, sampling protocol, or both and cross tabulation of results, scores were "agree" or "strongly agree" for approximately half of the papers, whereas for 8 criteria (items 4, 11, 12, 13, 15, 23, 25, and 26) including a description of the selection methods of animals and herds, and the use of blinding methods, 51 to 100% of the papers were marked "disagree" or "strongly disagree." Agreements between assessors for each criterion are described in Supplementary Table S2 (http://dx.doi.org/10.3168/jds.2013-7450). In total 19 (66%) criteria had a fair or higher agreement ($\kappa > 0.20$). As would be expected, agreement between both assessors was slightly higher when responses "agree" and "strongly agree," and "disagree" and "strongly disagree" were combined (Supplementary Table S2: http://dx.doi.org/10.3168/jds.2013-7450). Reduced agreement may be due to the review design, where the second assessor was any 1 of 7 individuals. Ideally, all observers would have rated all included papers, but because of time limitations, this design was not practical. In common with many published systematic reviews, the current study used 2 assessors during the selection and assessment process (Arlt et al., 2010; Haimerl et al., 2012; Roy and Keefe, 2012). It is recommended to use a minimum of 2 assessors for the evaluation of quality to improve objectivity (Higgins and Green, 2011). Additionally, it is unclear if inclusion of more assessors would improve agreement. Simoneit et al. (2012) found slightly higher interobserver agreements following use of a checklist on bovine reproduction papers when assessed by 14 observers. However, observers only reviewed 3 preselected papers and received additional information on use of the checklist, whereas for this review no additional information beyond the checklist was provided. In the absence of additional data about the effect of number of assessors on repeatability of quality, it is unclear if using more reviewers would have improved the paper selection and assessment process (Arlt et al., 2010). Also, no formal training was done to improve agreements among assessors in the current study. Thus the assessment of quality and statistical validity was reliant on the training and experience of those involved. On the other hand,

Table 1. Summary of cut points for the proportion of polymorphonuclear leukocytes (PMN) in uterine cytology, leukocyte esterase, protein and pH reagent test strips, and the optical density (OD) of fluid retrieved following uterine lavage for diagnosis of endometritis in dairy cows[1]

Diagnostic technique	DIM	Sample size	Excluded diseases	Statistical method	Reference outcome	Cut point	Reference
Endo cytobrush[2]	21 to 47	285	RFM,[3] metritis, PVD[4]	ROC analysis[5]	Pregnant by 90 DIM	6.7% PMN	Couto et al., 2013
Endo cytobrush	35	1,044	None	Survival analysis	Pregnant by 120 DIM	6% PMN	Dubuc et al., 2010a
Endo cytobrush	56	1,044	None	Survival analysis	Pregnant by 120 DIM	4% PMN	Dubuc et al., 2010a
Endo cytobrush	28 to 41	221	None	Survival analysis	Pregnant by 150 DIM	8% PMN	Barlund et al., 2008
Endo cytobrush	20 to 33	228	PVD	ROC analysis	Pregnant by 132 DIM	18% PMN	Kasimanickam et al., 2004
Endo cytobrush	34 to 47	228	PVD	ROC analysis	Pregnant by 132 DIM	10% PMN	Kasimanickam et al., 2004
Endo cytobrush	21 to 34	168	None	Survival analysis	Pregnant by 300 DIM	6.5% PMN	Deguillaume et al., 2012
Endo cytobrush	28	303	None	Descriptive[6]	—	9% PMN	McDougall et al., 2011
Endo cytobrush	42	303	None	Descriptive	—	7% PMN	McDougall et al., 2011
Uterine lavage	21	445	Pyometra, adhesions, abscesses	ROC analysis	Pregnant by 150 DIM	8.5% PMN	Galvão et al., 2009
Uterine lavage	35	445	Pyometra, adhesions, abscesses	ROC analysis	Pregnant by 150 DIM	6.5% PMN	Galvão et al., 2009
Uterine lavage	49	445	Pyometra, adhesions, abscesses	ROC analysis	Pregnant by 150 DIM	4% PMN	Galvão et al., 2009
Cervical cytobrush	21 to 34	168	None	Survival analysis	Pregnant by 300 DIM	5% PMN	Deguillaume et al., 2012
Uterine lavage	40 to 60	563	PVD	ROC analysis	Cytology (>10% PMN)	2+ (LE[7])	Cheong et al., 2012
Uterine lavage	40 to 60	563	PVD	ROC analysis	Cytology (>10% PMN)	3+ (Protein[8])	Cheong et al., 2012
Uterine lavage	40 to 60	563	PVD	ROC analysis	Cytology (>10% PMN)	7 (pH[9])	Cheong et al., 2012
Uterine lavage	34 to 36	1,742	Pyometra	ROC analysis	Gross uterine discharge[10]	0.058 (OD[11])	Machado et al., 2012
Uterine lavage	34 to 36	1,742	Pyometra	ROC analysis	Cytology (>18% PMN)	0.059 (OD)	Machado et al., 2012

[1]Only cut points that were analyzed using different statistical methods on original data are provided; described or referenced cut points are not included.

[2]Endo cytobrush: endometrial cytobrush.

[3]RFM: retained fetal membranes.

[4]PVD: purulent vaginal discharge.

[5]ROC analysis: receiver-operating characteristic analysis.

[6]Descriptive: upper quartile of the distribution.

[7]LE: leukocyte esterase reagent test strip [0 (no leukocytes), trace, + (small), 2+ (moderate), and 3+ (large)].

[8]Protein: protein reagent test strip [0 (no proteins), trace, + (30 mg/dL), 2+ (100 mg/dL), 3+ (300 mg/dL), and 4+ (>2,000 mg/dL)].

[9]pH: pH reagent test strip (5.0, 6.0, 6.5, 7.0, 7.5, 8.0, and 8.5).

[10]Visual gross uterine discharge in uterine lavage fluid.

[11]OD: optical density (wavelength = 620 nm).

all assessors were involved in the development and modification of the checklist. Additionally, some of the authors have previously published papers on literature assessments using different checklists (Arlt et al., 2010; Simoneit et al., 2012). In contrast, lower agreement among assessors may be attributable to unclear or incomplete reporting leading to difficulty in making an assessment more than to flaws in the assessment tool (Smidt et al., 2006b). It should also be noted that although kappa statistics are commonly calculated with 2 observers and dichotomous outcomes (e.g., diseased or not diseased), the current study had 6 possible scores (or 4 when combined). Hence, a score one unit apart is a relatively small difference and does not necessarily reflect substantial variation in assessment of the level of quality of reporting in a paper. Although it is common not to publish assessment of agreement between assessors of quality (Siddiqui et al., 2005; Zafar et al., 2008; Fontela et al., 2009), the current study describes fair to moderate agreement. Therefore, the level of agreement between assessors during the quality criteria assessment in the current study needs to be interpreted with some caution, although it is not clear if our assessments are any more divergent than is typical.

The majority of papers included in the current study did not provide data on all items of the checklist. The quality assessment of papers in our manuscript was modeled on the STRADAS-paraTB statement (Standards for Reporting of Animal Diagnostic Accuracy Studies for paratuberculosis; Gardner et al., 2011). That statement was in turn modified by independent experts from the STARD statement (Standards for Reporting of Diagnostic Accuracy), which aims to improve reporting of test accuracy studies in human medicine (www.stard-statement.org; Bossuyt et al., 2003). The impetus for the development of the STRADAS-paraTB was the review by Nielsen and Toft (2008) on diagnostic tests for paratuberculosis (Gardner et al., 2011). The conclusion of that review was that "the quality of design, implementation, and reporting of evaluations of tests for paratuberculosis was poor" (Nielsen and Toft, 2008). Others report similar concerns in other areas of veterinary and animal science (Arlt et al., 2010; Sannmann et al., 2012), as well as in human medical literature (Siddiqui et al., 2005; Zafar et al., 2008; Fontela et al., 2009).

Although the peer-review system is a good tool that enhances the quality of published manuscripts (Goodman et al., 1994; Purcell et al., 1998), this process has its limitations. The quality of papers varies even in peer-reviewed journals, and acceptance for publication does not guarantee the completeness, clarity, or credibility of papers, even in journals with a high impact factor (Kastelic, 2006; Benos et al., 2007; Arlt et al., 2010). Hence, even when published in peer-reviewed, high-impact journals, papers on reproductive diagnostic may lack sufficient information to allow critical appraisal. Besides the STARD and STRADAS-paraTB statements, other statements have been developed to improve publication standards, such as CONSORT (Schulz et al., 2010), STROBE (von Elm et al., 2007), and REFLECT (Sargeant et al., 2010). Despite discrepancies between assessors and the limitations of checklists when applied to diverse study types (Smidt et al., 2006b), quality of reporting and design has improved in journals adapting these guidelines (Moher et al., 2001; Smidt et al., 2006a). Therefore, authors of future papers on diagnostic methods of reproductive-tract disease are advised to use a guideline to improve the clarity and consistency of study design and reporting.

