University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Papers in Veterinary and Biomedical Science | Veterinary and Biomedical Sciences, Department of

1-1-2011

Consensus-based reporting standards for diagnostic test accuracy studies for paratuberculosis in ruminants

Ian A. Gardner
University of California - Davis

Søren S. Nielsen
University of Copenhagen

Richard J. Wittington
University of Sydney

Michael T. Collins
University of Wisconsin - Madison

Douwe Bakker
Central Veterinary Institute

See next page for additional authors

Follow this and additional works at: http://digitalcommons.unl.edu/vetscipapers

Gardner, Ian A.; Nielsen, Søren S.; Wittington, Richard J.; Collins, Michael T.; Bakker, Douwe; Harris, Beth; Sreevatsan, Srinand; Lombard, Jason E.; Sweeney, Raymond; Smith, David R.; Gavalchin, Jerrie; and Eda, Shigetoshi, "Consensus-based reporting standards for diagnostic test accuracy studies for paratuberculosis in ruminants" (2011). *Papers in Veterinary and Biomedical Science.* Paper 133.
http://digitalcommons.unl.edu/vetscipapers/133

This Article is brought to you for free and open access by the Veterinary and Biomedical Sciences, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Papers in Veterinary and Biomedical Science by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Authors
Ian A. Gardner, Søren S. Nielsen, Richard J. Wittington, Michael T. Collins, Douwe Bakker, Beth Harris, Srinand Sreevatsan, Jason E. Lombard, Raymond Sweeney, David R. Smith, Jerrie Gavalchin, and Shigetoshi Eda

This article is available at DigitalCommons@University of Nebraska - Lincoln: http://digitalcommons.unl.edu/vetscipapers/133

Preventive Veterinary Medicine 101 (2011) 18–34

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

Consensus-based reporting standards for diagnostic test accuracy studies for paratuberculosis in ruminants

Ian A. Gardner[a,*], Søren S. Nielsen[b], Richard J. Whittington[c], Michael T. Collins[d], Douwe Bakker[e], Beth Harris[f], Srinand Sreevatsan[g], Jason E. Lombard[h], Raymond Sweeney[i], David R. Smith[j], Jerrie Gavalchin[k], Shigetoshi Eda[l]

[a] Department of Medicine and Epidemiology, School of Veterinary Medicine, University of California, Davis, CA 95616, USA
[b] Department of Large Animal Sciences, Faculty of Life Sciences, University of Copenhagen, Grønnegardsvej 8, DK-1870 Frederiksberg C, Denmark
[c] Faculty of Veterinary Science, The University of Sydney, 425 Werombi Road, Camden 2570, NSW, Australia
[d] Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin, Madison, WI 53703-1102, USA
[e] Central Veterinary Institute, Edelhertweg 15, 8200 AB Lelystad, The Netherlands
[f] USDA, Animal and Plant Health Inspection Services (APHIS), Veterinary Services (VS), National Veterinary Services Laboratories, 1920 Dayton Avenue, Ames, IA 50010, USA
[g] Veterinary Population Medicine Department, College of Veterinary Medicine, University of Minnesota, 1971 Commonwealth Avenue, Saint Paul, MN 55108, USA
[h] USDA, Animal and Plant Health Inspection Service (APHIS), Veterinary Services (VS), Centers for Epidemiology and Animal Health, 2150 Centre Avenue, Bldg. B, Fort Collins, CO 80526-8117, USA
[i] Department of Clinical Studies-New Bolton Center, University of Pennsylvania School of Veterinary Medicine, 382 West Street Road, Kennett Square, PA 19348, USA
[j] School of Veterinary Medicine and Biomedical Sciences, University of Nebraska-Lincoln, PO Box 830905, Lincoln, NE 68583, USA
[k] Department of Animal Science, Cornell University, Ithaca, NY 14853, USA
[l] Center for Wildlife Health, Department of Forestry, Wildlife and Fisheries, 274 Ellington Plant Science, University of Tennessee Knoxville, Knoxville, TN 37996, USA

ARTICLE INFO

Article history:
Received 27 October 2010
Received in revised form 3 April 2011
Accepted 5 April 2011

Keywords:
Test accuracy
Sensitivity and specificity
Paratuberculosis
Reporting standards
STARD

ABSTRACT

The Standards for Reporting of Diagnostic Accuracy (STARD) statement (www.stard-statement.org) was developed to encourage complete and transparent reporting of key elements of test accuracy studies in human medicine. The statement was motivated by widespread evidence of bias in test accuracy studies and the finding that incomplete or absent reporting of items in the STARD checklist was associated with overly optimistic estimates of test performance characteristics. Although STARD principles apply broadly, specific guidelines do not exist to account for unique considerations in livestock studies such as herd tests, potential use of experimental challenge studies, a more diverse group of testing purposes and sampling designs, and the widespread lack of an ante-mortem reference standard with high sensitivity and specificity. The objective of the present study was to develop a modified version of STARD relevant to paratuberculosis (Johne's disease) in ruminants. Examples and elaborations for each of the 25 items were developed by a panel of experts using a consensus-based approach to explain the items and underlying concepts. The new guidelines, termed STRADAS-paraTB (Standards for Reporting of Animal Diagnostic Accuracy Studies for paratuberculosis), should facilitate improved quality of reporting of the design, conduct and results of paratuberculosis test accuracy studies which were identified as "poor" in a review published in 2008 in Veterinary Microbiology.
© 2011 Elsevier B.V. All rights reserved.

* Corresponding author at: Department of Medicine and Epidemiology, School of Veterinary Medicine, University of California, One Shields Avenue, Davis, CA 95616, USA. Tel.: +1 530 752 6992; fax: +1 530 752 0414.
E-mail address: iagardner@ucdavis.edu (I.A. Gardner).

0167-5877/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.prevetmed.2011.04.002

This article is a U.S. government work, and is not subject to copyright in the United States.

1. Introduction

The Standards for Reporting of Diagnostic Accuracy (STARD) statement (www.stard-statement.org) was published in 2003 in several leading medical journals with an overall goal of improving reporting of test accuracy studies in human medicine (Bossuyt et al., 2003a,b). The statement includes a checklist of 25 key items and a flow diagram that accounts for patient losses, including lack of reference standard testing during the diagnostic work-up. Although focused primarily on clinical patients in hospitals and referral clinics, the principles apply equally well to subclinical diseases regardless of whether they are infectious (Peeling et al., 2006; TDR Diagnostics Evaluation Expert Panel, 2006) or non-infectious. Absence or underreporting of STARD items has been associated with overly optimistic estimates of test accuracy (Lijmer et al., 1999).

Methodological aspects of studies, including their design and statistical analysis, are not explicitly mentioned in STARD and a separate 14-item tool, named QUADAS (Quality Assessment of Diagnostic Accuracy Studies) is used for such assessments (Whiting et al., 2003). Different study designs can be used for the same purpose and application of a test. In general, prospective designs likely are preferable to retrospective studies using repository samples because of improved data quality and completeness of information on covariates that influence variability in test accuracy. However, poor reporting may occur in both well-designed and flawed test accuracy studies.

A structured approach to the design of test evaluation studies for chronic infections in animals has recently been proposed (Nielsen et al., 2011) but the manuscript does not address reporting guidelines. In a prior study, one of us proposed that a modification of STARD checklist items was necessary to increase applicability to and perhaps adoption of its use in animal studies (Gardner, 2010). Suggested changes included modification of terminology to be consistent with the REFLECT (Reporting Guidelines for Randomized Control Trials) statement for trials in livestock and food safety (O'Connor et al., 2010), inclusion of herd tests and other testing purposes, potential use of challenge studies, an expanded concept of clinical utility of findings, and more variable sampling designs as examples.

Over the last 2 decades, there has been increased interest in developing novel tests or improving existing tests for detection of paratuberculosis (Johne's disease), caused by infection with *Mycobacterium avium* subsp. *paratuberculosis* (MAP). An important goal of much of the research has been to identify tests of greater sensitivity and specificity that detect infection at an earlier stage, thereby minimizing production losses. Nielsen and Toft (2008) reviewed 85 test validation studies for paratuberculosis that mostly involved serum ELISA tests in cattle and identified multiple design and reporting flaws. Many of the published studies were considered by the authors to be of poor quality.

This paper describes a consensus-based reporting list of items, based on a modification of STARD, for field-based test accuracy studies for individual and herd classification of MAP infection status. The new guidelines are termed STRADAS-paraTB (Standards for Reporting of Animal Diagnostic Accuracy Studies for paratuberculosis). As a prerequisite for designation of appropriate checklist items, we assumed that the test under evaluation (TUE) had been optimized in laboratory experiments and preliminary estimates of sensitivity, specificity, repeatability and reproducibility were available using a limited number of well-characterized samples. In addition, we assumed that results of the validation to this level, which corresponds to Stage I in the World Organisation for Animal Health (OIE) pathway (OIE, 2010), had been published in a peer-reviewed journal.

2. Methods and processes

The STARD statement was developed by a 9-person steering committee and a group of invited experts (Bossuyt et al., 2003a). A similar approach was used for the STRADAS-paraTB initiative. Two preliminary face-to-face meetings were held in Minnesota in 2008 (4 co-authors) and 2009 (7 co-authors) to decide if modification of STARD was necessary to increase relevance to paratuberculosis test accuracy studies, to define changes if modification was necessary, and to develop a plan to obtain additional scientific input prior to publication.

As a sequel to these initial efforts, a meeting was held in Orlando in March 2010 with a goal of finalizing the checklist and identifying examples and elaborations. The final list and manuscript was compiled by 12 participants (9 from USA and 1 each from Australia, Denmark and the Netherlands) who attended at least one of the meetings. The expertise of the contributors was broadly classified as basic scientist/researcher, epidemiologist, and clinician/diagnostician and several participants were associate editors or editorial board members of journals. Nine of 12 participants had authored or co-authored a manuscript on paratuberculosis test accuracy and 8 had authored a paper that was evaluated by Nielsen and Toft (2008). At the Orlando meeting, consensus about changes in the original STARD items was deemed to have occurred when at least 80% of participants voted for a suggested change or for no change in the wording of the item. A final opportunity to review and approve the wording changes was made approximately one month prior to submission of this manuscript for publication and during manuscript revision.

3. Examples, explanations and elaborations

Of the original 25 STARD checklist items, we modified 18 (2– 6, 8–11, 13, 15–18, and 22–25) to reflect input of the experts (Table 1). Relevant examples and explanations for each item were based on expert input where each co-author contributed based on their experience and knowledge of published studies. In the examples, we use square brackets when the original text was changed to improve clarity e.g. spelling out of acronyms and insertion of additional words. For 3 items (14, 17 and 20), we only provide an explanation because the STARD example is sufficient (Bossuyt et al., 2003b). Definitions and terms used in the manuscript are in Appendix 1.

Item:1 Identify the article as a study of diagnostic accuracy (recommend MeSH heading of "sensitivity and specificity").

