Preventive Veterinary Medicine 138 (2017) 37–47

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

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

STARD-BLCM: Standards for the Reporting of Diagnostic accuracy studies that use Bayesian Latent Class Models

Polychronis Kostoulas[a,*], Søren S. Nielsen[b], Adam J. Branscum[c], Wesley O. Johnson[d], Nandini Dendukuri[e], Navneet K. Dhand[f], Nils Toft[g], Ian A. Gardner[h]

[a] Laboratory of Epidemiology, Biostatistics and Animal Health Economics, Faculty of Veterinary Medicine, University of Thessaly, Karditsa GR43100, Greece
[b] Department of Veterinary and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Grønnegårdsvej 8, DK-1870 Frederiksberg C, Denmark
[c] Biostatistics Program, Oregon State University, Corvallis, OR, 97331, USA
[d] Department of Statistics, University of California, Irvine, CA, 92697, USA
[e] McGill University Health Centre, McGill University, Montréal, QC, Canada
[f] Faculty of Veterinary Science, The University of Sydney, 425 Werombi Road, Camden, 2570 NSW, Australia
[g] Technical University of Denmark, National Veterinary Institute, Bülowsvej 27, DK-1870 Frederiksberg C, Denmark
[h] Department of Health Management, Atlantic Veterinary College, University of Prince Edward Island, Charlottetown, Prince Edward Island C1A4P3, Canada

ARTICLE INFO

Article history:
Received 1 April 2016
Received in revised form
21 December 2016
Accepted 9 January 2017

Keywords:
Sensitivity
Specificity
Bayesian analysis
Latent class models

ABSTRACT

The Standards for the Reporting of Diagnostic Accuracy (STARD) statement, which was recently updated to the STARD2015 statement, was developed to encourage complete and transparent reporting of test accuracy studies. Although STARD principles apply broadly, the checklist is limited to studies designed to evaluate the accuracy of tests when the disease status is determined from a perfect reference procedure or an imperfect one with known measures of test accuracy. However, a reference standard does not always exist, especially in the case of infectious diseases with a long latent period. In such cases, a valid alternative to classical test evaluation involves the use of latent class models that do not require a priori knowledge of disease status. Latent class models have been successfully implemented in a Bayesian framework for over 20 years. The objective of this work was to identify the STARD items that require modification and develop a modified version of STARD for studies that use Bayesian latent class analysis to estimate diagnostic test accuracy in the absence of a reference standard. Examples and elaborations for each of the modified items are provided. The new guidelines, termed STARD-BLCM (Standards for Reporting of Diagnostic accuracy studies that use Bayesian Latent Class Models), will facilitate improved quality of reporting on the design, conduct and results of diagnostic accuracy studies that use Bayesian latent class models.

1. Introduction

The Standards for Reporting of Diagnostic Accuracy (STARD) initiative (http://www.equator-network.org/reporting-guidelines/stard/) developed a checklist that should be followed in diagnostic accuracy studies. The original STARD checklist was simultaneously published in 2003 in 7 journals (Bossuyt et al., 2003) and comprised 25 key items for reporting. An updated list of 30 items was recently released (Bossuyt et al., 2015). Although STARD principles broadly apply – more than 200 biomedical journals encourage its use in their instructions to authors – the checklist is limited to studies designed to evaluate the accuracy of one or more tests when disease status is determined from a perfect reference (gold standard) procedure or when disease status is unknown and tests are evaluated against an imperfect reference procedure with known measures of test accuracy (e.g., known sensitivity and specificity). In such cases, standard statistical methods for making inferences about test accuracy apply (e.g., Broemeling, 2007).