Distribution of Diseases

In this review, the diseases were described on a pathological basis (i.e., inflammation of some part of the reproductive tract). This approach was taken to ensure that diagnostic methods that encompassed clinical diseases (e.g., detection of grossly evident purulent material in the vagina) as well as subclinical disease (e.g., definition of endometritis based on PMN% in endometrial cytology) would be included. Purulent vaginal discharge is a symptom or condition of an inflammatory process, as defined by a variety of diagnostic methods, rather than being a specific etiological or pathological diagnosis. Therefore, papers on diagnostic methods for PVD were included in papers describing vaginitis as this is where material is collected for assessment. Thus, even though PVD is often used as proxy for uterine inflammation, it is an assumption that the purulent material originates only from within the uterine lumen; regardless of origin, it is identified clinically in the vagina. It is clear, from recent studies, that PVD is not always coincident with endometritis and may occur independently (Dubuc et al., 2010a), suggesting that the inflammation originates from the cervix or vagina. Also, even if the primary source of (muco)purulent material is the uterus, presence of such material in the vagina might induce vaginitis. Recent studies have clearly demonstrated that endometritis, cervicitis, and vaginitis (including PVD) are related but not synonymous conditions (Dubuc et al., 2010a,b; Deguillaume et al., 2012).

The most common diagnosis described in the reviewed papers was endometritis (n = 45; 88%), followed by vaginitis (n = 29; 57%) and cervicitis (n = 4; 8%). It is only recently that a research interest has developed in the effects of inflammation of the cervix, which likely explains the relatively low number of studies of

this condition. No diagnostic studies were included on salpingitis and oophoritis. It is unlikely that these 2 diseases totally coincide with other reproductive-tract diseases; therefore, specific diagnostic methods may be required for their diagnosis. The prevalence of these conditions is assumed to be low, but diagnosis is difficult practically and methods for use in live animals have not been validated.

Distribution of Diagnostic Methods

Cervical diameter, uterine horn size, thickness of the uterine wall, and volume of intrauterine fluids can be assessed by transrectal palpation or ultrasonography (LeBlanc et al., 2002; Barlund et al., 2008) and were reported in 18 (35%) and 10 (20%) papers, respectively. Purulent material in the vagina may be visualized by the use of a speculum (LeBlanc et al., 2002) and detected by the introduction and retraction of a clean gloved hand (gloved hand; Plöntzke et al., 2011; n = 5; 10%) or a stainless steel rod with a rubber hemisphere attached at the end (Metricheck, Simcrotech, Hamilton, New Zealand; McDougall et al., 2007; n = 11; 22%). Cervical and uterine samples to evaluate inflammatory cells or bacterial growth can be obtained by a cytobrush or swab (Yavari et al., 2009; Deguillaume et al., 2012; n = 21; 41%), cervical fluid aspiration or uterine lavage (Gilbert et al., 2005; Yavari et al., 2009; n = 13; 25%), and biopsy (Bonnett et al., 1991; n = 6; 12%).

Cut Points and Reported Cut-Point Analyses

Similar to many other diagnostic methods in medicine, no gold standard test is available for reproductive-tract diseases in cows (Sheldon et al., 2006). This complicates the determination of cut points as well as the evaluation of diagnostic tests (Gardner and Greiner, 2006). The recommended receiver-operating characteristic analysis for the determination of cut points (Gardner and Greiner, 2006) was used in half of the papers. However, without the availability of a gold standard, these papers instead used pregnancy by 90 to 150 DIM or >10% PMN by uterine lavage as reference outcomes (Table 1). Reproductive performance as a reference outcome has the advantage of being tangible and of economic importance, but it has the disadvantage of being influenced by a multitude of factors other than the disease condition of interest, and it does not directly measure pathological processes. Mucopurulent vaginal discharge or a cervical diameter >7.5 cm at >20 DIM or PVD at >26 DIM was predictive for nonpregnancy by 120 DIM (LeBlanc et al., 2002). Four other papers calculated cut points using similar time to event (pregnancy) techniques, and one divided the continuous scoring scale into quartiles (Table 1). Bayesian methods can also be used to determine cut points (Choi et al., 2006). It is interesting to note that none of the studies used these methods (although McDougall et al., 2007, applied this approach for the assessment of the Metricheck device). Prior assessment of likely test performance is used in Bayesian methodology (Gardner and Greiner, 2006). The Bayesian approach may be a more robust approach for using a distribution of possible values instead of a fixed (unknown) parameter used in classical methods (Enøe et al., 2000). Potential reasons for the limited use of Bayesian analysis are the difficulty to understand and implement these methods, and analytical convergence issues when the range of selected priors is not wide enough. However, it may be valuable to pursue these methods in future research.

Intra- and Interobserver Agreements

Of the 51 included papers, only one reported intra- and interobserver agreements of a diagnostic test (vaginoscopy; Table 2). Moderate agreement was calculated between 3 operators, of which one operator was a group of inexperienced veterinary students (Leutert et al., 2012). Within- and between-reader agreements and correlations for the evaluation of cytological microscope slides generated by samples from the cytobrush or uterine lavage techniques were determined by 6 studies (Table 2). These studies generally assessed the repeatability of multiple readings of a single slide created by a single cow-side operator at one time point. Hence, these assessments are limited to laboratory variability of the test and not complete methodology (e.g., the on-farm, between-cow, between-operator variability). Low repeatability is associated with lower Se and Sp of the diagnostic method. The repeatability study of vaginoscopy was performed in one herd, and cows were examined at one time point (Leutert et al., 2012). Therefore, unfortunately, no data are available for tests performed at different time points and between different populations.

Comparison Among Tests

Agreement between various diagnostic methods was reported in 12 of the 51 included papers. In 11 of these papers, more than one agreement was assessed. Intertest agreements between vaginoscopy and other tests were reported in 5 papers, whereas agreements between Metricheck, ultrasonography, cytobrush, swab, uterine lavage, biopsy, and leukocyte esterase test and other tests were reported in 6, 2, 7, 1, 1, 2, and 1 papers, respectively (Table 3). Agreement measures (i.e., κ and r) compare diagnostic methods, irrespective of the tests

Table 2. Reported intra- and interobserver agreements of diagnostic methods for reproductive-tract disease in papers (n = 7) included in a systematic review

		Intraobserver		Interobserver			
Diagnostic method	DIM	Statistic[1]	Value	No. of observers	Statistic	Value	Reference
Vaginoscopy	21 to 27	κ	0.55–0.60	3	κ	0.44	Leutert et al., 2012
Cytology slide (cytobrush)[2]	35	κ	0.82	2	κ	0.77	Dubuc et al., 2010b
Cytology slide (cytobrush)	20 to 47	r	0.84	2	r	0.84	Kasimanickam et al., 2004
Cytology slide (cytobrush)	28 to 41	r	0.85				Barlund et al., 2008
Cytology slide (cytobrush)	29 and 43			2	r	0.82	McDougall et al., 2011
Cytology slide (uterine lavage)[2]	40 to 60	κ	0.86				Gilbert et al., 2005
Cytology slide (uterine lavage)	28 to 41	r	0.76				Barlund et al., 2008
Cytology slide (both)[2]	20 to 47			2	r	0.90	Kasimanickam et al., 2005

[1]The kappa statistic gives a value between −1 and 1, where ≤0 is no agreement and 1 is perfect agreement beyond chance. The correlation coefficient (r) gives a value between −1 and 1, where −1 is perfect negative association and 1 is perfect positive association. Complete independence has a value of 0.

[2]Cytology slide: These studies assessed the repeatability of multiple readings of a single cytology slide created by a single cow-side operator at one time point using the cytobrush, uterine lavage, or both techniques. Hence, these assessments are limited to laboratory variability of the test and not the complete methodology (e.g., on-farm, between-cow, between-operator variability).

being correct or not. Sensitivity and Sp are measurements used for validation of diagnostic methods. Seven papers reported Se and Sp, with other diagnostic tests as the reference method (Table 4). These reference methods were previously validated as being associated with reproductive performance. The Bayesian approach for TAGS was used in one study to determine Se and Sp and reported reasonable to high Se and Sp for vaginoscopy and Metricheck (Table 4).