Table 1
Checklist of items for reporting of diagnostic test accuracy studies for paratuberculosis in ruminants based on the STARD checklist (www.stard-statement.org).

Section and topic	Item	Description of item	On page
Title/abstract/keywords	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	
Introduction	2	State the research question or study aims such as estimation of diagnostic accuracy or comparison of accuracy between tests ***in a specified matrix (specimen type) for a defined purpose at the animal or herd level***	
Materials and Methods			
Animals and herds	3	***Describe study sampling frame: Describe the source population*** and inclusion and exclusion criteria, setting and locations where data were collected ***for all relevant levels of the study sample (animals and herds)***	
	4	***Describe selection of animals and herds: Describe sample selection methods (random, convenience,*** **etc.)** ***within each level of the sampling hierarchy (*****e.g.*** ***regions, farms, barns, cows) including exclusion criteria and number of study animals and herds.***	
	5	***Describe sampling protocol: Describe the collection, specimen size, transportation, handling and storage of specimens prior to the performance of the test under evaluation (TUE) and the reference standard***.	
	6	Describe ***study design***: Was data collection planned before the TUE and reference standard were performed (prospective study) or after (retrospective study)?	
Test methods	7	Describe the reference standard and its rationale.	
	8	Describe technical specifications of materials and methods involved including how and when measurements were taken, and/or cite references for TUE and reference standards. ***Specify quality control samples for TUE and reference standard and specimen/analytical unit size of tested samples.***	
	9	Describe the ***outcome measure and rationale for the cutoffs*** and/or categories of the results of the TUE and reference standard.	
	10	Describe the ***name, location, and qualifications of the laboratory***, including the number, training and expertise of persons executing the TUE and reference standard.	
	11	Describe whether or not the readers of the TUE and reference standard were blind (masked) to the results of the other test and describe ***any individual or herd-level information available to the readers.***	
Statistical methods	12	Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	
	13	Describe methods for calculating test ***repeatability and*** reproducibility, if done.	
Results			
Animals and herds	14	Report when study was done, including beginning and end dates of recruitment	
	15	***Report demographic and other biologically relevant characteristics of the study sample at the individual (e.g. age, sex, breed, and risk factors) and at the herd levels (*****e.g.*** ***production system).***	
	16	Report the ***number of animals and herds*** satisfying the criteria for inclusion that did or did not undergo the TUE and/or the reference standard: describe ***why animals and herds*** failed to receive either test.	
Test results	17	Report time interval ***between collection of samples for the TUE and the reference standard, and interventions*** administered between.	
	18	Report distribution of severity of disease or stage of infection (define criteria), ***and other relevant diagnoses or treatments in animals in the study sample.***	
	19	Report a cross tabulation of the results of the TUE (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	
	20	Report any adverse events from performing the TUE or the reference standard.	
Estimates	21	Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	
	22	Report how indeterminate results, missing responses and outlier values ***of the TUE and the reference standard*** were handled. If additional testing of animals and herds is done to resolve discrepant results, then describe the rationale and approach (a flow diagram is strongly recommended).	
	23	Report estimates of variability of diagnostic accuracy between ***relevant subpopulations***, readers, ***or testing sites***, if done.	
	24	Report estimates of test ***repeatability and*** reproducibility, if done.	
Discussion	25	***Discuss the utility of the TUE in various settings (clinical, research, surveillance*** **etc.)** ***in the context of the currently available tests.***	

Modifications of text from the STARD checklist are in italics and bold. TUE (test under evaluation) is used instead of index test as used in STARD.

Example (Objective statement from a structured abstract)

To estimate the sensitivity (Se) and specificity (Sp) for an enhanced direct-fecal PCR procedure, bacterial culture of feces (BCF), and a serum ELISA for detecting *Mycobacterium avium* subsp. *paratuberculosis* (MAP) infection in adult dairy cattle (Scott et al., 2007).

Explanation

Electronic retrieval of relevant manuscripts is a key component of primary diagnostic test accuracy research and is necessary for systematic literature reviews. The Medical Subject Headings (MeSH: www.ncbi.nlm.nih.gov/mesh) includes the terms paratuberculosis, Johne's disease and *Mycobacterium avium* subsp. *paratuberculosis* and lists 4 possible headings or subheadings relevant to diagnostic test evaluation: (1) sensitivity and specificity; (2) predictive value of tests; (3) ROC curves; and (4) limit of detection. The term "diagnostic accuracy" is not a MeSH term. Hence, the terms "sensitivity and specificity" should be included in at least

one of the 3 designated locations (title/abstract/keywords) as is done in the structured abstract in the example. The terms "herd", "flock" or "aggregate" are not listed in MeSH. PubMed searches for "herd OR sensitivity" yield similar numbers of manuscripts to those obtained when the only search term is "sensitivity".

Item:2 State the research question or study aims such as estimation of diagnostic accuracy or comparison of accuracy between tests in a specified matrix (specimen type) for a defined purpose at the animal or herd level.

Example

The aim of this study was to develop and evaluate a method for culturing of fecal samples pooled from a number of sheep in order to provide an economical test for *M. avium* subsp. *paratuberculosis* in flocks. Specific aims were to determine an acceptable rate of pooling of fecal samples, to compare the sensitivities of pooled fecal culture and an AGID [agar gel immunodiffusion] test, to evaluate the practicality of sample collection, and to develop recommendations for sampling rates for confirmation of *M. avium* subsp. *paratuberculosis* infection in flocks (Whittington et al., 2000).

Explanation

A clearly defined research objective relative to the TUE enables the reader to determine the validity of the test evaluation study in the context of the purpose of testing. The OIE endorses the concept of "fitness for purpose" in validation of diagnostic tests and lists 6 purposes: (1) to demonstrate population 'freedom' from infection (zero prevalence); (2) to demonstrate freedom from infection or agent in individual animals or products for trade purposes; (3) to eradicate infection; (4) to confirm a diagnosis of clinical cases; (5) to estimate prevalence of infection to facilitate risk analysis; and 6) to determine immune status in individual animals or populations (OIE, 2010). In the context of paratuberculosis, additional purposes, e.g. to control MAP to maximize profit, are relevant. Examples of MAP-specific testing purposes are given in Collins et al. (2006) and Nielsen and Toft (2008). The purpose for testing (i.e. proposed use or application of the test) should be explicitly stated to assist readers in making inferences about application of results. Generally, test accuracy estimates are considered to be valid only for the purpose for which the test has been validated and purpose-specific test validation in one population cannot be generalized unconditionally to other populations without careful scrutiny prior to acceptance (Greiner and Gardner, 2000). The example above would be improved by specification of geographic location in the abstract or objective, although this information can be ascertained by further examination of the publication.

Item:3 Describe study sampling frame: Describe the source population and inclusion and exclusion criteria, setting and locations where data were collected for all relevant levels of the study sample (animals and herds).

Example (Hypothetical based on revision of information in Lombard et al., 2006)

The target and source populations for the National Animal Health Monitoring System's Dairy 2002 study were operations and cows in 21 states that represented at least 70% of the dairy cows and dairy operations in the US. The median size of dairy herds in these states was 110 cows (range, 60–1800). A cow was defined as an adult bovid that had given birth to at least one calf. The study sample for Phase II was a stratified random sample of operations in the 21 states and was restricted to operations with at least 30 cows and where the operator was willing to participate in biologic sample collection.

Explanation

For paratuberculosis test accuracy studies, information about the target population may only be available for a limited number of descriptors, mostly at the herd level (e.g. species and for cattle, whether dairy or beef (non-dairy), herd size, and geographic location such as region or state) and is unlikely to be readily available for paratuberculosis testing history etc. Paratuberculosis prevalence and distribution in the target and source populations should be reported, if known. The relevant information for the herds can be presented as means (medians) and standard deviations (ranges) for quantitative variables, and frequencies for categorical variables. The example represents a modification of a prior report (Lombard et al., 2006) to more clearly define the target and source populations. Another example (see item 4) is Pitt et al. (2002) which reports the number of dairy herds in a geographically defined target population. Location of the target and source populations is important since this factor may affect specificity of antibody tests for paratuberculosis. Detailed descriptive information relevant to the study sample is easier to obtain compared with the target and source populations and should be reported for biologically important variables described in item 15.

Item:4 Describe selection of animals and herds: Describe sample selection methods (random, convenience, etc.) within each level of the sampling hierarchy (e.g. regions, farms, barns, cows) including exclusion criteria and number of study animals and herds.

Example

The dairy cattle industry in north Queensland is confined to three shires on the Atherton Tablelands and has 202 herds in total. Herds included in the study were those assessed as being most at risk of having JD [Johne's disease]. This assessment was done in consultation with advisers and private veterinarians and by review of herd health records, past JD ELISA and complement fixation testing, interstate import records and local knowledge. Risk factors included cases of chronic diarrhea, wasting, unexplained deaths and introductions from states with endemic JD. Twenty-five dairy herds were identified as at risk by this initial assessment and, of these, 18 were sampled in the preliminary round of testing that was conducted between August and November 1995. A total of 475 dairy cows.....were included in the study. At least 25 mature cows

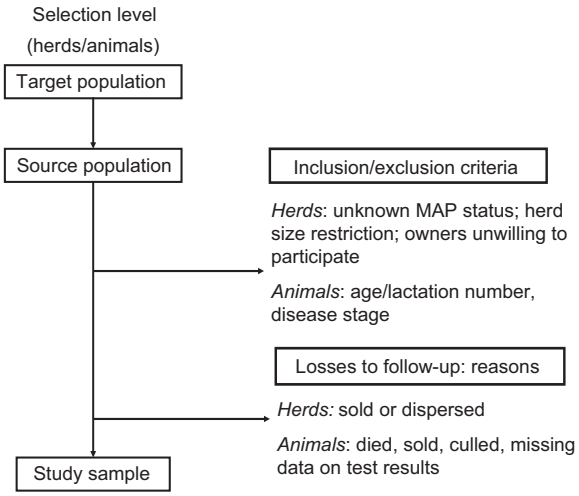

Fig. 1. Generic flowchart for reporting of relevant information for the source population and study sample in a diagnostic test evaluation study for paratuberculosis in ruminants.

per herd were sampled, and cows that had a positive or high negative result in the absorbed ELISA were subject to further investigation..... which included three collections of feces for bacteriological culture and blood sampling for absorbed ELISA..... Retesting [fecal culture and ELISA testing for 19 cows] was performed in late November 1995, in March 1996 and in June 1996 (Pitt et al., 2002).

Explanation

Two-stage sampling (herds selected first and then animals within herds) is often used for test evaluation studies for paratuberculosis. If the term "random" is used to describe herd and animal selection, a formal description of the procedure should be used. An example that includes description of a stratified random sampling protocol is Scott et al. (2007). Although the STARD statement uses the term "participant" in several items, "participant" should be reserved for the owner or manager of the herd and its animals and who gives permission to collect samples for testing (O'Connor et al., 2010). Herd owners (participants) are recruited but it is the animals or herds to which the tests are applied. There may be instances where herd owners are asked, but decline, to participate which could bias a study, e.g. registered breeders afraid of a negative impact on their reputation/sales if positive test results are found.