An affordable, reliable, noninvasive reference standard does not always exist, especially in the case of infectious diseases with a long latent period (e.g. chronic infections). In such cases, a natural alternative for valid statistical evaluation of diagnostic test accuracy involves the use of latent class models (LCMs) that do not require knowledge of disease status (i.e., disease status is a latent variable) or the application of an imperfect reference procedure with known

* Corresponding author.
E-mail addresses: pkost@vet.uth.gr, polychronis.kostoulas@gmail.com (P. Kostoulas).

http://dx.doi.org/10.1016/j.prevetmed.2017.01.006
0167-5877/

test accuracy characteristics. Most applications of LCMs to diagnostic accuracy studies are based on cross-classified test-outcome data from two or more tests. Among the pioneering research on LCMs for diagnostic accuracy studies was the development of the two-test, two-population model introduced in Hui and Walter (1980). A thorough discussion of the applicability of latent class methods in diagnostic accuracy studies was first given by Walter and Irwig (1988). The World Organisation for Animal Health (OIE) has endorsed the use of LCMs to estimate sensitivity (Se) and specificity (Sp) based on test-outcome data from animals of unknown disease status (OIE, 2016).

Latent class analysis for diagnostic accuracy studies has been successfully implemented in a Bayesian framework for over 20 years (see, for example, Joseph et al., 1995Johnson et al., 2001; Branscum et al., 2005; Collins and Huynh, 2014; Enøe et al., 2000). Bayesian LCMs (BLCMs) are widely used because of their flexibility, the ease of interpretation of their results, and the availability of user-friendly software such as OpenBUGS (Lunn et al., 2009), which has popularized Bayesian data analysis. A Bayesian approach is necessary when using non-identifiable LCMs (i.e., statistical models that, broadly speaking, contain test accuracy and/or prevalence parameters that cannot be uniquely estimated by the data alone). A crucial aspect of any Bayesian analysis is proper justification of the prior distributions used in the primary and sensitivity analysis, a task that has increased importance when using non-identifiable models because the prior information will always influence the conclusions.

In addition to the required assumptions that need to hold in the case of any LCM approach for diagnostic accuracy studies, Bayesian inferential procedures require additional care primarily due to the incorporation of prior information in the analysis (Johnson et al., 2009; Jones et al., 2010). Hence, there is the necessity to adapt the STARD checklist in order to establish reporting requirements for diagnostic test accuracy studies that use LCMs in a Bayesian framework. To achieve this goal, we initially identified the STARD items that required modification for the case of BLCMs (Section 2). Subsequently, to provide an example for the necessity of the proposed modifications two authors (IG and SN) searched for published papers that used BLCMs to estimate the accuracy of tests for *Mycobacterium avium* subsp. *paratuberculosis* (MAP) infection in ruminants (Section 3). The example demonstrated the need for the proposed modifications and facilitated further refinement of these modifications. Finally, the modified items – termed STARD-BLCM – with examples and elaborations are in Section 4. A brief description of the statistical/methodological considerations of any Bayesian analysis is given in Section 5. In this paper, the term "diagnostic test" or "test" is used to describe any classification procedure (e.g., classifiers based on a single biological marker or a composite score from multiple biomarkers) designed to detect a specific target condition. Moreover, the terms "test accuracy" and "diagnostic accuracy" are used interchangeably.

2. Identification of the STARD items that require modification in the case of BLCMs

A review of the STARD statement revealed items that required modification in order to cover the additional information that should be reported when using BLCMs for the validation of diagnostic tests. Criteria for identifying such items were based on the principle of providing the necessary information that would be needed (i) to assess the internal and external validity of the tests and associated results, (ii) to infer the situations where results from a test accuracy study can be extrapolated as in the case of actually using the validated tests for disease control programs in comparable populations, (iii) to direct medical and veterinary professionals and other test users to statistically valid methods in Bayesian latent class analysis, and (iv) to enhance the filtering of validation studies that use BLCMs during the systematic review in meta-analysis projects.

A recent review and meta-analysis of studies that used LCM methods emphasized the importance of verifying that model assumptions are valid and proper reporting of methods that assessed the validity of these assumptions, since violations can lead to biased estimates of diagnostic accuracy (van Smeden et al., 2013). The authors found that 28% of the studies included in their meta-analysis failed to report any evidence that assumptions were verified or that the underlying models were of adequate fit to the data at hand. In their work, studies that used either Bayesian or frequentist methods were considered.