Studies comparing vaginoscopy, Metricheck and ultrasonography with cytobrush as the reference test report overall low Se and moderate to high Sp (Table 4), indicating a high number of false negatives and a low number of false positives. For example, approximately half of the cows were PVD negative by vaginoscopy but had a PMN score >8% (Barlund et al., 2008). The Se and Sp for vaginoscopy and Metricheck with cytobrush as the reference were similar but were only described in 2 papers. The Sp in these 2 papers was slightly higher than calculated with the TAGS approach, whereas Se analyzed with the TAGS approach was considerably higher. The Se (poor to high) and Sp (mediocre to high) of ultrasonography relative to cytology are inconsistent (Table 4).

Controversy exists about the magnitude and direction of association between the bacterial species isolated from the uterus and the different diagnostic methods for PVD and PMN% as detected by the cytobrush technique. In some studies isolation of any bacteria, or even specific bacterial species, is not directly or consistently associated with the degree of inflammation within the uterus or vagina (Table 3), whereas others found positive associations between isolation of specific bacterial species and PVD score (Williams et al., 2005). Similar and positive correlations are reported for the agreement between the inflammation score for biopsies and isolation of bacteria (Table 3). The potential lack of association between bacteriology and other tests may not be surprising given the different biological bases upon which the tests are based i.e., assessing different elements of the immune or inflammatory response to bacterial infection and tissue trauma.

In the search for systemic diagnostic tests for reproductive-tract diseases, studies compared outcomes of reproductive tract–based diagnostic methods (e.g., PVD, cytology) with indirect or systemic tests such as hematology and biochemistry (Green et al., 2009), local and circulating concentrations of cytokines (Ishikawa et al., 2004; Kim et al., 2005; Fischer et al., 2010), acute phase proteins (Williams et al., 2005; Dubuc et al., 2010b), NEFA and BHBA (Dubuc et al., 2010b; Senosy et al., 2012), and hormones (e.g., progesterone and prostaglandin $F_{2\alpha}$ metabolite; Seals et al., 2002; Senosy et al., 2011). Although associations may be found among these tests, the direction of causality is not clear. For example, reproductive-tract inflammation may directly result in elevated acute phase protein concentrations. Alternatively, systemic inflammation may stimulate release of proinflammatory cytokines and acute phase proteins, which may contribute to impaired immune defenses or reproductive-tract inflammation. Hence, where such associations are found, these outcomes may not be specific enough to be a predictive test for reproductive-tract diseases. Additionally, intrauterine concentrations of cytokines may be more closely associated with uterine disease than the circulating concentrations (Galvão et al., 2011).

Complications with Comparison Among Tests

The determination of PMN% is most commonly used as the near–gold standard test for the calculation of Se

Table 3. Reported agreements between diagnostic methods of reproductive-tract disease in papers included in a systematic review (n = 12)

Diagnostic method	Outcome	Comparison method	Outcome	DIM	Statistic[1]	Value	Reference
Metricheck	PVD[2]	Vaginoscopy	PVD	33 ± 16[3]	κ	0.45	McDougall et al., 2007
Metricheck	VDS (0–5)[4]	Vaginoscopy	VDS (0–5)	33 ± 16[3]	κ	0.27	McDougall et al., 2007
Metricheck	PVD	Vaginoscopy	PVD	7 to 28	κ	0.73	Runciman et al., 2009
Metricheck	VDS (0–3)[5]	Vaginoscopy	VDS (0–3)	7 to 28	κ	0.59	Runciman et al., 2009
Metricheck	VDS (0–3)	Ultrasound	Uterine horn diameter	18 to 46	r	0.52[6]	Senosy et al., 2009
Metricheck	VDS (0–3)	Ultrasound	IUF[7] (yes/no)	18	r	0.49[6]	Senosy et al., 2009
Metricheck	VDS (0–3)	Cytobrush	% PMN[8]	25 and 32	r	0.59	Senosy et al., 2009
Metricheck	VDS (0–5)	Cytobrush	≥6% PMN	35	κ	0.14–0.20	Dubuc et al., 2010a
Metricheck	VDS (0–5)	Cytobrush	≥9% PMN	28	κ	0.29	McDougall et al., 2011
Metricheck	VDS (0–5)	Cytobrush	≥7% PMN	42	κ	0.12	McDougall et al., 2011
Metricheck	VDS (0–5)	Cytobrush	≥8% PMN	28 to 41	κ	0.30	Peter et al., 2011
Vaginoscopy	PVD	Cytobrush	≥8% PMN	28 to 41	κ	0.52	Barlund et al., 2008
Vaginoscopy	VDS (0–3)	Cytobrush	≥5% PMN	21 to 27	r	0.30	Westermann et al., 2010
Vaginoscopy	VDS (0–3)	Cytobrush	≥18% PMN	21 to 27	r	0.30	Westermann et al., 2010
Vaginoscopy	PVD	Biopsy[9]	Hist[10] score	28 to 35	r	0.36	Studer and Morrow, 1978
Vaginoscopy	PVD	Biopsy[11]	Hist score	28 to 35	r	0.42	Studer and Morrow, 1978
Ultrasound	IUF (>3 mm)	Cytobrush	% PMN (2 cut points)[12]	20 to 42	κ	0.28	Kasimanickam et al., 2004
Uterine lavage	≥8% PMN	Cytobrush	≥8% PMN	28 to 41	κ	0.74	Barlund et al., 2008
Uterine lavage	≥8% PMN	Cytobrush	≥8% PMN	28 to 41	r	0.66	Barlund et al., 2008
Cervical LE[13]	0–3+	Uterine LE	0–3+	21 to 47	κ	0.37	Couto et al., 2013
Metricheck	VDS (0–5)	Cytobrush	5 Bacteria[14]	29	κ	0.05	McDougall et al., 2011
Metricheck	VDS (0–5)	Cytobrush	5 Bacteria	42	κ	0.00	McDougall et al., 2011
Vaginoscopy	PVD	Uterine swab	Bacteria	28 to 35	r	0.44	Studer and Morrow, 1978
Vaginoscopy	VDS (0–3)	Cytobrush	*T. pyogenes*	21 to 27	r	0.40	Westermann et al., 2010
Cytobrush	≥5% PMN	Cytobrush	*T. pyogenes*	21 to 27	r	0.42	Westermann et al., 2010
Cytobrush	≥18% PMN	Cytobrush	*T. pyogenes*	21 to 27	r	0.42	Westermann et al., 2010
Cytobrush	≥9% PMN	Cytobrush	5 Bacteria	29	κ	0.14	McDougall et al., 2011
Cytobrush	≥7% PMN	Cytobrush	5 Bacteria	42	κ	0.12	McDougall et al., 2011
Biopsy	Hist score	Uterine swab	Bacteria	28 to 35	r	0.27	Studer and Morrow, 1978
Biopsy	Hist score	Biopsy	*T. pyogenes*	26	r	0.25	Bonnett et al., 1991
Biopsy	Hist score	Biopsy	*T. pyogenes*	40	r	0.37	Bonnett et al., 1991
Biopsy	*T. pyogenes*[15]	Biopsy	*T. pyogenes*	40	r	0.63	Bonnett et al., 1991
Biopsy	*T. pyogenes*[15]	Biopsy	Hist lesions	40	r	0.27	Bonnett et al., 1991

[1]The kappa statistic (κ) gives a value between −1 and 1, where ≤0 is no agreement and 1 is perfect agreement beyond chance. The correlation coefficient (r) gives a value between −1 and 1, where −1 is perfect negative association and 1 is perfect positive association. Complete independence has a value of 0.

[2]PVD: presence of purulent vaginal discharge with a dichotomous outcome (positive or negative).

[3]"At-risk" cows (i.e., those with retained fetal membranes, metritis, or twins, i.e., a population in which a high prevalence of PVD would be expected) were assessed 35 d before the start of the breeding season.

[4]VDS (0–5): vaginal discharge score = no mucus to >50% purulent vaginal discharge and odor.

[5]VDS (0–3): vaginal discharge score = clear or translucent mucus to >50% purulent vaginal discharge.

[6]Agreement only in cows with a corpus luteum.

[7]IUF: intrauterine fluid.

[8]PMN: polymorphonuclear leukocytes.

[9]Biopsy taken from right horn.

[10]Hist: histopathological.

[11]Biopsy taken from left horn.

[12]The 2 cut points: >18% PMN at 21–33 DIM and >10% PMN at 34–47 DIM.

[13]LE test: leukocyte esterase test [0 (no leukocytes), trace, + (small), 2+ (moderate), and 3+ (large)].

[14]*Trueperella pyogenes*, *Fusobacterium necrophorum*, *Prevotella melaninogenica*, *Proteus* spp., or *Escherichia coli*.