Accounting for inclusion and exclusion of herds and animals can be done in the text or in a simple flowchart (Fig. 1). To our knowledge, flowcharts have not been used for reporting of paratuberculosis test accuracy studies but could enhance reader understanding of complex sampling protocols. Fig. 2 shows the herd and animal selection process for a study that evaluated a milk ELISA for MAP (Lombard et al., 2006) and was constructed by the lead author of that paper for purposes of this report. The study was conducted as part of a large-scale national study which makes the flowchart more complicated than the majority of test accuracy studies. Criteria for inclusion at the 2 critical steps (National Agricultural Statistics Service (NASS) sample and selection for MAP testing) are shown. Although the study started with a random sample of dairy operations, the operations that participated in the evaluation were a convenience sample. Selection of cows for testing was also not random. The study sample selected for serum ELISA and fecal culture testing was cows that were primarily second lactation or greater and cows for milk testing were only lactating cows.

Item:5 Describe sampling protocol: Describe the collection, specimen size, transportation, handling and storage of specimens prior to the performance of the TUE and the reference standard.

Example

There were five groups of samples for this study (Table 1). The first two were pooled fecal samples used in an earlier study. The remaining three were pooled fecal samples, individual fecal samples and tissue samples submitted by veterinarians during routine surveillance for ovine paratuberculosis or collected during field-based epidemiological studies or experimental infections.....Fecal samples were submitted chilled (approximately 4 °C),sometimes via other laboratories. Upon receipt at the main laboratory, samples were placed at 4 °C or at −80 °C if they could not be processed within 4 days. Some samples were homogenised upon receipt then stored at −80 °C if culture was to be delayed.....The interval between collection of samples from sheep on the farm and receipt of those samples at the laboratory ranged from 1 to 7 days (median 3 days). Processing capacity was sometimes exceeded at the laboratory, necessitating storage of samples at −80 °C upon receipt. Consequently there were variable receipt-to-homogenisation (range 0–54 days, median 1 day) and homogenisation-to-decontamination (range 0–29 days, median 1 day) intervals (Whittington, 2009).

Explanation

Sample handling procedures may affect test accuracy estimates. Specific information on sample collection volume, use of individual sleeves for collection of feces, shipping information such as delays, on ice, etc. should also be included in the Materials and Methods section of the manuscript, wherever applicable. An example where there is complete description of sample collection and handling for serum, tissues and feces is McKenna et al. (2005a). For pooled samples, the weights of fecal material or number of pellets should be specified as well as the sample sizes and additional handling procedures, if relevant. Scott et al. (2007) provides a good example of the necessary information for pooled sampling.

Item:6 Describe study design. Was data collection planned before the TUE and reference standard were performed (prospective study) or after (retrospectively)?

Example

Three hundred and seventy-one animals.....were included in the present study and retrospectively categorized as follows: (1) infected animals that produced at least one culture-positive fecal sample in the study period ($n = 42$ animals), or were culture-positive from tissue samples at slaughter ($n = 2$, only a limited number of animals

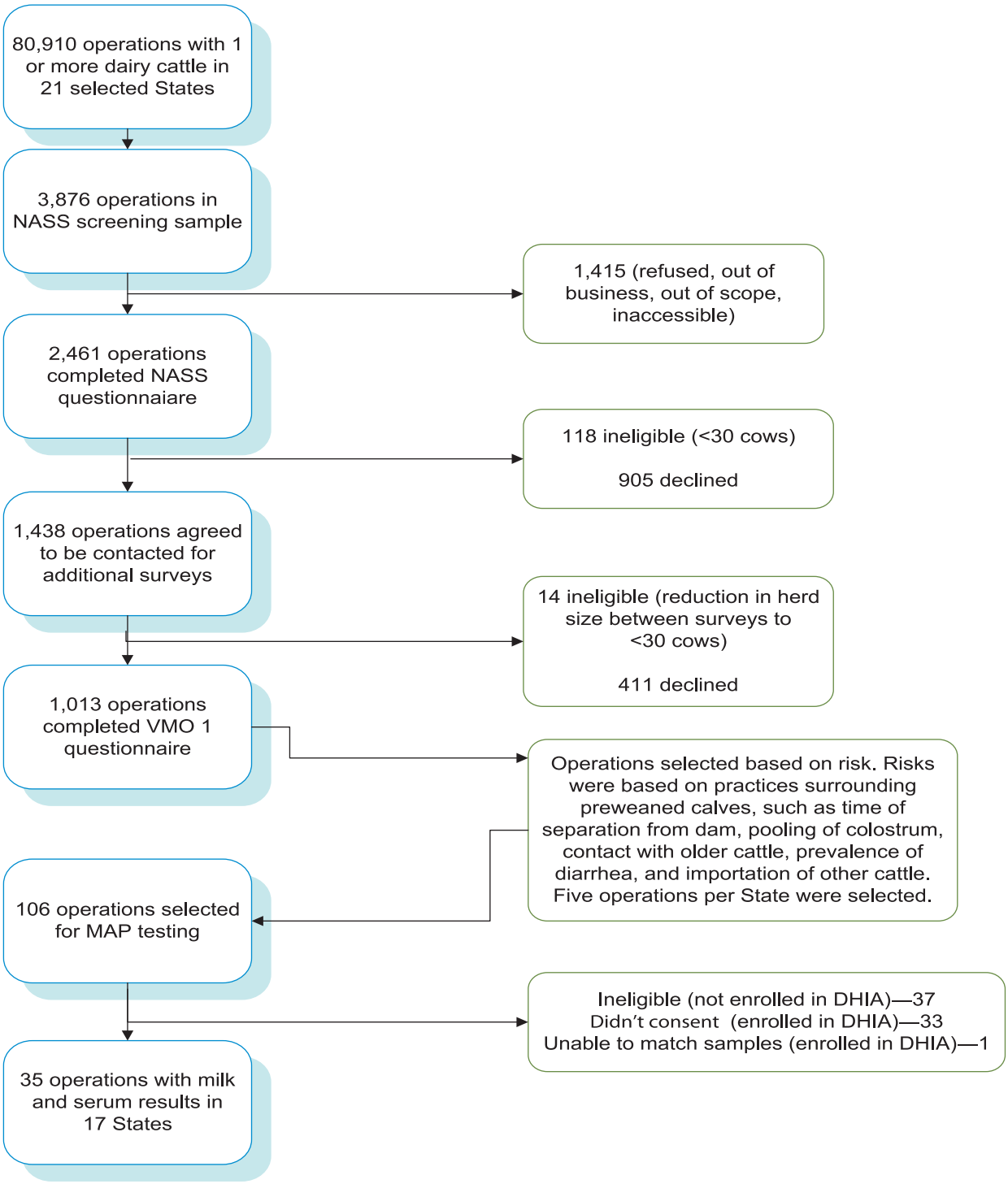

Fig. 2. Flowchart for herd and animal selection based on the National Animal Health Monitoring System's Dairy 2002 study (Lombard et al., 2006). The chart was prepared by one of the coauthors (JL) based on the original manuscript and other reports and data. DHIA = Dairy Herd Improvement Association; VMO = Veterinary Medical Officer; NASS = National Agricultural Statistics Service.

were tissue cultured); (2) exposed animals ($n = 208$), originated from infected herds, but without culture-positive fecal samples in the sampling period; (3) non-infected animals ($n = 119$), originating from five non-infected herds that were without evidence of paratuberculosis at the time of herd selection.....and without any culture-positive samples in the following 2-year study period (Huda et al., 2004).

Explanation

Retrospective designs use the results of the reference standard and TUE from samples collected in the past. For example, the performance of various strategies to classify herd-level MAP infection status (herd-level TUEs) could be evaluated using test results of individual cattle from herds previously tested (Jordan, 1996). In prospective designs, samples are collected and the TUE and reference standard are evaluated after the initiation of the study and, presumably, for the purpose of test evaluation. Results of retrospective studies may be valid, but because the test samples often were collected for a different purpose, it is possible for selection bias to affect the estimates of test performance. Prospective studies have the advantage of forethought in subject and sample selection. The sequence of testing is important when the study animals are selected. The required sample size often is reduced if some selection criteria, including historical information about the study units, are included based on retrospective sampling. However, these selection criteria, which could favor selection of animals with advanced stages of infection, can potentially lead to spectrum bias and overestimation of test performance, unless the data are interpreted within the selection strata (Ransohoff and Feinstein, 1978; Nielsen et al., 2011). Therefore, prospective sampling including a completely random inclusion of animals would generally be more appropriate but make the required sample size larger. Hence, it is important to specify the sequence of data collection and the study design (Nielsen et al., 2011) so the reader can appropriately judge the inferences.

Item:7 Describe the reference standard and its rationale.

Example

The advantage to the use of tissue culture as a comparison standard represents a wider spectrum of animals representing all stages of disease for determination of sensitivity.....It is apparent that the use of tissue culture as a gold standard results in a lower estimate of [ELISA] sensitivity.....The specificity estimates determined here were included for completeness; however, a more accurate comparison standard for specificity estimation would be animals from herds known from their history not to be infected with Mptb [MAP]. This type of information was not available for the cattle used in this study. Therefore, the specificity estimates determined here would be more appropriately viewed as apparent specificity based on a non-conventional comparison standard (McKenna et al., 2005a).

Explanation

Choice of and justification for the reference standard is a critical issue in test evaluation studies. In the McKenna et al. (2005a) study, two different reference standards (tissue culture results and fecal culture results) were used based on samples collected at slaughter. Each is described in detail in the text, including sample collection, preparation, storage conditions, and culture methodology and isolate confirmation. The authors report the sensitivity and specificity and 95% confidence intervals (CI) of 3 ELISAs relative to the two described reference standards. The discussion section of the cited manuscript includes considerations of the relative merits of each positive reference standard and discussion of the negative reference standard in this study (tissue culture- or fecal culture-negative) versus a more stringent standard. The reference standard at the herd-level may involve multiple criteria and risk-based sampling may be used to establish the true herd status.

Item:8 Describe technical specifications of materials and methods involved including how and when measurements were taken, and/or cite references for TUE and reference standards. Specify quality control samples for TUE and reference standard and specimen/analytical unit size of tested samples.

Example

Two types of controls were used for the fecal PCR assay. Extraction controls were tested once per extraction. For negative extraction controls, DNA was extracted from fecal samples from cattle from dairy herds known to be uninfected (level 4 of the Voluntary Johne's Disease Test Negative Program for Cattle) and tested with the MAV2 TaqMan PCR assay. Positive extraction controls were created by extracting DNA from fecal samples from cattle known to be uninfected that was then spiked with *M. avium* subsp. *paratuberculosis* and tested with the MAV2 TaqMan PCR assay. In the MAV2 TaqMan PCR assay, positive template controls consisted of dilutions of *M. avium* subsp. *paratuberculosis* DNA (previously extracted), with two positive template controls per plate. Eight no-template controls (NTC) consisting of nuclease-free water were tested per test plate (Wells et al., 2006).