[15]*Trueperella pyogenes* cultured at 26 DIM.

and Sp (Table 4). However, no data are available that support this, and it seems that endometritis assessed by cytology is only one element of reproductive-tract inflammatory disease, along with PVD and cervicitis (Dubuc et al., 2010a; Deguillaume et al., 2012). Also, the use of different cut points makes it difficult to compare test validation studies. Inflammation can also be evaluated by histopathological assessment of tissue obtained by biopsy (Bonnett et al., 1991). The extent of variability in intra- and interobserver repeatability of both tests in cows is unknown. Most papers do not describe how often and with what pressure the cytobrush is rolled on the cervix or endometrium. It can be hypothesized that this may influence the PMN:epithelial cell ratio depending on the depth of cells that are removed. Moreover, the cytobrush and biopsy techniques sample

Table 4. Sensitivity (Se) and specificity (Sp) of tests for reproductive-tract disease relative to other, validated diagnostic tests reported in papers included in a systematic review (n = 7)

Diagnostic method	Reference method/outcome	DIM	Se	Sp	Reference
Vaginoscopy (≥flecks of pus)	Bayesian TAGS[1] approach	33 ± 16	72	87	McDougall et al., 2007
Vaginoscopy (≥mucopurulent)	Cytobrush (>8% PMN[2])	28 to 41	54	96	Barlund et al., 2008
Metricheck (≥flecks of pus)	Bayesian TAGS approach	33 ± 16	96	78	McDougall et al., 2007
Metricheck (≥flecks of pus)	Cytobrush (>8% PMN)	28 to 41	44	89	Peter et al., 2011
Ultrasound (ET[3] >7 mm)	Cytobrush (>8% PMN)	28 to 41	23	75	Barlund et al., 2008
Ultrasound (ET >8 mm)	Cytobrush (>8% PMN)	28 to 41	4	89	Barlund et al., 2008
Ultrasound (IUF[4] >1 mm)	Cytobrush (>8% PMN)	28 to 41	39	78	Barlund et al., 2008
Ultrasound (IUF >3 mm)	Cytobrush (>8% PMN)	28 to 41	31	93	Barlund et al., 2008
Ultrasound (IUF present)	Cytobrush (2 cut points[5])	21 to 47	88	62	Meira et al., 2012
Ultrasound (cervix >5.0 cm)	Cytobrush (2 cut points)	21 to 47	56	73	Meira et al., 2012
Biopsy (score <15)	Cytobrush (2 cut points)	21 to 47	44	92	Meira et al., 2012
Combination[6]	Cytobrush (2 cut points)	21 to 47	44	97	Meira et al., 2012
Uterine lavage (>8% PMN)	Cytobrush (>8% PMN)	28 to 41	92	94	Barlund et al., 2008
ULOSD[7] (>0.058)	Pus IUF[8]	34 to 36	76	78	Machado et al., 2012
ULOSD (>0.059)	Uterine lavage (>18% PMN)	34 to 36	100	82	Machado et al., 2012
LE[9] (>2+)	Cytobrush (>10.2% PMN)	21 to 47	69	73	Couto et al., 2013
LE (>2+)	Uterine lavage (>10% PMN)	40 to 60	77	52	Cheong et al., 2012
pH (7.0)	Uterine lavage (>10% PMN)	40 to 60	45	79	Cheong et al., 2012
Protein (3+)	Uterine lavage (>10% PMN)	40 to 60	58	56	Cheong et al., 2012
LE (3+) and pH (7.0)	Uterine lavage (>10% PMN)	40 to 60	19	97	Cheong et al., 2012
NEFA (serum; ≥0.5 mmol/L)	Metricheck (PVD)	−7 to −1	54	53	Dubuc et al., 2010b
NEFA (serum; ≥1.0 mmol/L)	Metricheck (PVD)	1 to 7	41	66	Dubuc et al., 2010b
NEFA (serum; ≥0.9 mmol/L)	Metricheck (PVD)	8 to 14	43	64	Dubuc et al., 2010b
BHBA (serum; ≥1,100 μmol/L)	Metricheck (PVD)	1 to 7	28	84	Dubuc et al., 2010b
BHBA (serum; ≥700 μmol/L)	Metricheck (PVD)	8 to 14	59	48	Dubuc et al., 2010b
Haptoglobin (serum; ≥0.8 g/L)	Metricheck (PVD)	1 to 7	39	80	Dubuc et al., 2010b
Haptoglobin (serum; ≥0.3 g/L)	Metricheck (PVD)	8 to 14	47	67	Dubuc et al., 2010b

[1]TAGS: tests in absence of a gold standard.

[2]PMN: polymorphonuclear leukocytes.

[3]ET: endo thickness.

[4]IUF: intrauterine fluid.

[5]The 2 cut points: >18% PMN at 21 to 33 DIM and >10% PMN at 34 to 47 DIM.

[6]Combination: ultrasound (intrauterine fluid present), ultrasound (cervix diameter >5.0 cm), and biopsy (score <15).

[7]ULOSD: uterine lavage sample optical density.

[8]Pus IUF: gross uterine discharge in uterine lavage fluid.

[9]LE: leukocyte esterase test.

only a small area. In horses, diagnosing endometritis using these techniques was not representative for the entire endometrium and seemed to have a low within-horse repeatability (Overbeck et al., 2013). It is unclear if this is also true for the bovine uterus. It has also been reported that up to 41% of the biopsies taken can be unsatisfactory for histological evaluation (Meira et al., 2012). This high failure rate may create bias; cows with more severe intrauterine pathology might be less likely to yield biopsy material that is considered assessable.

Uterine lavage may sample a larger area than the cytobrush. However, the period during which fluid is left in the uterus and how often and with what pressure the uterus is massaged have not been clearly reported. Similar to the pressure used with the cytobrush, it may be difficult to measure the pressure used while massaging the uterus in vivo. In one study, in 17% of the cases the operator was unable to recover uterine lavage fluid (Kasimanickam et al., 2005). In contrast, others have reported 100% successful sampling (Gilbert et al., 2005; Galvão et al., 2009). Kasimanickam et al. (2005) also reported a larger number of deformed cells recovered by uterine lavage in comparison with the cytobrush technique. Even though both techniques (cytobrush and uterine lavage) report the same outcome, PMN%, they are different methods. The cytobrush likely removes adhered endometrial and inflammatory cells, whereas uterine lavage may collect proportionally more cells that are free within the uterine lumen. Neither test is designed to evaluate the endometrium. The reported intra- and intercytology slide reader agreements suggest that the reading and scoring aspects of these techniques are robust (Table 2). Also, one study reported a high Se and Sp comparing uterine lavage with cytobrush as the reference test (Table 4). Unfortunately, no cytology slide reading protocols are described, which may influence comparison of test results between various studies.

Considering these limitations, the cytobrush, uterine lavage, and biopsy technique are not perfect diagnostic methods. Therefore, when reference tests, such as

PMN% determined by cytobrush or uterine lavage, and histopathological results by biopsy are used and errors of these tests are ignored, bias in the evaluation of the accuracy of test under evaluation is likely. Also, it is not surprising that the overall level of agreement between tests was only fair to moderate given the basis of the tests is different (Table 3). The low agreement in this case may be interpreted that one test is a poor surrogate for the other, likely because, despite previous assumptions, these techniques assess different aspects of reproductive-tract disease.

Test Validation Using Reproductive Outcomes

Although it could be argued that histopathological findings are the gold standard against which clinical tests should be assessed, validation of histopathological techniques either against other tests or against reproductive outcomes has been limited. For veterinary practitioners and producers, the use of reproductive outcomes either as a dichotomous outcome, e.g., pregnancy by an economically based target interval, or as a continuous variable (e.g., calving-to-conception interval) seems the preferred reference for cut-point analysis and validation of reproductive-tract diagnostic methods where no gold standard exists. Diagnostic tests for reproductive-tract diseases in production animals have limited utility without clinical effect or economic rationale. However, reproductive performance is not only influenced by reproductive-tract diseases. Confounders may have a negative effect on reproduction, e.g., poor heat detection, older cows, poor semen quality, diseases other than those of the reproductive tract, production systems, and nutrition. Thus, these confounders need to be taken into account when using biological outcomes from field trials to validate diagnostic tests.

Tests for Grossly Evident Purulent Material in the Vagina. Six papers compared reproductive performance in cows with PVD to unaffected ones. Impaired reproduction outcomes were described in 5 of the 6 papers diagnosing PVD by vaginoscopy (n = 5), Metricheck (n = 1), and transrectal palpation determining the diameter of the cervix (n = 1) mainly between 14 and 35 d, but ranging from 7 and 60 DIM.