Explanation

Complete description of culture methods is needed because laboratories often modify methods to suit their needs including availability of media and reagents, and there are substantial differences in sensitivity and contamination associated with these methodological variations (Whittington, 2010). Unambiguous definitions of positive and negative test results are required. In contrast to serum ELISA testing where use of quality control samples plate on each 96-well plate is standard practice in most laboratories, control samples (e.g. feces from a low MAP shedder) are typically not used by most laboratories evaluating culture-based methods. A description of quality control (QC) samples used in a fecal PCR evaluation study (Wells et al., 2006) is used as the example for this item to show the needed information.

Item:9 Results of the TUE and reference standard.

Example

The ELISA A reports the analyzed optical densities (OD) as an S/P ratio (sample OD to positive control OD ratio). The ELISA B reports as a score value, which is determined in relation to the cut-off that is determined by the mean of the negative controls plus 0.100. The ELISA C reports a pp-value (percent positive), which is based on a regression analysis of log–log transformed OD values. The calculation involves generating a linear regression of the blanked OD values and "log–log" transformed OD values, and using the inverse slope of this line multiplied by the log of the OD of the sample to arrive at the pp-value. This calculation

is performed to standardize the linear relationship of the test values with corresponding increases in antibody levels (McKenna et al., 2005b).

Explanation

To improve comparability of test results from different runs of the same ELISA assay or results from different laboratories, the data often are standardized (or normalized) based on values of positive and negative quality control samples which are included on each plate. The example from McKenna et al. (2005b) provides a description of calculations and reporting of data transformations involved with ELISA. The criteria for interpretation of normalized data must be prescribed; for example, a positive result for a paratuberculosis ELISA was defined as a S/P ratio >70% in Gumber et al. (2006). When herd status for MAP is determined based on ELISA test results of multiple individual samples, there are 2 cutoff values – the cutoff (threshold) for interpretation of each individual test results and the number (or percentage) of positive test results – to designate the herd-level TUE result as positive (Christensen and Gardner, 2000). Regardless of whether an indirect detection test (IDT) or a direct detection test (DDT) is evaluated, the number of individual samples tested by the TUE and reference standard for herd classification should be reported.

Real-time PCR data may be normalized by various means but the number of targeted genomic copies derived from a standard curve of Ct (cycle threshold) values is considered to be the most appropriate method (Bustin et al., 2009). The standard curve is based on a series of internal controls, and it provides within-assay assessment of the limit of detection in that particular run. Lower limits of detection usually are not well defined and are stochastically limited (Bustin et al., 2009), which is a problem for selection of a positive–negative cutoff value. Technically the detection limit should lie within the range of Ct of the standard curve.

Item:10 Describe the name, location, and qualifications of the laboratory, including the number, training and expertise of persons executing the TUE and reference standard.

Example

Each of the technicians involved [in 2 laboratories that participated in an evaluation of interlaboratory variation in test results of a commercial ELISA kit] had passed a national proficiency test for the ELISA method of detecting antibodies to MAP in bovine serum (Adaska et al., 2002).

Explanation

The number, training, and expertise of those carrying out the TUE should be reported, not only in order to allow evaluation of the investigators' expertise, but also to facilitate comparisons between studies using the same TUE and to make possible inferences with regard to potentially discordant results. The diagnostic accuracy of a particular TUE or reference test may be affected by variability in the manipulation, processing or reading of the test results by the individual(s) conducting the assays and interpreting their results (Elmore and Feinstein, 1992). Information about the training and expertise of these individuals can provide a reader with an idea of data quality. The more extensive the expertise and training of the individual(s) conducting the test, the more confidence the reader can have in the interpretation of the assay (Brealey et al., 2002). Appropriate examples for paratuberculosis are rare with the exception of the example shown (Adaska et al., 2002).

Item:11 Describe whether or not the readers of the TUE and reference standard were blind (masked) to the results of the other test and describe any individual or herd-level information available to the readers.

Example

Laboratory personnel were blinded to the common identity of the milk and fecal samples (Hendrick et al., 2005).

Explanation

Information about blinding of personnel performing the TUE and the reference standard test is essential so readers can assess potential bias in the study. Knowledge of the results of the reference standard can influence the interpretation of the TUE, and vice versa. Such knowledge is likely to increase the agreement between results of the TUE and those of the reference standard, leading to bias, and possibly an inflated measure of diagnostic accuracy (Philbrick et al., 1980). Withholding information from the individual(s) conducting the test, including results of other tests, is critical to minimize or prevent bias. Studies have shown that inappropriate masking may have substantial effects and produce an inaccurate measure of test accuracy (Detrano et al., 1989; Lijmer et al., 1999). If a blinded third party broke code after the analysis was done, this should also be mentioned. Blinding might be difficult to achieve for small research teams but is highly desirable and should be the recommended standard of practice. Strategies that can assist blinding include bar-coding of samples and use of an outside source or personnel specifically dedicated to that task.

Item:12 Describe methods for calculating and comparing measures of diagnostic accuracy and statistical methods used to quantify uncertainty in the estimates (e.g. 95% confidence intervals).

Example 1 (Hypothetical based on modification of Collins et al. (2005)

The sensitivity and specificity of the 5 ELISA tests were estimated with 95% confidence intervals (CI) accounting for the clustered sampling of animals in 7 infected and 7 non-infected herds, respectively (Dohoo et al., 2009). McNemar's test for correlated proportions was used to evaluate whether pairs of tests had different sensitivities and specificities at $p = 0.05$. Likelihood ratios (LR) for serum-to-positive intervals of assay A results were computed with 95% confidence intervals (CI) using the Med-Calc software (www.medcalc.be/). In Medcalc, CI for LR are estimated using a Poisson approximation with no adjustment for clustering.

Example 2

We applied latent-class models.....in a Bayesian framework to estimate the Se and Sp of the ELISA and the

FC....., separately in sheep and in goats. The estimation of Se and Sp of the ELISA and the FC was based on their cross-classified results. For the method to be valid, three main assumptions need to be met: (a) the diagnostic tests should be conditionally independent of each other.....; (b) the target population should consist of two or more sub-populations with different prevalence (P); (c) in these sub-populations, the Se and Sp of the diagnostic tests should be constant (Kostoulas et al., 2006).

Explanation

A unique aspect of test evaluation studies in animal compared with human medicine is use of clustered designs which typically require accounting for the design effect and inflation of variance estimates when CI are estimated (Dargatz and Hill, 1996; Dohoo et al., 2009). In a longitudinal design, repeated testing using the TUE will yield correlated results that need to be accounted for in the case definition for positive/negative results or in the statistical analysis. To our knowledge, adjustment for clustering has not been used for paratuberculosis studies but could be reported as shown in example 1. For CI calculation, the method used should be specified. Normal approximations are unlikely to be appropriate for small samples and when estimates are close to zero or unity. Exact methods, which are increasingly used, are preferable to normal approximations but provide conservative interval estimates (Newcombe, 1998).

Because latent-class statistical methods are used frequently in paratuberculosis test accuracy studies for samples collected ante-mortem (see for example, Nielsen et al., 2002 and Wells et al., 2006), the methods and underlying assumptions inherent in the approach should be described as in the second example. For Bayesian analyses, the prior distributions should be specified and justified. For more complex models with covariates (Norton et al., 2010), the modeling section of the paper should provide greater detail and code for Bayesian models should be included in an appendix or referenced.

Item:13 Describe methods for calculating test repeatability and reproducibility, if done.

Example

Variance components were calculated for the S/P [sample OD to positive control OD] ratios. Four factors were considered as sources of variability in the S/P ratios: laboratory, kit lot, wells of the 96-well microtiter plate in which a particular sample was tested, and date of testing. The date factor represents day-to-day variability. The contribution of each of the factors to the overall variability in S/P ratios was estimated [using Proc Varcomp program in SAS]. Ratios for the P and HP samples were analyzed separately. All factors were assumed to be random, and the well in a plate was nested within kit lot (Dargatz et al., 2004).

Explanation

Repeatability of a test at the sample level and at the laboratory analytical level and analytical or procedural reproducibility, defined as between laboratory variation, are important considerations for assessing how well a test might perform when used within the same laboratory over time and among laboratories, respectively. Most studies of repeatability and reproducibility for paratuberculosis involve ELISA tests (see for example Collins et al., 1993; Dargatz et al., 2004; Gumber et al., 2006). Data analysis typically involves estimation of the standard deviation (SD) or the CV of replicate samples among plates and laboratories for ultimate use as a quality assurance criterion. The CV is most appropriately used when the SD is proportional to the mean of the replicates. Sometimes, more complex designs with variance components models are used to quantify the relative contributions of different sources of variation (Dargatz et al., 2004) as shown in the example. Repeatability and reproducibility for dichotomous or ordinal test results can be estimated using kappa values which measure the chance-corrected agreement within and between laboratories, respectively (Dohoo et al., 2009). If repeatability of the TUE has been estimated and reported in a prior study, the relevant citation should be given.

Evaluations of the repeatability and reproducibility of DDT, especially culture and PCR, are rare; an exception is Kawaji et al. (2007). Lack of sample homogeneity may be a very influential factor affecting the reproducibility of a DDT, especially in samples transported to other laboratories. Although repeatability and reproducibility of DDT can be evaluated using spiked samples, use of naturally contaminated fecal samples from low to moderate MAP shedders would likely better represent the combined effects of handling, storage, and transportation in addition to potential clumping of MAP which might be impossible to replicate in a spiking experiment.

Item:14 Report when study was done, including beginning and end dates of recruitment.

Explanation

For most published paratuberculosis studies, the beginning and end dates of the study are included with information about sampling protocol in the Materials and Methods section as shown in the item 4. However in longitudinal studies, reporting of details of recruitment may be necessary. Huda et al. (2004), for example, indicated that animals entered their study at 12 months of age.

Item:15 Report demographic and other biologically relevant characteristics of the study sample at the individual (e.g. age, sex, breed, and risk factors) and at the herd levels (e.g. production system).

Example

Dairy cattle from 14 herds were included in the study. All herds were comprised of Holstein cattle, except one that had Jerseys. The uninfected population was comprised of 359 adult cattle from seven Minnesota dairy herds designated status level 4 according to the criteria of the U.S. Voluntary Bovine Johne's Disease Herd Status Program.....Seven known *M. paratuberculosis*-infected Wisconsin herds, comprised of 2094 adult cattle, were used to find cases of bovine paratuberculosis. These herds had no previous history of systematically testing for paratuberculosis or removal of test-positive cattle. The infected and non-infected herds were similar in many respects,

including their standardized risk assessment scores for *M. paratuberculosis* infection transmission (Table 1) (Collins et al., 2005).