Cows with PVD required more inseminations per pregnancy (LeBlanc et al., 2002; Gautam et al., 2009) and had a decreased first-service conception risk (LeBlanc et al., 2002; Runciman et al., 2009). A positive correlation has been reported between purulence score and time to conception (r = 0.22; $P < 0.05$; Studer and Morrow, 1978). A decrease in the proportion of cows pregnant by 6 wk after the mating start date (0.32 vs. 0.55; $P = 0.005$; Runciman et al., 2009) and an increase in time to conception of 119 to 151 d ($P = 0.001$; LeBlanc et al., 2002) and 120 to 325 d ($P < 0.001$; Gautam et al., 2009) was found in PVD-negative versus PVD-positive cows, respectively. Cows with PVD had an increased time to pregnancy (LeBlanc et al., 2002; Gautam et al., 2010) and decreased pregnancy by 210 DIM (Gautam et al., 2009). No unfavorable reproduction effect of PVD was reported by one study, where PVD was diagnosed with a gloved hand between 18 and 52 DIM in 243 pasture-based cows (Plöntzke et al., 2011). Although this is only one study conducted in 3 herds, it highlights that further validation of this commonly used diagnostic method may be required, particularly relative to the timing of examination to calving and to first insemination.

The Se and Sp for various diagnostic tests validated with reproductive outcomes were reported in 6 papers and are described in Table 5. The reproductive outcome in all studies was the pregnancy status, but this was assessed at different DIM. Often it was unclear whether pregnancy or nonpregnancy was used as the reference outcome. Only moderate Se (61 and 65%) and Sp (63 and 61%) are reported for vaginoscopy and Metricheck, respectively, in seasonal calving herds when using reproductive outcomes as the reference (Table 5). For example, the false-positive and false-negative rates following classification of cows based on PVD were approximately 40% using nonpregnancy by 6 wk after the start of breeding season as the outcome variable. In contrast, poor Se (15 and 18%) and high Sp (90 and 92%) were found in nonseasonal herds for Metricheck at 2 different time points (Table 5). The combination of measuring cervical diameter by transrectal palpation and vaginoscopy had a poor Se (20%); that is, only a few cows that were detected positive by either transrectal palpation or vaginoscopy were not pregnant by 120 DIM, and a high Sp (88%); many test negative cows were pregnant by 120 DIM (Table 5).

Depending on the management system (i.e., seasonal and nonseasonal calving systems), different levels of Se and Sp may be optimal. In seasonally managed herds, cows not detected pregnant at the end of the breeding period will likely be culled. Therefore, using a diagnostic test with a high Se may be more important in seasonal herds than in nonseasonal herds because time to intervene is limited. Hence, accepting a high false-positive rate may be more cost effective than a high Sp. A diagnostic method with a low Se (for example, defining a high Metricheck score threshold as test positive) will result in fewer cows being test positive and more truly diseased cows being incorrectly defined as test negative (i.e., this cut point increases the proportion of false-negative results). These cows will not have the benefit of intervention and may be at higher risk of not conceiving by the end of the breeding program

Table 5. Sensitivity (Se) and specificity (Sp) of tests for reproductive-tract disease with reproductive outcomes as the reference outcome reported in papers included in a systematic review (n = 6)

Diagnostic method	Outcome	DIM	Se	Sp	Reference
Vaginoscopy (≥mucopurulent)	Pregnant by 6 wk after MSD[1]	7 to 28	61	63	Runciman et al., 2009
Metricheck (≥mucopurulent)	Pregnant by 6 wk after MSD	7 to 28	65	61	Runciman et al., 2009
Metricheck (≥mucopurulent)	Pregnant by 120 DIM	35 ± 3	18	90	Dubuc et al., 2010a
Metricheck (≥mucopurulent)	Pregnant by 120 DIM	56 ± 3	15	92	Dubuc et al., 2010a
Vaginoscopy and cervix diameter[2]	Nonpregnancy beyond 120 DIM	20 to 33	20	88	LeBlanc et al., 2002
Cytobrush (>6.7% PMN[3])	Pregnant by 90 DIM	21 to 47	86	42	Couto et al., 2013
Cytobrush (>6% PMN)	Pregnant by 120 DIM	35 ± 3	24	86	Dubuc et al., 2010a
Cytobrush (>4% PMN)	Pregnant by 120 DIM	56 ± 3	17	91	Dubuc et al., 2010a
Cytobrush (>18% PMN)	Pregnant by 132 DIM	20 to 33	36	94	Kasimanickam et al., 2004
Uterine lavage (>10% PMN)	Pregnant by 210 DIM	40 to 60	79	43	Cheong et al., 2011

[1]MSD: mating start date; start of breeding season.

[2]Vaginoscopy and cervix diameter: cut points are mucopurulent discharge at >26 DIM diagnosed by vaginoscopy and cervix diameter >7.5 cm at >20 DIM diagnosed by transrectal palpation.

[3]PMN: polymorphonuclear leukocytes.

and thus being culled. In a nonseasonal herd the same cow may get more time to conceive; hence, the lack of Se may be less critical. The optimal cut point for tests thus depends on the cost of the test, the Se and Sp of that test, the economic effect of false-negative and false-positive results, and the cost and efficacy of the treatments.

Cytological and Histopathological Testing Methods. Twelve papers reporting on diagnostic methods measuring inflammatory response by cytology or histopathology compared with reproductive outcomes were included in this review. Diagnosis was made by biopsy, cytobrush, uterine lavage, and leukocyte esterase test in 1, 6, 4, and 2 papers, respectively.

Of 6 papers diagnosing reproductive-tract inflammation by cytobrush, 5 reported impaired reproduction outcomes associated with increased proportions of PMN. Of these papers, one included cervical and uterine inflammation in the reproductive analysis (Deguillaume et al., 2012), whereas the other papers diagnosed uterine inflammation only (Kasimanickam et al., 2004; McDougall et al., 2011; Senosy et al., 2012; Couto et al., 2013). Cows were examined between 20 and 49 DIM. Cut points for inflammation were between 5 and 9% in 4 papers, whereas 1 paper derived higher cut points (i.e., 10 and 18%) between 20 to 33 and 34 to 47 DIM (Kasimanickam et al., 2004).

Between 7 and 19 percentage-point reductions in first-service conception risk are described in cows diagnosed with endometritis by cytobrush (Kasimanickam et al., 2004; Senosy et al., 2012). Cows diagnosed with endometritis took longer to conceive from start of breeding (13 to 23 d) and from calving (29 to 62 d) compared with unaffected cows (Kasimanickam et al., 2004; McDougall et al., 2011). A decrease in proportion pregnant by the end of the breeding season was described in cows with endometritis diagnosed by cytobrush at 28 and 42 DIM compared with those unaffected at these days (McDougall et al., 2011). Also, in nonseasonal systems, reduced pregnancy rates were reported (Kasimanickam et al., 2004; Deguillaume et al., 2012; Couto et al., 2013). In contrast, in one study in pasture-based herds, no differences in reproductive performance were found between cows affected with endometritis diagnosed by cytobrush compared with those unaffected (Plöntzke et al., 2010).

All 4 studies on endometritis diagnosed by uterine lavage reported unfavorable reproduction outcomes. Cows were examined between 21 and 60 DIM. Cut points for inflammation were between 4 and 10% PMN (Gilbert et al., 2005; Galvão et al., 2009; Bacha and Regassa, 2010; Cheong et al., 2011). A 25 percentage-point reduction in first-service conception risk was observed in cows with endometritis diagnosed by uterine lavage (Gilbert et al., 2005; Bacha and Regassa, 2010). Also, affected cows took 30 to 88 d longer to conceive (Gilbert et al., 2005; Galvão et al., 2009; Cheong et al., 2011). Cows with endometritis were less likely to be pregnant by 180 and 300 DIM (Gilbert et al., 2005; Bacha and Regassa, 2010).

Couto et al. (2013) found no associations between a leukocyte esterase test performed on cytobrush samples taken from the cervix and the uterus and reproductive outcomes over a range of cut points. The testing procedure involved suspending the cytobrush in a small volume of saline. Subsequently, a leukocyte esterase test strip was dipped in the saline. It is unclear what the dilution effect may have been using this method. In contrast, multiparous cows that were test positive (≥3+) following testing of uterine lavage fluid with leukocyte esterase test took 39 d longer to conceive compared with cows below this cut point (Cheong et al., 2012). Further validation of the leukocyte esterase test is required.

Complications with Cytological and Histopathological Testing Methods. The use of different cut points and diagnostic techniques makes it challenging to compare the association between PMN% and reproductive outcomes. Moreover, no consistent reproductive outcome variable is reported, partially because of the use of different outcomes in different management systems. It is also unclear when an outcome variable is not reported whether no difference was found, and therefore was assumed not to be a valuable result, or if that the variable was not analyzed. This problem exists due to lack of clarity in reporting study methods. Additionally, several studies have excluded cows diagnosed with PVD. Using methods to diagnose 2 potentially different diseases (Dubuc et al., 2010a,b), this effectively is testing cows in series (i.e., performing diagnostic tests one after the other). Therefore, measurements of reproductive outcomes will be biased by excluding cows with PVD. Future validation studies on diagnosing endometritis using only cytobrush or uterine lavage should include all cows but specifically report on those affected with PVD. In contrast, Se and Sp of identifying cows with reproductive-tract disease may increase by combining tests (Barlund et al., 2008; Dubuc et al., 2010a); however, limited data are available on this. The optimal test strategy in clinical practice may be to use a rapid, inexpensive method with high Se for whole-herd testing, followed by testing disease-positive cows with a more specific test.

The only study that did not find any effect of increased endometrial PMN on reproductive outcomes (Plöntzke et al., 2010) excluded cows with PVD, and the chosen cut point (5% PMN) at 18 to 38 DIM may be too low. However, studies that did find a difference in some but not all reproductive outcomes used similar cut points at similar sampling times (8.5, 5, and 5% PMN diagnosed at 3, 4, and 5 wk postpartum, respectively; Galvão et al., 2009; Bacha and Regassa, 2010; Senosy et al., 2012). Of these studies, Galvão et al. (2009) included cows with PVD and found examination at 35 and 49 DIM using 6.5 and 4.0% cut points, respectively, to be predictive for reproductive failure. Another study that excluded cows with PVD did not find a cut point for PMN% that affected pregnancy by 90 DIM, when cows were examined between 21 and 31 DIM, whereas those examined between 32 and 47 DIM with >6.7% PMN had a decreased pregnancy rate (Couto et al., 2013).

Variation exists in the reported Se and Sp of the cytobrush and uterine lavage techniques assessed against the proportion of cows pregnant at given times. Poor (<36%) Se and good (>86%) Sp for predicting pregnancy by 120 and 132 DIM were found for cytobrush results at 3 different cut and time points. Predicting pregnancy by 90 DIM by cytobrush or by 210 DIM by uterine lavage had reasonable Se (79 and 86%) but poor Sp (42 and 43%; Table 5). Only one study calculated Se and Sp of uterine lavage with pregnancy status at 210 DIM as the reference outcome, whereas 4 studies measured Se and Sp for the cytobrush technique with similar reference outcomes at different DIM. Also, different cut points for endometritis were used. With the available data, it is not possible to determine which of these 2 diagnostic tests is a better predictor of reproductive performance. However, the present data indicate that >5% PMN in the uterus after 4 wk postpartum was associated with worse reproductive performance. Before this time point many cows may be included with physiological inflammation that might be associated with the process of postpartum uterine involution.

Bacteriological Tests. Uterine bacterial growth was compared with reproduction performance in 2 papers included in this systematic review. Isolation of *Trueperella pyogenes*, coliforms, or streptococci from the uterus increased the number of services required per conception (3.53, 3.45, and 3.36, respectively) in comparison with no bacterial growth (2.14; $P < 0.05$; Studer and Morrow, 1978). Conversely, bacterial infection of the uterus with *T. pyogenes*, *Fusobacterium necrophorum*, *Prevotella melaninogenica*, *Proteus* spp., and *Escherichia coli* did not influence reproductive performance in a study conducted 3 decades later (McDougall et al., 2011). However, technical difficulties with bacteriology of uterine samples (i.e., presence of multiple bacterial species including aerobes and anaerobes as well as gram-positive and gram-negative isolates) and the generally small number of cases assessed mean that associations may be missed. Recent development of (meta)genomic tests that allow multiple bacterial species to be detected without the cost and time associated with culture may allow re-assessment of these relationships (Santos and Bicalho, 2012). Additionally, studies not selected for this systematic review have reported decreased reproductive performance in cows with uterine infection with *T. pyogenes* (Bonnett et al., 1993; Huszenicza et al., 1999). In contrast, others did not find these associations and only described decreased reproductive performance in cows infected with *E. coli* possessing certain virulence factors (Bicalho et al., 2010; Bicalho et al., 2012). Further investigations on bacteriological tests and the association between bacterial isolation and reproductive failure are needed.

Practical Applications of Diagnostic Tests

The utility of a test will depend on the purpose of the test, e.g., if the test is being used in a research

or clinical context, and if an effective therapy is available. Treatment efficacy will also influence the time of examination. The use of tests that require penetration of the cervix are likely to have more limited application in the clinical environment because of the time and skill required and the requirement for laboratory support for subsequent testing. In the search for new diagnostic methods, a cow-side test, e.g., on milk or blood, using systemic and reproductive-specific biomarkers appear attractive options for further investigation. Costs, sampling time, on-farm convenience of the diagnostic method, requirement for laboratory skills, laboratory costs, and time to report are important considerations to justify examination of reproductive-tract disease in clinical practice. For example, the Metricheck method is faster and easier than a vaginal speculum (McDougall et al., 2007; Runciman et al., 2009). Unfortunately, no data are available on the economics of various diagnostic methods. However, it is likely that some reduction of Se and Sp in the clinical environment is acceptable to reduce the costs and test-result turn-around time.

CONCLUSIONS

Various reproductive-tract diseases and diagnostic methods have been described in the literature. However, the quality of reporting of disease definitions, validation, and diagnostic methods is inconsistent and generally low. The majority of the papers reviewed did not contain enough information to thoroughly assess the validity of the tests used. Hence future authors are encouraged to use a checklist for quality of design and reporting as a guide to improve the clarity, completeness, and utility of their manuscripts. Based on the evaluated literature, vaginoscopy or Metricheck are likely to remain the preferred cow-side diagnostic methods for detecting reproductive-tract disease in the clinical environment. However, further studies are required that meet the criteria of high-quality research. The ideal time for diagnostic examination and which cut point to use may depend on management system (e.g., seasonal and nonseasonal calving systems). Cytological assessment of endometritis presently may be better suited to research, where economical and time factors may be less stringent. However, development and validation of simple, inexpensive point-of-care tests would be highly desirable. Between 35 and 40 DIM, uterine cytology with >5% PMN cut point is generally associated with impaired reproductive outcomes. It was outside the scope of this review to perform a meta-analysis on cut points and time of examination, but as more data become available, that may be helpful. No gold standard test is available for reproductive-tract disease. Therefore, the use of reproductive outcomes in clinical trials is the most logical way to validate tests. To improve comparisons between studies, authors are encouraged to report more reproductive outcomes including those without significant differences. Use of newer statistical techniques, such as the Bayesian approach for TAGS, may be a potential path to improve the understanding of the validity of current and future tests. Furthermore, more data on intra- and inter-observer agreement are needed to determine the precision and sources of variability of the evaluated diagnostic methods. Additionally, further work is needed to more clearly optimize the timing of diagnosis relative to calving and to breeding, and to establish diagnostic cut points and criteria in this context. Cut points may vary between different management systems (e.g., housing conditions and milk yield), and approaches that define the optimal Se and Sp in the different production systems are topics for future research. To be able to improve uterine health and reproductive performance in dairy cows, a better understanding of diagnostic methods is required. Such progress will be aided by rigorous, comprehensive, clear, and consistent reporting of study methods and outcomes.

ACKNOWLEDGMENTS

This systematic review was performed as part of the PhD program of the first author and was partly funded by New Zealand dairy farmers through DairyNZ (AN808) and Cognosco, Anexa Animal Health, New Zealand. The meeting of the "International consortium on inflammation and immunity in the bovine reproductive system" facilitated by Martin Sheldon during the International Congress on Animal Reproduction (ICAR) meeting in Vancouver in August 2012 is acknowledged, where a review on diagnostic methods of reproductive-tract diseases was instigated.

REFERENCES

Arlt, S., V. Dicty, and W. Heuwieser. 2010. Evidence-based medicine in canine reproduction: Quality of current available literature. Reprod. Domest. Anim. 45:1052–1058.

Bacha, B., and F. G. Regassa. 2010. Subclinical endometritis in Zebu × Friesian crossbred dairy cows: Its risk factors, association with subclinical mastitis and effect on reproductive performance. Trop. Anim. Health Prod. 42:397–403.

Barlund, C. S., T. D. Carruthers, C. L. Waldner, and C. W. Palmer. 2008. A comparison of diagnostic techniques for postpartum endometritis in dairy cattle. Theriogenology 69:714–723.

Benos, D. J., E. Bashari, J. M. Chaves, A. Gaggar, N. Kapoor, M. LaFrance, R. Mans, D. Mayhew, S. McGowan, A. Polter, Y. Qadri, S. Sarfare, K. Schultz, R. Splittgerber, J. Stephenson, C. Tower, R. G. Walton, and A. Zotov. 2007. The ups and downs of peer review. Adv. Physiol. Educ. 31:145–152.