Explanation

MAP infections are chronic with several infection stages. The age distribution and distribution of animals in different infection stages can greatly affect the sensitivity estimates and should be reported separately by subpopulation (item 18), wherever possible. Furthermore, the reader should be able to understand the production system in which the animals were kept so that the applicability to other populations can be assessed. Prior testing history and actions taken on the basis of test results can affect generalizability of results and should be described as in the example. Risk assessment scores also provide a useful summary assessment of implemented herd management practices to control paratuberculosis. When data from a few herds (<20) are included in the study, relevant information may be represented in tabular format as done in the example and in Gumber et al. (2006), or it may be summarized as percentages as in Hope et al. (2000).

For the study sample, the following animal and herd information was considered essential by the experts for inclusion under this item. Other information may be relevant depending on the specific circumstances:

Animal: clinical status (clinical vs. subclinical); age or lactation number; stage of lactation at testing for dairy cattle.

Herd: geographic location (country and region/state), herd size with the criterion e.g. animals greater than 2 years (for herd-level diagnostics at least); whether paratuberculosis vaccination used (yes/no and if yes, brand name and manufacturer of vaccine used, and date of last vaccination); and production and/or housing system (e.g. freestall, drylot, tiestalls, stanchion barn, pasture etc.). This was considered to be essential for herd-level tests but only useful for individual tests. Skin testing history for *Mycobacterium bovis* may impact antibody test results for paratuberculosis (Varges et al., 2009) and should be reported. Prior paratuberculosis testing history is an important consideration since selective removal of test-positive animals, whether true or false-positive, can substantially impact sensitivity and specificity estimates (Whitlock et al., 2000). An example of reporting of within-herd test prevalence for study herds is provided in Aly et al. (2010).

Item:16 Report the number of animals and herds satisfying the criteria for inclusion that did or did not undergo the TUE and/or the reference standard: describe why animals and herds failed to receive either test.

Example

Of the 35 [dairy] operations, 21 had individual animal fecal culture results. Three of the 21 operations had no cattle that tested fecal culture positive, and so were not included in the fecal culture analysis.....Milk and serum ELISA results were available for 8552 and 6874 animals, respectively..... A total of 6349 animals had [both] milk and serum ELISA results, and of these, 1921 animals had fecal culture results (Lombard et al., 2006).

Explanation

In cross-sectional designs where testing by the TUE and reference standard do not occur simultaneously and in longitudinal studies, there can be multiple reasons why some cows and/or herds do not undergo all testing. For example, in the study by Lombard et al. (2006), cows that were not lactating at the time of the milk ELISA could not have been tested. Since all testing was not conducted simultaneously, some animals may have been removed or died in the period between tests. Additionally according to study protocols, first lactation animals were to be excluded from serum ELISA and fecal culture testing. Also, the number of herds with fecal cultures done was lower than for ELISA testing because of laboratory capacity issues. These discrepancies in numbers tested by each of the assays evaluated could be shown in a table (see item 19 for a modified example using the data from Lombard et al., 2006), in a flowchart or in a simple case, described in the text. An example where death loss and reason for death are reported is Pitt et al. (2002).

Item:17 Report time interval between collection of samples for the TUE and the reference standard, and interventions administered between.

Explanation

For practical purposes, some sample types, e.g. milk, often are collected on different days from blood and feces collections because milk sampling is regularly scheduled by dairy industry associations in some countries. In addition, it is well known that fecal culture and ELISA results for paratuberculosis fluctuate over time and therefore investigators sometimes use a composite reference standard based on longitudinal test results. In these cases, the time (mean or median and range) between sequential tests of the same test type should be reported in either the results or methods section. A further problem arises when necropsy data are used as the reference standard and the TUE was based on samples collected earlier in life rather than at the time of necropsy when disease would be expected to have progressed to more advanced stages. Although there are few published data, sequential biopsy results from the terminal ileum and associated lymph nodes of naturally exposed sheep have shown that histopathological status can remain unchanged or change from negative to positive (most animals), positive to negative (few animals), or paucibacillary to multibacillary within 6 months (Dennis et al., 2011).

Treatments for paratuberculosis are not used in commercial operations but there is evidence that use of monensin sodium in dairy rations reduces ELISA positivity and fecal shedding of MAP (Hendrick et al., 2006a,b) and hence, its use should be reported. Other interventions (paratuberculosis vaccination or skin testing for tuberculosis) might affect IDT results (see items 18 and 20).

Item:18 Report distribution of severity of disease or stage of infection (define criteria), and other relevant diagnoses or treatments in animals in the study sample.

Example

The distribution of shedding rates among the culture-positive cows shows a preponderance of cows (45%) in the very low shedding category (Fig. 1).....There was a direct relationship between the level of fecal shedding of *M. paratuberculosis* and the percentage of positive assays (table 3). [For the 415 culture-positive cattle, the numbers of cattle were 229, 68, 36, and 82 in shedding level categories of >0–1, >1–2, >2–3, and >3–4, respectively]. Positive ELISAs were found for 6.9–28.6% (mean, 13.3%) of cows with low numbers of *M. paratuberculosis* in their feces (fecal scores > 0–1). At progressively higher fecal culture scores, the mean percentages of positive antibody assays for all five assays were 27.3, 54.9, and 78.4%, respectively (Collins et al., 2005).

Explanation

Spectrum bias affects sensitivity (Ransohoff and Feinstein, 1978). For chronic infections such as paratuberculosis, the timetable of infection progression makes disease spectrum readily evident (Nielsen et al., 2011). Generally, diagnostic tests perform better in late-stage paratuberculosis when the number of organisms in feces and tissues is high or when antibody concentration in serum and milk is high. Conversely, early in the course of infection, these same analytes are low in number or even absent making sensitivity poor. Thus, papers evaluating paratuberculosis diagnostic tests should explicitly describe the spectrum of disease in the study sample. Ideally, the spectrum of infection stages (proportion of animals in each infection stage) in the study sample should reflect the spectrum of infection stages in the target population. For paratuberculosis, a common surrogate for infection severity is the number of MAP in fecal samples, as used in Collins et al. (2005). A better indicator of disease severity is the histopathogical score at necropsy as reported for 152 infected sheep in Gumber et al. (2006) and for 224 sheep used for evaluation of ELISA in Sergeant et al. (2003).

Item:19 Report a cross tabulation of the results of each TUE (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.

Example

Distribution of milk and serum ELISA results and MAP culture results by fecal shedding level (table 5) from Lombard et al. (2006).

Explanation

Lack of clarity in presentation of the joint results of multiple diagnostic tests is common in many test evaluations for paratuberculosis. An exception is Alinovi et al. (2009) where results of 4 tests were cross-tabulated. Table 5 in Lombard et al. (2006), as shown in the example, presents a cross classification of serum ELISA and milk ELISA results individually and jointly by level of fecal culture shedding, as evaluated by parallel interpretation of results from 3 culture methods ($n = 1921$). However, discordant test results (milk positive and serum negative, and vice versa) were not explicitly reported. Greater transparency would have been achieved had results of all 3 tests been reported for all 6097 cattle that were tested by any of the 3 methods. The modified table (shown below) for dichotomized fecal culture results was reconstructed from the original data. This table also serves to completely account for milk and serum ELISA when no fecal sample was cultured (Not tested (NT) = 4176). A similar table could have been generated for the 105 herds, if considered warranted.

Milk ELISA	Serum ELISA	Fecal culture	Total
+	+	+	28
+	+	–	6
+	–	+	8
+	–	–	21
–	+	+	12
–	–	+	122
–	+	–	26
–	–	–	1698
+	+	NT	93
+	–	NT	56
–	+	NT	106
–	–	NT	3921

A novel way of presenting detailed test results across multiple populations was recently published (Caraguel et al., 2009). The same approach might have utility in paratuberculosis test accuracy studies.

Item:20 Report any adverse events from performing the TUE or the reference standard.

Explanation

Only in rare circumstances, would the TUE or reference standard have adverse consequences. If the reference standard is an invasive procedure such as full thickness ileum biopsy, there is the risk of standard post-surgical complications, and the presence or absence of these should be reported with the diagnostic test evaluation or separately (McConnel et al., 2004). Other situations where there might be unintended consequences include instances where skin test antigens could result in allergic reactions in the tested subject or alter immune responses to subsequent tests as was demonstrated with *M. bovis* (Buddle et al., 2010).

Fecal culture shedding		Milk ELISA			Serum ELISA			Milk and serum ELISA
	Total	Strong positive (%)	Positive (%)	Negative (%)	Strong positive (%)	Positive (%)	Negative (%)	Positives (%)
Heavy	13	61.5	15.4	23.1	53.8	38.5	7.7	69.3
Moderate	26	38.5	15.4	46.1	38.5	11.5	50.0	46.1
Low	83	7.2	3.6	89.2	4.8	4.8	90.4	6.0
Very low	48	2.1	4.2	93.7	4.2	10.4	85.4	4.2
Negative	1751	0.3	1.2	98.5	0.2	1.6	98.2	0.3
Total	1921	1.6	1.7	96.7	1.4	2.3	96.3	1.8

Item:21 Report estimates of diagnostic accuracy and a measure of statistical uncertainty (e.g. 95% confidence intervals).

Example 1 (Hypothetical example based on modification of Collins et al., 2005)

The sensitivities of the ELISA to detect cows shedding MAP in feces did not differ for 4 of the ELISAs ($P > 0.05$) and ranged from 28.0% to 28.9% but all these assays were lower than the sensitivity for assay D (44.5%, $P < 0.05$). Overall specificity among the 5 ELISAs evaluated ranged from 84.7% for assay D to 100% for assay C. Assay D had significantly ($P < 0.05$) lower specificity than all other assays, and assay A had significantly lower specificity than B, C and E. Sensitivity and specificity estimates and clustered-adjusted 95% confidence intervals (CI) for each assay are in table X. For assay A, the LR (95% CI) for the 3 categories of test results above the manufacturer recommended cutoff of 0.25 were 1.5 (95% CI = 0.8–3.0), 11.3 (95% CI = 2.7–47.1) and 63.2 (95% CI = 8.8–452.1) for serum-to-positive ratios in the intervals of >0.25–0.4, >0.4–1.0, and >1.0, respectively.

Example 2

Of the 400 animals initially sampled, results on both FC [fecal culture] and ELISA for 368 were obtained. For each species, cross-classified results of the ELISA and FC by sub-population are in table 3. At the recommended ELISA cutoff (S/P value = 0.4), posterior medians obtained under the independence model for the SeELISA, SpELISA, SeFC and SpFC, were 63% (95% CrIs [credibility intervals]: 42, 93%), 95% (90, 98%), 8% (2, 17%) and 98% (95, 100%) in goats and 37% (10, 80%), 97% (93, 99%), 16% (2, 48%) and 97% (95, 99%) in sheep, respectively (Kostoulas et al., 2006).

Explanation

Regardless of the choice of measure of diagnostic accuracy (sensitivity, specificity, likelihood ratios, area under the ROC curve), test performance results should be reported with 95% CI (or probability intervals in a Bayesian analysis) to capture the uncertainty in parameter estimates. The ease with which this can been done has increased with on-line calculators, freeware software programs, and the incorporation of estimators in commercial packages (e.g. SAS, Stata, Medcalc) and freeware (e.g. R, WinBUGS). The two examples demonstrate concise yet complete reporting of salient information based on the statistical methods described in item 12. The approaches for herd-level sensitivity and specificity estimation are similar to those for animal-level estimation and CI should be calculated based on the number of infected and non-infected herds that were tested, respectively.

Item:22 Report how indeterminate results, missing responses and outlier values of the TUE and the reference standard were handled. If additional testing of animals and herds is done to resolve discrepant results, then describe the rationale and approach (a flow diagram is strongly recommended).

Example

All DNA templates extracted from fecal samples were tested by the QPCR assay in duplicate, and in the case of contradiction (1 of 2 wells positive) the fecal sample was retested from the beginning of the DNA extraction. If either well was positive in the second test, the sample was defined as positive.(Kawaji et al., 2007).

Explanation

The example based on PCR testing provides a straightforward description of the necessary information. For culture, contamination with organisms other than MAP may be an issue irrespective of whether solid or liquid media are used. A retrospective analysis of test performance by Whittington (2009), in which irrelevant microorganisms were observed on 16% of the primary cultures on 7H10+MJ agar from the 1535 samples and in 76% of the 551 subcultures from BACTEC medium, provides an adequate description of the needed information. If additional testing or retesting is used to ultimately classify initial results that are indeterminate or outliers, the criteria for selection of samples should be described. An example of plate and sample-level quality control criteria for determining the need for ELISA retesting is Gumber et al. (2006).

There are different implications of applying additional testing depending on whether it is done on the reference test or the TUE. For example, adding a confirmatory test to the reference standard will increase specificity but reduce the sensitivity of the reference standard. This reduced pool of "true disease positives" will likely generate a higher estimate of sensitivity for the TUE. Adding a confirmatory test to the TUE will reduce its sensitivity and increase its specificity.

Item:23 Report estimates of variability of diagnostic accuracy between relevant subpopulations, readers, or testing sites, if done.

Example

The overall sensitivity of ELISA was 34.9% (95% confidence limits, 27.3–43.0) from 152 infected sheep but it varied from 19.3% to 50.0% depending on the extent of histopathological lesions [table 3 includes details]. The sensitivity of ELISA was greater for animals having histopathological lesion types HS 2 and HS 3 than for animals with other lesion categories. The sensitivity on the basis of positivity of both tissue culture and histopathology results was 32.8%.....The sensitivity of ELISA increased uniformly as the duration of exposure of animals increased.....(table 5).....The specificity of ELISA for sheep from Western Australia and NSW was 99.4% and 98.3%, respectively. The overall specificity of ELISA irrespective of states was 98.8% (table 6) (Gumber et al., 2006).

Explanation

Examples where subpopulations have influenced reported test accuracy values for paratuberculosis are common and predominantly involve bias in prevalence or stage of disease in the study animals. Although there is no formal relationship between prevalence and sensitivity in epidemiological theory, in practice paratuberculosis test sensitivity is affected by prevalence as it drives the contact rate with infected animals and the level of environmental contamination. The MAP load in the environment affects the degree of exposure and hence the incubation period, and therefore the proportion of individuals at a given

age which have reached a stage of disease consistent with fecal shedding or immunological response (reviewed in Whittington and Sergeant, 2001). For this scenario, apparent sensitivity of both fecal tests and serum ELISA for testing of any age stratum is improved.

Histological examination of intestine and associated lymph nodes provides one way of classifying/standardizing animals by stage of disease, as does body condition score, so that inferences can then be made to other herds based on expected proportions of cases at each disease stage. Although few reports would meet all criteria of reporting subpopulation detail, examples with adequate geographic origin, age and stage of disease data for ELISA, AGID or fecal culture/PCR are Gumber et al. (2006) and Kawaji et al. (2007). Examples where prevalence and stage of disease data have been reported with test accuracy estimates include Sergeant et al. (2002, 2003). Infections by, or exposures to, microbes that share antigens with MAP may cause variability in estimates of specificity. For paratuberculosis in goats, concomitant infection or vaccination with *Corynebacterium pseudotuberculosis* impacted the diagnostic specificity of one ELISA kit for paratuberculosis more than another (Manning et al., 2007). Geographic location may impact specificity of ELISA in cattle (Pitt et al., 2002), presumably due to exposure to environmental mycobacteria. Effects of covariates on sensitivity and specificity can be assessed by stratification or logistic regression modeling (Coughlin et al., 1992).

Herd-level sensitivity of a TUE is likely to be strongly dependent on within-herd prevalence of MAP and reporting of estimates by prevalence should be done. An example where the performance of a commercial ELISA used on bulk-tank milk samples was related to within-herd ELISA seroprevalence is Van Weering et al. (2007).

Item:24 Report estimates of test repeatability and reproducibility, if done.

Example

Modeling the S/P ratio for the P sample showed that the largest amount of variation was attributed to the kit lot (37.5%), followed by random error (27.0%) and interlaboratory variation (18.3%) (table 4). Modeling the S/P ratio for the HP sample showed a somewhat different distribution for the sources of variation (table 5). The largest proportion of the variation was attributed to random (unexplained) error (55.0%) followed by date (21.4%) and laboratory (17.1%). To assess the impact of plates with a low degree of separation between the negative and positive control means, 2 additional analyses were conducted for the S/P ratios for the P samples (table 6). In the first analysis, 19 plates in which the separation was less than 0.3 were not included and the percent contribution of the factors to the variance reassessed. This analysis was repeated once more, not including all plates with a separation of less than 0.4 (136 observations) (Dargatz et al., 2004).

Explanation

There is at least one study of repeatability at the sample level to evaluate how a single fecal sample represents feces from the cow at that time (Eamens et al., 2007). For detection of antigen or nucleic acid from MAP cells in low abundance, stochastic events associated with taking a subsample from a larger fecal sample may affect repeatability and should be reported as the frequency of occurrence of discrepant results among duplicates, the method of resolution of discrepancies, and the definition of positive and negative test outcomes where retests are required; an example is provided by Kawaji et al. (2007). In addition, it is well known that histopathological examination becomes more sensitive as more levels or sections of intestine or lymph node are examined because a single sample may not be representative of the tissue as a whole.

Analytical or procedural reproducibility among laboratories can be reported as SD or as the relative contribution from different sources of S/P ratio variability as in the example by Dargatz et al. (2004).

Item:25 Discuss the utility of the TUE in various settings (clinical, research, surveillance etc.) in the context of the currently available tests.

Example

Selection of an assay (e.g., among serum ELISA, BCF [bacterial culture of feces], and PCR assay) is not straightforward and will depend on such factors as cost; laboratory capacity, performance, and efficiency; and diagnostic test accuracy. Despite apparent relative improvements in Se and Sp for the qRT-PCR assay over values for the serum ELISA and Se for the qRT-PCR assay over that for BCF, other factors may be important when considering the appropriate test for a given herd situation and diagnostic laboratory. For example, although the enhanced direct-fecal qRT-PCR assay performed best for the commercial herd described here, other herds that contain many more young and preclinical (e.g., non-shedders or low shedders) cattle or cattle of unknown infection status may not be suitable for testing with this assay. A large herd with unknown MAP infection status would be more suited to screening by use of the serum ELISA, with follow-up confirmation of positive results by use of BCF, rather than to initiate testing with a relatively expensive alternative such as the direct qRT-PCR assay. On the other hand, in a herd of known infection status that contains older cows and in which culling decisions need to be made in a timely manner, the direct-fecal qRT-PCR assay may play a more important role (Scott et al., 2007).

Explanation

The concept of utility of a diagnostic test, as demonstrated in the example, has much broader implications in veterinary medicine, and especially livestock health, as compared to human medicine. Tests must be affordable for end-users in addition to having high accuracy. One way of addressing this is to incorporate economic outcomes into methods for setting the optimal assay cutoff for a specified purpose. Alternatively, conventional methods of assay accuracy, e.g. ROC analysis, can be used to establish assay sensitivity and specificity, which then can be incorporated into more sophisticated economic decision analysis models (Dorshorst et al., 2006). Regardless, of the method employed, a thorough diagnostic assay evaluation should include discussion of assay utility. Producers should expect that the benefits of use of the assay for a designated

purpose will outweigh the costs of testing. Laboratories should expect that the assay can be performed with sufficient throughput to meet client expectations, and at a laboratory fee that clients consider affordable. The paper should discuss, with regards to these indicators, how the current TUE compares to other commercially available tests in common use.

4. Conclusions and recommendations

STRADAS-paraTB is an expert-derived list of items based on STARD for improving the quality of reporting of test accuracy studies for paratuberculosis. Use of the checklist can potentially benefit many end-users of paratuberculosis tests including livestock owners, veterinarians, and animal health officials for the purpose of developing and implementing surveillance and testing programs for control of paratuberculosis. Moreover, the checklist and guidelines can assist researchers developing tests, reviewers and journal editors who subsequently consider manuscripts reporting research results for publication. We recommend that authors complete the STRADAS-paraTB checklist and include with their submitted manuscript. Journals publishing paratuberculosis test accuracy studies should "strongly encourage" the use of the checklist at all stages of the review process and could publish the list as a supplemental file. Because STRADAS-paraTB focuses on reporting standards and not design, we recommend that authors of paratuberculosis test accuracy studies consult Nielsen et al. (2011) for guidance about design aspects and strategies to prevent common biases evident in many published paratuberculosis validation studies.

Conflict of interest statement

None.

Acknowledgements

We thank Vivek Kapur and Tiffany Cunningham for technical assistance and the Johne's Disease Integrated Program (USDA-NIFA Award No. 2008-55620-18710) for funding the meetings of coauthors of the manuscript.

Appendix A. Terms and definitions

Paratuberculosis (Johne's disease): The MeSH definition is "a chronic gastroenteritis in ruminants caused by *Mycobacterium avium* subsp. *paratuberculosis*". However, the definition is restrictive and should include any subclinical stage of paratuberculosis infection.

Mycobacterium avium subsp. *paratuberculosis (MAP)*: A subspecies of Gram-positive, aerobic bacteria of the genus *Mycobacterium*. It is the etiologic agent of paratuberculosis (Johne's disease), a chronic gastroenteritis in ruminants (MeSH definition). Additional genomic considerations are described in Turenne et al. (2007).

Target condition: Underlying MAP infection status of interest (e.g. infected, infectious, or affected states of individual animals or herds). The target condition is normally measured based on quantification of an analyte or biological marker (e.g. MAP organisms or serum antibodies to MAP) that the test under evaluation (TUE) or reference standard detects as an indicator of the target condition.

Affected: A MAP-infected animal exhibiting one or more clinical signs of disease, such as weight loss, diarrhea, or edema.

Subclinical infection: A MAP-infected animal that appears clinically normal, i.e. does not have overt disease signs (observable by the herd owner or veterinarian) compatible with paratuberculosis such as diarrhea and low body condition, or reduced milk production. Most MAP-infected cows in a MAP-infected herd will be in the subclinical stage of infection.