Bicalho, M. L. S., V. S. Machado, G. Oikonomou, R. O. Gilbert, and R. C. Bicalho. 2012. Association between virulence factors of *Escherichia coli*, *Fusobacterium necrophorum*, and *Arcanobacte-*

rium pyogenes and uterine diseases of dairy cows. Vet. Microbiol. 157:125–131.

Bicalho, R. C., V. S. Machado, M. L. S. Bicalho, R. O. Gilbert, A. G. V. Teixeira, L. S. Caixeta, and R. V. V. Pereira. 2010. Molecular and epidemiological characterization of bovine intrauterine *Escherichia coli*. J. Dairy Sci. 93:5818–5830.

Bonnett, B. N., S. W. Martin, and A. H. Meek. 1993. Associations of clinical findings, bacteriological and histological results of endometrial biopsy with reproductive performance of postpartum dairy cows. Prev. Vet. Med. 15:205–220.

Bonnett, B. N., R. B. Miller, W. G. Etherington, S. W. Martin, and W. H. Johnson. 1991. Endometrial biopsy in Holstein-Friesian dairy cows. 1. Technique, histological criteria and results. Can. J. Vet. Res. 55:155–161.

Bossuyt, P. M., J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. M. Irwig, D. Moher, D. Rennie, H. C. W. de Vet, and J. G. Lijmer. 2003. The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. Clin. Chem. 49:7–18.

Centre for Reviews and Dissemination. 2009. Systematic reviews—CRD's guidance for undertaking reviews in health care. University of York. www.york.ac.uk/inst/crd/pdf/Systematic_Reviews.pdf.

Cheong, S. H., D. V. Nydam, K. N. Galvao, B. M. Crosier, and R. O. Gilbert. 2011. Cow-level and herd-level risk factors for subclinical endometritis in lactating Holstein cows. J. Dairy Sci. 94:762–770.

Cheong, S. H., D. V. Nydam, K. N. Galvao, B. M. Crosier, A. Ricci, L. S. Caixeta, R. B. Sper, M. Fraga, and R. O. Gilbert. 2012. Use of reagent test strips for diagnosis of endometritis in dairy cows. Theriogenology 77:858–864.

Choi, Y. K., W. O. Johnson, M. T. Collins, and I. A. Gardner. 2006. Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard. J. Agric. Biol. Environ. Stat. 11:210–229.

Couto, G. B., D. H. Vaillancourt, and R. C. Lefebvre. 2013. Comparison of a leukocyte esterase test with endometrial cytology for diagnosis of subclinical endometritis in postpartum dairy cows. Theriogenology 79:103–107.

Deguillaume, L., A. Geffre, L. Desquilbet, A. Dizien, S. Thoumire, C. Vorniere, F. Constant, R. Fournier, and S. Chastant-Maillard. 2012. Effect of endocervical inflammation on days to conception in dairy cows. J. Dairy Sci. 95:1776–1783.

Dohoo, I., W. Martin, and H. Stryhn. 2009. Veterinary Epidemiologic Research. 2nd ed. AVC Inc., Charlottetown, Prince Edward Island, Canada.

Dubuc, J., T. F. Duffield, K. E. Leslie, J. S. Walton, and S. J. LeBlanc. 2010a. Definitions and diagnosis of postpartum endometritis in dairy cows. J. Dairy Sci. 93:5225–5233.

Dubuc, J., T. F. Duffield, K. E. Leslie, J. S. Walton, and S. J. LeBlanc. 2010b. Risk factors for postpartum uterine diseases in dairy cows. J. Dairy Sci. 93:5764–5771.

Enøe, C., M. P. Georgiadis, and W. O. Johnson. 2000. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Prev. Vet. Med. 45:61–81.

Fischer, C., M. Drillich, S. Odau, W. Heuwieser, R. Einspanier, and C. Gabler. 2010. Selected pro-inflammatory factor transcripts in bovine endometrial epithelial cells are regulated during the oestrous cycle and elevated in case of subclinical or clinical endometritis. Reprod. Fertil. Dev. 22:818–829.

Földi, J., M. Kulcsar, A. Pecsi, B. Huyghe, C. de Sa, J. Lohuis, P. Cox, and G. Huszenicza. 2006. Bacterial complications of postpartum uterine involution in cattle. Anim. Reprod. Sci. 96:265–281.

Fontela, P. S., N. P. Pai, I. Schiller, N. Dendukuri, A. Ramsay, and M. Pai. 2009. Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: Evaluation using QUADAS and STARD standards. PLoS ONE 4:e7753.

Galvão, K. N., M. Frajblat, S. B. Brittin, W. R. Butler, C. L. Guard, and R. O. Gilbert. 2009. Effect of prostaglandin $F_{2\alpha}$ on subclinical endometritis and fertility in dairy cows. J. Dairy Sci. 92:4906–4913.

Galvão, K. N., N. R. Santos, J. S. Galvão, and R. O. Gilbert. 2011. Association between endometritis and endometrial cytokine expression in postpartum Holstein cows. Theriogenology 76:290–299.

Gardner, I. A., and M. Greiner. 2006. Receiver-operating characteristic curves and likelihood ratios: Improvements over traditional methods for the evaluation and application of veterinary clinical pathology tests. Vet. Clin. Pathol. 35:8–17.

Gardner, I. A., S. S. Nielsen, R. J. Whittington, M. T. Collins, D. Bakker, B. Harris, S. Sreevatsan, J. E. Lombard, R. Sweeney, D. R. Smith, J. Gavalchin, and S. Eda. 2011. Consensus-based reporting standards for diagnostic test accuracy studies for paratuberculosis in ruminants. Prev. Vet. Med. 101:18–34.

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

Gautam, G., T. Nakao, M. Yusuf, and K. Koike. 2009. Prevalence of endometritis during the postpartum period and its impact on subsequent reproductive performance in two Japanese dairy herds. Anim. Reprod. Sci. 116:175–187.

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

Goodman, S. N., J. Berlin, S. W. Fletcher, and R. H. Fletcher. 1994. Manuscript quality before and after peer-review and editing at Annals of Internal Medicine. Ann. Intern. Med. 121:11–21.

Green, M. P., A. M. Ledgard, M. C. Berg, A. J. Peterson, and P. J. Back. 2009. Prevalence and identification of systemic markers of sub-clinical endometritis in postpartum dairy cows. Proc. New Zeal. Soc. Anim. 69:37–42.

Greiner, M., and I. A. Gardner. 2000a. Application of diagnostic tests in veterinary epidemiologic studies. Prev. Vet. Med. 45:43–59.

Greiner, M., and I. A. Gardner. 2000b. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev. Vet. Med. 45:3–22.

Grindlay, D. J. C., M. L. Brennan, and R. S. Dean. 2012. Searching the veterinary literature: A comparison of the coverage of veterinary journals by nine bibliographic databases. J. Vet. Med. Educ. 39:404–412.

Haimerl, P., S. Arlt, and W. Heuwieser. 2012. Evidence-based medicine: Quality and comparability of clinical trials investigating the efficacy of prostaglandin $F_{2\alpha}$ for the treatment of bovine endometritis. J. Dairy Res. 79:287–296.

Higgins, J., and S. Green, ed. 2011. Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. The Cochrane collaboration. www.cochrane-handbook.org.

Huszenicza, G., M. Fodor, M. Gacs, M. Kulcsar, M. J. W. Dohmen, M. Vamos, L. Porkolab, T. Kegl, J. Bartyik, J. Lohuis, S. Janosi, and G. Szita. 1999. Uterine bacteriology, resumption of cyclic ovarian activity and fertility in postpartum cows kept in large-scale dairy herds. Reprod. Domest. Anim. 34:237–245.

Ishikawa, Y., K. Nakada, K. Hagiwara, R. Kirisawa, H. Iwai, M. Moriyoshi, and Y. Sawamukai. 2004. Changes in interleukin-6 concentration in peripheral blood of pre- and post-partum dairy cattle and its relationship to postpartum reproductive diseases. J. Vet. Med. Sci. 66:1403–1408.

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

Kasimanickam, R., T. F. Duffield, R. A. Foster, C. J. Gartley, K. E. Leslie, J. S. Walton, and W. H. Johnson. 2005. A comparison of the cytobrush and uterine lavage techniques to evaluate endometrial cytology in clinically normal postpartum dairy cows. Can. Vet. J. 46:255–259.

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

Kim, I. H., K. J. Na, and M. P. Yang. 2005. Immune responses during the peripartum period in dairy cows with postpartum endometritis. J. Reprod. Dev. 51:757–764.

LeBlanc, S. J. 2008. Postpartum uterine disease and dairy herd reproductive performance: A review. Vet. J. 176:102–114.