Infected (non-infected) animal: An animal that has (does not have) MAP in its tissues.

Infected (non-infected) herd: A herd that has at least one MAP-infected animal (zero MAP-infected animals).

Infectious (non-infectious) animal: An animal that excretes (does not excrete) sufficient MAP bacteria to potentially infect one or more non-infected animals. Infectious animals are a sub-population of all MAP-infected animals.

Case definition: A practical definition of the target condition typically defined by the results of the reference standard.

Reference standard: A highly accurate diagnostic test, or combination of tests or observations that is used to classify the status of animals or herds according to the target condition. The term "gold standard" applies to a reference standard that is 100% sensitive and specific, or virtually so. Often, the term leads to unrealistic expectations and its use as an absolute standard precludes demonstration that a TUE has superior accuracy (Wilks, 2001). Ante-mortem "gold standards" do not exist for paratuberculosis.

Test under evaluation (TUE): A diagnostic method proposed for use to classify animals or herds with regard to the target condition. In STARD, the TUE is referred to as the "index test".

Test types

Direct detection tests (DDT): Tests that detect the live MAP organism or any subcomponent unique to MAP such as a gene by PCR, or antigen by immunoassay. DDT yield binary (positive/negative) or semi-quantitative results. The latter includes measurements such as the number of colony-forming units (CFU) per gram of cultured specimen on solid media, days-to-positive in liquid media or the number of cycle threshold (Ct) values for quantitative real-time PCR assays.

Indirect detection tests (IDT): Tests that quantify immune response such as antibodies to MAP in matrices such as serum or milk. IDT typically provide semi-quantitative results such as optical densities (OD) or sample-to-positive (S/P) ratios. These results are often interpreted as positive or negative at a single cutoff (threshold) value or using multiple cutoffs (e.g. low and high), as recommended by a test-kit manufacturer or diagnostic laboratory.

Test parameters

Animal-level sensitivity: The probability that an animal with the target condition (e.g. MAP infection) will test positive with the TUE.

Herd-level sensitivity: The probability that a herd with at least one animal with the target condition (e.g. MAP infection) will test positive with the TUE.

Animal-level specificity: The probability that an animal without the target condition (e.g. MAP infection) will test negative with the TUE.

Herd-level specificity: The probability that a herd with zero animals with the target condition (e.g. MAP infection) will test negative with the TUE.

Repeatability: A measure of the variability of test results in a single laboratory usually conducted by the same person (e.g. within run, run-to-run, or day-to-day). For tests measured on a continuous scale, repeatability is often expressed as standard deviations (SD) or coefficients of variation, if the SD is proportional to the mean of the replicates. For categorical tests, kappa is used for binary results (positive/negative) or in a weighted form when there are more than 2 categories (e.g. positive, intermediate, and negative).

Reproducibility: A measure of the variability of test results among laboratories following the same test protocol and estimated as for repeatability. Generally, among laboratory variation (reproducibility) in test results will exceed within laboratory variation (repeatability). Both repeatability and reproducibility (at least for direct tests) will vary enormously depending on the number of colony-forming units (CFU) in the fecal sample and/or how uniform or clustered MAP organisms (or the target) are in the sample.

Populations, samples and sampling

Populations and study sample: The *target population* is the population to which the estimates of test accuracy might be extrapolated. The *source population* is the population from which the study sample is drawn. The *study sample* (*sometimes termed study population*) is the sample of animals or herds which are included in the study (Dohoo et al., 2009).

Convenience sample: A sample from the source population not based on random selection methods.

Herd-level sample: Any type and number of samples used to classify the MAP status of a herd (a geographically defined population (cluster) of animals). Samples can be collected from individual animals and tested individually or in pools. In this case, herd-level test interpretation requires designation of the number or percentage of positive test results (threshold or cutoff value) required to classify the herd as MAP-infected. Alternatively, a herd-level test may utilize a single herd-level sample (e.g. bulk-tank milk).

Pooled sample: A composite sample (pool) that is obtained from at least 2 animals. The investigator controls the pooling procedure e.g. the number of animals that contribute samples to the pool and the amount or volume of the specimen.

Random sample: A set of animals drawn from a population (e.g. herd) using a formal random selection process such that each time an animal is selected, every animal in the population has a known, non-zero probability of inclusion in the sample.

Risk-based sample: An approach in which the sampled population is classified into subpopulations (e.g. by age or lactation number) with different prevalences or risks of MAP infection. The population with highest prevalence or risk of MAP infection is targeted for sampling to detect MAP, if present.

Sample size: The number of herds and animals from which samples were collected for testing.

Specimen size: The volume, weight or dimensions of a sample matrix submitted for testing by the TUE and/or reference standard. The amount of material tested (analytic unit size) in the assay, e.g. PCR, may be much smaller than the amount of specimen material submitted.

Prevalence

Within-herd test prevalence: The apparent (test) prevalence of the target condition (e.g. MAP infection) in a herd, typically calculated as the number of test-positive results by the TUE out of the total number of animals tested.

Within-herd true prevalence: The estimated true prevalence of the target condition (e.g. MAP infection) in a herd. This value is obtained by correcting the within-herd test prevalence estimate for the sensitivity and specificity of the TUE.

Test result classification

Positive: Test result indicative of MAP infection (e.g. test signal for an IDT above a defined cutoff value).

Negative: Test result not indicative of MAP infection (e.g. test signal for an IDT below a defined cutoff value).

Intermediate: Non-positive, non-negative result (also termed inconclusive, borderline, suspicious, or suspect in some test evaluation studies). This classification is based on construction of a three-zone partition of test results using two cutoff values. The cutoffs define three categories of test results (positive, intermediate, and negative).

Indeterminate: Test result that is not acceptable for technical reasons. Examples include insufficient response in positive control wells in gamma interferon assays, anti-complementary activity in a serum complement fixation test, and overgrowth of MAP cultures with contaminants precluding counting or observation of CFU.

References

Adaska, J.M., Munoz-Zanzi, C.A., Hietala, S.K., 2002. Evaluation of result variability with a commercial Johne's disease enzyme-linked immunosorbent assay kit and repeat testing of samples. J. Vet. Diagn. Invest. 14, 423–426.

Alinovi, C.A., Ward, M.P., Lin, T.L., Moore, G.E., Wu, C.C., 2009. Real-time PCR, compared to liquid and solid culture media and ELISA, for the detection of *Mycobacterium avium* ssp. *paratuberculosis*. Vet. Microbiol. 136, 177–179.

Aly, S.S., Mangold, B.L., Whitlock, R.H., Sweeney, R.W., Anderson, R.J., Jiang, J., Schukken, Y.H., Hovingh, E., Wolfgang, D., Van Kessel, J.A., Karns, J.S., Lombard, J.E., Smith, J.M., Gardner, I.A., 2010. Correlation between Herrold egg yolk medium culture and real-time quantitative polymerase chain reaction results for *Mycobacterium avium* subspecies *paratuberculosis* in pooled fecal and environmental samples. J. Vet. Diagn. Invest. 22, 677–683.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D., de Vet, H.C.W., 2003a. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin. Chem. 49, 1–6.

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

Brealey, S., Scally, A.J., Thomas, N.B., 2002. Review article: methodological standards in radiographer plain film reading performance studies. Br. J. Radiol. 75, 107–113.

Buddle, B.M., Wilson, T., Denis, M., Greenwald, R., Esfandiari, J., Lyashchenko, K.P., Liggett, S., Mackintosh, C.G., 2010. Sensitivity, specificity, and confounding factors of novel serological tests used for the rapid diagnosis of bovine tuberculosis in farmed red deer (*Cervus elaphus*). Clin. Vaccine Immunol. 17, 626–630.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., Vandesompele, J., Wittwer, C.T., 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin. Chem. 55, 611–622.

Caraguel, C., Stryhn, H., Gagné, N., Dohoo, I., Hammell, L., 2009. Traditional descriptive analysis and novel visual representation of diagnostic repeatability and reproducibility: application to an infectious salmon anaemia virus RT-PCR assay. Prev. Vet. Med. 92, 9–19.

Christensen, J., Gardner, I.A., 2000. Herd-level interpretation of diagnostic tests. Prev. Vet. Med. 45, 83–106.

Collins, M.T., Angulo, A., Buergelt, C.D., Hennager, S.G., Hietala, S.K., Jacobson, R.H., Whipple, D.L., Whitlock, R.H., 1993. Reproducibility of a commercial enzyme-linked immunosorbent assay for bovine paratuberculosis among eight laboratories. J. Vet. Diagn. Invest. 5, 52–55.

Collins, M.T., Wells, S.J., Petrini, K.R., Collins, J.E., Schultz, R.D., Whitlock, R.H., 2005. Evaluation of five antibody detection tests for diagnosis of bovine paratuberculosis. Clin. Diagn. Lab. Immunol. 12, 685–692.

Collins, M.T., Gardner, I.A., Garry, F.B., Roussel, A.J., Wells, S.J., 2006. Consensus recommendations on diagnostic testing for the detection of paratuberculosis in cattle in the United States. J. Am. Vet. Med. Assoc. 229, 1912–1919.

Coughlin, S., Trock, B., Criqui, M., Pickle, L., Browner, D., Tefft, M., 1992. The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test. J. Clin. Epidemiol. 45, 1–7.

Dargatz, D.A., Hill, G.W., 1996. Analysis of survey data. Prev. Vet. Med. 28, 225–237.

Dargatz, D.A., Byrum, B.A., Collins, M.T., Goyal, S.M., Hietala, S.K., Jacobson, R.H., Kopral, C.A., Martin, B.M., McCluskey, B.J., Tewari, D., 2004. A multilaboratory evaluation of a commercial enzyme-linked immunosorbent assay test for the detection of antibodies against *Mycobacterium avium* subsp. *paratuberculosis*. J. Vet. Diagn. Invest. 16, 509–514.

Dennis, M.M., Reddacliff, L.A., Whittington, R.J., 2011. Longitudinal study of clinicopathological features of Johne's disease in sheep naturally exposed to *Mycobacterium avium* subspecies *paratuberculosis*. Vet. Pathol. 48, 565–575.

Detrano, R., Gianrossi, R., Froelicher, V., 1989. The diagnostic accuracy of the exercise electrocardiogram: a meta-analysis of 22 years of research. Prog. Cardiovasc. Dis. 32, 173–206.

Dohoo, I., Martin, W., Stryhn, H., 2009. Veterinary Epidemiologic Research. VER Inc., Charlottetown, Prince Edward Island, Canada.

Dorshorst, N.C., Collins, M.T., Lombard, J.E., 2006. Decision analysis model for paratuberculosis control in commercial dairy herds. Prev. Vet. Med. 75, 92–122.

Eamens, G.J., Turner, M.J., Whittington, R.J., 2007. Sampling and repeatability of radiometric faecal culture in bovine Johne's disease. Vet. Microbiol. 119, 184–193.