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

Leutert, C., X. von Krueger, J. Plöntzke, and W. Heuwieser. 2012. Evaluation of vaginoscopy for the diagnosis of clinical endometritis in dairy cows. J. Dairy Sci. 95:206–212.

Machado, V. S., W. A. Knauer, M. L. S. Bicalho, G. Oikonomou, R. O. Gilbert, and R. C. Bicalho. 2012. A novel diagnostic technique to determine uterine health of Holstein cows at 35 days postpartum. J. Dairy Sci. 95:1349–1357.

McDougall, S., H. Hussein, D. Aberdein, K. Buckle, J. Roche, C. Burke, M. Mitchell, and S. Meier. 2011. Relationships between cytology, bacteriology and vaginal discharge scores and reproductive performance in dairy cattle. Theriogenology 76:229–240.

McDougall, S., R. Macaulay, and C. Compton. 2007. Association between endometritis diagnosis using a novel intravaginal device and reproductive performance in dairy cattle. Anim. Reprod. Sci. 99:9–23.

Meira, E. B. S., L. C. S. Henriques, L. R. M. Sa, and L. Gregory. 2012. Comparison of ultrasonography and histopathology for the diagnosis of endometritis in Holstein-Friesian cows. J. Dairy Sci. 95:6969–6973.

Moher, D., A. Jones, L. Lepage, and C. Grp. 2001. Use of the CONSORT statement and quality of reports of randomized trials—A comparative before-and-after evaluation. JAMA 285:1992–1995.

Nielsen, S. S., and N. Toft. 2008. Ante mortem diagnosis of paratuberculosis: A review of accuracies of ELISA, interferon-gamma assay and faecal culture techniques. Vet. Microbiol. 129:217–235.

Overbeck, W., K. Jager, H. A. Schoon, and T. S. Witte. 2013. Comparison of cytological and histological examinations in different locations of the equine uterus—An in vitro study. Theriogenology 79:1262–1268.

Peter, A. T., G. M. Jarratt, and D. W. Hanlon. 2011. Accuracy of diagnosis of clinical endometritis with Metricheck™ in postpartum dairy cows. Clin. Theriogenol. 3:461–465.

Plöntzke, J., L. V. Madoz, R. L. De la Sota, M. Drillich, and W. Heuwieser. 2010. Subclinical endometritis and its impact on reproductive performance in grazing dairy cattle in Argentina. Anim. Reprod. Sci. 122:52–57.

Plöntzke, J., L. V. Madoz, R. L. De la Sota, W. Heuwieser, and M. Drillich. 2011. Prevalence of clinical endometritis and its impact on reproductive performance in grazing dairy cattle in Argentina. Reprod. Domest. Anim. 46:520–526.

Pouillot, R., G. Gerbier, and I. A. Gardner. 2002. "TAGS", a program for the evaluation of test accuracy in the absence of a gold standard. Prev. Vet. Med. 53:67–81.

Purcell, G. P., S. L. Donovan, and F. Davidoff. 1998. Changes to manuscripts during the editorial process—Characterizing the evolution of a clinical paper. JAMA 280:227–228.

Roy, J.-P., and G. Keefe. 2012. Systematic review: What is the best antibiotic treatment for *Staphylococcus aureus* intramammary infection of lactating cows in North America? Vet. Clin. North Am. Food Anim. Pract. 28:39–50.

Runciman, D. J., G. A. Anderson, and J. Malmo. 2009. Comparison of two methods of detecting purulent vaginal discharge in postpartum dairy cows and effect of intrauterine cephapirin on reproductive performance. Aust. Vet. J. 87:369–378.

Sannmann, I., S. Arlt, and W. Heuwieser. 2012. A critical evaluation of diagnostic methods used to identify dairy cows with acute postpartum metritis in the current literature. J. Dairy Res. 79:436–444.

Santos, T. M. A., and R. C. Bicalho. 2012. Diversity and succession of bacterial communities in the uterine fluid of postpartum metritic, endometritic and healthy dairy cows. PLoS ONE 7:e53048.

Sargeant, J. M., A. M. O'Connor, I. A. Gardner, J. S. Dickson, M. E. Torrence, I. R. Dohoo, S. L. Lefebvre, P. S. Morley, A. Ramirez, and K. Snedeker. 2010. The REFLECT statement: Reporting guidelines for randomized controlled trials in livestock and food safety: Explanation and elaboration. J. Food Prot. 73:579–603.

Sargeant, J. M., A. Rajic, S. Read, and A. Ohlsson. 2006. The process of systematic review and its application in agri-food public-health. Prev. Vet. Med. 75:141–151.

Schulz, K. F., D. G. Altman, D. Moher, and C. Grp. 2010. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. J. Clin. Epidemiol. 63:834–840.

Seals, R. C., I. Matamoros, and G. S. Lewis. 2002. Relationship between postpartum changes in 13,14-dihydro-15-keto-PGF(2 alpha) concentrations in Holstein cows and their susceptibility to endometritis. J. Anim. Sci. 80:1068–1073.

Senosy, W., M. Uchiza, N. Tameoka, Y. Izaike, and T. Osawa. 2011. Impact of ovarian and uterine conditions on some diagnostic tests output of endometritis in postpartum high-yielding dairy cows. Reprod. Domest. Anim. 46:800–806.

Senosy, W. S., Y. Izaike, and T. Osawa. 2012. Influences of metabolic traits on subclinical endometritis at different intervals postpartum in high milking cows. Reprod. Domest. Anim. 47:666–674.

Senosy, W. S., M. Uchiza, N. Tameoka, Y. Izaike, and T. Osawa. 2009. Association between evaluation of the reproductive tract by various diagnostic tests and restoration of ovarian cyclicity in high-producing dairy cows. Theriogenology 72:1153–1162.

Sheldon, I. M., G. S. Lewis, S. LeBlanc, and R. O. Gilbert. 2006. Defining postpartum uterine disease in cattle. Theriogenology 65:1516–1530.

Siddiqui, M. A. R., A. Azuara-Blanco, and J. Burr. 2005. The quality of reporting of diagnostic accuracy studies published in ophthalmic journals. Br. J. Ophthalmol. 89:261–265.

Simoneit, C., W. Heuwieser, and S. P. Arlt. 2012. Inter-observer agreement on a checklist to evaluate scientific publications in the field of animal reproduction. J. Vet. Med. Educ. 39:119–127.

Smidt, N., A. W. S. Rutjes, D. van der Windt, R. Ostelo, P. M. Bossuyt, J. B. Reitsma, L. M. Bouter, and H. C. W. de Vet. 2006a. The quality of diagnostic accuracy studies since the STARD statement—Has it improved? Neurology 67:792–797.

Smidt, N., A. W. S. Rutjes, D. A. W. M. van der Windt, R. W. J. G. Ostelo, P. M. Bossuyt, J. B. Reitsma, L. M. Bouter, and H. C. w. de Vet. 2006b. Reproducibility of the STARD checklist: An instrument to assess the quality of reporting of diagnostic accuracy studies. BMC Med. Res. Methodol. 6:12.

Studer, E., and D. A. Morrow. 1978. Postpartum evaluation of bovine reproductive potential: Comparison of findings from genital tract examination per rectum, uterine culture, and endometrial biopsy. J. Am. Vet. Med. Assoc. 172:489–494.

Tranfield, D., D. Denyer, and P. Smart. 2003. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. Br. J. Manage. 14:207–222.

von Elm, E., D. G. Altman, M. Egger, S. J. Pocock, P. C. Gotzsche, J. P. Vandenbroucke, and S. Initiative. 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement guidelines for reporting observational studies. Epidemiology 18:800–804.

Westermann, S., M. Drillich, T. B. Kaufmann, L. V. Madoz, and W. Heuwieser. 2010. A clinical approach to determine false positive findings of clinical endometritis by vaginoscopy by the use of uterine bacteriology and cytology in dairy cows. Theriogenology 74:1248–1255.

Williams, E. J., D. P. Fischer, D. U. Pfeiffer, G. C. W. England, D. E. Noakes, H. Dobson, and I. M. Sheldon. 2005. Clinical evaluation of postpartum vaginal mucus reflects uterine bacterial infection and the immune response in cattle. Theriogenology 63:102–117.

Yavari, M., M. Haghkhah, M. R. Ahmadi, H. R. Gheisari, and S. Nazifi. 2009. Comparison of cervical and uterine cytology between different classification of postpartum endometritis and bacterial isolates in Holstein dairy cows. Int. J. Dairy Sci. 4:19–26.

Zafar, A., G. I. Khan, and M. A. R. Siddiqui. 2008. The quality of reporting of diagnostic accuracy studies in diabetic retinopathy screening: A systematic review. Clin. Experiment. Ophthalmol. 36:537–542.