Elmore, J.G., Feinstein, A.R., 1992. A bibliography of publications on observer variability. J. Clin. Epidemiol. 45, 567–580.

Gardner, I.A., 2010. Quality standards are needed for reporting of test accuracy studies for animal diseases. Prev. Vet. Med. 97, 136–143.

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

Gumber, S., Eamens, G., Whittington, R.J., 2006. Evaluation of a Pourquier ELISA kit in relation to agar gel immunodiffusion (AGID) test for assessment of the humoral immune response in sheep and goats with and without *Mycobacterium paratuberculosis* infection. Vet. Microbiol. 115, 91–101.

Hendrick, S.H., Duffield, T.E., Kelton, D.E., Leslie, K.E., Lissemore, K.D., Archambault, M., 2005. Evaluation of enzyme-linked immunosorbent assays performed on milk and serum samples for detection of paratuberculosis in lactating dairy cows. J. Am. Vet. Med. Assoc. 22, 424–428.

Hendrick, S.H., Kelton, D.F., Leslie, K.E., Lissemore, K.D., Archambault, M., Bagg, R., Dick, P., Duffield, T.E., 2006a. Efficacy of monensin sodium for the reduction of fecal shedding of *Mycobacterium avium* subsp. *paratuberculosis* in infected dairy cattle. Prev. Vet. Med. 75, 206–220.

Hendrick, S.H., Duffield, T.E., Leslie, K.E., Lissemore, K.D., Archambault, M., Bagg, R., Dick, P., Kelton, D.F., 2006b. Monensin might protect Ontario, Canada dairy cows from paratuberculosis milk-ELISA positivity. Prev. Vet. Med. 76, 237–248.

Hope, A.F., Kluver, P.F., Jones, S.L., Condron, R.J., 2000. Sensitivity and specificity of two serological tests for the detection of ovine paratuberculosis. Aust. Vet. J. 78, 850–856.

Huda, A., Jungersen, G., Lind, P., 2004. Longitudinal study of interferon-gamma, serum antibody and milk antibody responses in cattle infected with *Mycobacterium avium* subsp. *paratuberculosis*. Vet. Microbiol. 104, 43–53.

Jordan, D., 1996. Aggregate testing for the evaluation of Johne's disease herd status. Aust. Vet. J. 73, 16–19.

Kawaji, S., Taylor, D.L., Mori, Y., Whittington, R.J., 2007. Detection of *Mycobacterium avium* subsp. *paratuberculosis* in ovine faeces by direct quantitative PCR has similar or greater sensitivity compared to radiometric culture. Vet. Microbiol. 125, 36–48.

Kostoulas, P., Leontides, L., Enoe, C., Billinis, C., Florou, M., Sofia, M., 2006. Bayesian estimation of sensitivity and specificity of serum ELISA and faecal culture for diagnosis of paratuberculosis in Greek dairy sheep and goats. Prev. Vet. Med. 76, 56–73.

Lijmer, J.G., Mol, B.W., Heisterkamp, S., Bonsel, G.J., Prins, M.H., van der Meulen, J.H., Bossuyt, P.M.M., 1999. Empirical evidence of design-related bias in studies of diagnostic tests. J. Am. Med. Assoc. 282, 1061–1066.

Lombard, J.E., Byrem, T.M., Wagner, B.A., McCluskey, B.J., 2006. Comparison of milk and serum enzyme-linked immunosorbent assays for diagnosis of *Mycobacterium avium* subspecies *paratuberculosis* infection in dairy cattle. J. Vet. Diagn. Invest. 18, 448–458.

Manning, E.J., Cushing, H.F., Hietala, S., Wolf, C.B., 2007. Impact of *Corynebacterium pseudotuberculosis* infection on serologic surveillance for Johne's disease in goats. J. Vet. Diagn. Invest. 19, 187–190.

McConnel, C.S., Churchill, R.C., Richard, M.M., Corkhill, C.M., Reddacliff, L.A., Whittington, R.J., 2004. Surgical method for biopsy of terminal ileum and mesenteric lymph node of sheep for detection of *Mycobacterium avium* subsp *paratuberculosis*. Aust. Vet. J. 82, 149–151.

McKenna, S.L., Keefe, G.P., Barkema, H.W., Sockett, D.C., 2005a. Evaluation of three ELISAs for *Mycobacterium avium* subsp. *paratuberculosis* using tissue and fecal culture as comparison standards. Vet. Microbiol. 110, 105–111.

McKenna, S.L., Sockett, D.C., Keefe, G.P., McClure, J., VanLeeuwen, J.A., Barkema, H.W., 2005b. Comparison of two enzyme-linked immunosorbent assays for diagnosis of *Mycobacterium avium* subsp. *paratuberculosis*. J. Vet. Diagn. Invest. 17, 463–466.

Newcombe, R.G., 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat. Med. 17, 857–872.

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

Nielsen, S.S., Grønbæk, C., Agger, J.F., Houe, H., 2002. Maximum-likelihood estimation of sensitivity and specificity of ELISAs and faecal culture for diagnosis of paratuberculosis. Prev. Vet. Med. 53, 191–204.

Nielsen, S.S., Toft, N., Gardner, I., 2011. Structured approach to design of diagnostic test evaluation studies for chronic progressive infections in animals. Vet. Microbiol. 150, 115–125.

Norton, S., Johnson, W.O., Jones, G., Heuer, C., 2010. Evaluation of diagnostic tests for Johne's disease (*Mycobacterium avium* subspecies paratuberculosis) in New Zealand dairy cows. J. Vet. Diagn. Invest. 22, 341–351.

O'Connor, A.M., Sargeant, J.M., Gardner, I.A., Dickson, J.S., Torrence, M.E., Dewey, C.E., Dohoo, I.R., Evans, R.B., Gray, J.T., Greiner, M., Keefe, G., Lefebvre, S.L., Morley, P.S., Ramirez, A., Sischo, W., Smith, D.R., Snedeker, K., Sofos, J., Ward, M.P., Wills, R., 2010. The REFLECT Statement: methods and processes of creating reporting guidelines for randomized control trials for livestock and food safety by modifying the CONSORT Statement. Zoonoses Public Health 57, 95–104.

Office International des Epizooties (OIE), 2010. Manual of Diagnostic Tests and Vaccines for Terrestrial Animals 2010. Available at http://www.oie.int/en/international-standard-setting/terrestrial-manual/access-online (accessed February 23, 2011).

Peeling, R.W., Smith, P.G., Bossuyt, P.G.P.M., 2006. A guide for diagnostic evaluations. Nat. Rev. Microbiol. 4 (12 Suppl), S2–S6.

Philbrick, J.T., Horwitz, R.I., Feinstein, A.R., 1980. Methodologic problems of exercise testing for coronary artery disease: groups, analysis and bias. Am. J. Cardiol. 46, 807–812.

Pitt, D.J., Pinch, D.S., Janmaat, A., Condron, R.J., 2002. An estimate of specificity for a Johne's disease absorbed ELISA in northern Australian cattle. Aust. Vet. J. 80, 57–60.

Ransohoff, D.F., Feinstein, A.R., 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. New Eng. J. Med. 299, 926–930.

Scott, H.M., Fosgate, G.T., Libal, M.C., Sneed, L.W., Erol, E., Angulo, A.B., Jordan, E.R., 2007. Field testing of an enhanced direct-fecal polymerase chain reaction procedure, bacterial culture of feces, and a serum enzyme-linked immunosorbent assary for detecting *Mycobacterium avium* subsp. *paratuberculosis* in adult dairy cattle. Am. J. Vet. Res. 68, 236–245.

Sergeant, E.S., Whittington, R.J., More, S.J., 2002. Sensitivity and specificity of pooled faecal culture and serology as flock-screening tests for detection of ovine paratuberculosis in Australia. Prev. Vet. Med. 52, 199–211.

Sergeant, E.S., Marshall, D.J., Eamens, G.J., Kearns, C., Whittington, R.J., 2003. Evaluation of an absorbed ELISA and an agar-gel immunodiffusion test for ovine paratuberculosis in sheep in Australia. Prev. Vet. Med. 61, 235–248.

TDR, TDR Diagnostics Evaluation Expert Panel, 2006. Evaluation of diagnostic tests for infectious diseases: general principles. Nat. Rev. Microbiol. 4 (Suppl. 12), S20–S32.

Turenne, C.Y., Wallace, R., Behr, M.A., 2007. *Mycobacterium avium* in the postgenomic era. Clin. Microbiol. Rev. 20, 205–229.

Van Weering, H., van Schaik, G., van der Meulen, A., Waal, M., Franken, P., van Maanen, K., 2007. Diagnostic performance of the Pourquier ELISA for detection of antibodies against *Mycobacterium avium* subspecies *paratuberculosis* in individual milk and bulk milk samples of dairy herds. Vet. Microbiol. 125, 49–58.

Varges, R., Marassi, C.D., Oelemann, W., Lilenbaum, W., 2009. Interference of intradermal tuberculin tests on the serodiagnosis of paratuberculosis in cattle. Res. Vet. Sci. 86, 371–372.

Wells, S.J., Collins, M.T., Faaberg, K.S., Wees, C., Tavornpanich, S., Petrini, K.R., Collins, J.E., Cernicchiaro, N., Whitlock, R.H., 2006. Evaluation of a rapid fecal PCR test for detection of *Mycobacterium avium* subsp. *paratuberculosis* in dairy cattle. Clin. Vaccine Immunol. 13, 1125–1130.

Whiting, P., Rutjes, A.W.S., Reitsma, J.B., Bossuyt, P.M., Kleijnen, J., 2003. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med. Res. Methodol. 3, 25.

Whitlock, R.H., Wells, S.J., Sweeney, R.W., Van Tiem, J., 2000. ELISA and fecal culture for paratuberculosis (Johne's disease): sensitivity and specificity of each method. Vet. Microbiol. 77, 387–398.

Whittington, R.J., 2009. Factors affecting isolation and identification of *Mycobacterium avium* subsp. *paratuberculosis* from faecal and tissue samples in a liquid culture system. J. Clin. Microbiol. 47, 614–622.

Whittington, R.J., 2010. Cultivation of *Mycobacterium avium* subsp. *paratuberculosis*. In: Behr, M.A., Collins, D.M. (Eds.), Paratuberculosis. Organism, Disease Control. CABI, Wallingford, pp. 244–266.

Whittington, R.J., Sergeant, E.S., 2001. Progress towards understanding the spread, detection and control of *Mycobacterium avium* subsp. *paratuberculosis* in animal populations. Aust. Vet. J. 79, 267–278.

Whittington, R.J., Fell, S., Walker, D., McAllister, S., Marsh, I., Sergeant, E., Taragel, C.A., Marshall, D.J., Links, I.J., 2000. Use of pooled fecal culture for sensitive and economic detection of *Mycobacterium avium* subsp. *paratuberculosis* infection in flocks of sheep. J. Clin. Microbiol. 38, 2550–2556.

Wilks, C., 2001. Editorial: Gold standards and fool's gold. Aust. Vet. J. 79, 115.