Some of the proposed modifications of the STARD are relevant to both Bayesian and frequentist estimation of LCMs. However, the focus is on Bayesian methods due to the complexity of Bayesian models, the additional element of prior information and the fact that – in the case of non-identifiable models – a meaningful solution can be obtained only by the incorporation of informative priors that have to be properly selected. Key elements that must be addressed in BLCMs are: (i) the absence of a reference test, (ii) the need to identify the condition that the tests under evaluation are targeting, (iii) an explicit description of the specified BLCM structure and (iv) a clear and justified specification of priors. To this end, items 1, 4, 6, 10, 11, 12, 14, 22, 24 and 27 of the STARD2015 statement were identified as components that require modification (Table 1).

3. Poor reporting in BLCM studies: the paratuberculosis example

3.1. Literature review and identification of published papers that used BLCM for the validation of diagnostic tests for Mycobacterium avium *subsp.* paratuberculosis

Mycobacterium avium subsp. *paratuberculosis* (MAP) causes a chronic intestinal infection of domestic and wild ruminants. Infections occur worldwide and are responsible for severe losses throughout the productive life of infected individuals. MAP infected animals can be less productive than herd-mates and, ultimately, die or are prematurely removed from a herd. MAP is a chronic infection with a long latent period and no perfect ante-mortem diagnostic test. These characteristics of MAP infection have contributed to the development and application of various BLCMs for investigating the accuracy of various combinations of conditionally independent or dependent MAP tests.

Two search engines, PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and Web of Science (Thomson Reuters, New York City, USA) were used to identify publications that used a BLCM for the validation of diagnostic tests for MAP. We used the Medical Subject Headings (MeSH) search terms "Bayesian, sensitivity, specificity, paratuberculosis" to identify papers published between 1 January 2009 and 23 March 2014, the latter being the date of the search. The resulting papers were then evaluated independently by two of the authors (IG and SN). Estimation of Se and Sp was required to be a specified objective, while prevalence studies where estimates of test accuracy occurred as a by-product were excluded. Studies that focused on methodological aspects of test evaluation, where the data were previously used in other studies or did not involve MAP infection, were also excluded.

The literature search yielded 23 peer-reviewed papers, 2 through PubMed alone, 8 through Web of Science alone, and the remaining 13 were identified by both search engines. Among these 23 studies, both authors identified 9 that met the inclusion criteria. Fourteen studies were excluded: 7 were classified as focusing

Table 1
Checklist of items for the Standards for the Reporting of Diagnostic accuracy studies by the use of Bayesian Latent Class Models (STARD-BLCM) based on the STARD2015 checklist. Modifications are in bold and either accompanied by further elaboration presented in this work or are textual changes. With the exception of the items relating to the incorporation of prior information the proposed modifications pertain beyond Bayesian estimation (i.e. maximum likelihood).

SECTION & TOPIC	ITEM	STARD2015	STARD-BLCM	Elaboration
Title/Abstract/ Keywords	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	Identification as a study of diagnostic accuracy, using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC) **and Bayesian latent class models**	Yes
Abstract	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	As is	No
Introduction	3	Scientific and clinical background, including the intended use and clinical role of the index test	Scientific and clinical background, including the intended use and clinical role of the **tests under evaluation**	No
	4	Study objectives and hypotheses	Study objectives and hypotheses, **such as estimation of diagnostic accuracy of the tests for a defined purpose through BLCM**	Yes
Methods				
Study Design	5	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	Whether data collection was planned before the **tests** were performed (prospective study) or after (retrospective study)	No
Participants	6	Eligibility criteria	Eligibility criteria **and description of the source population**	Yes
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	As is	No
	8	Where and when potentially eligible participants were identified (setting, location, and dates)	As is	No
	9	Whether participants formed a consecutive, random or convenience series	As is	No
Test methods	10a	Index test, in sufficient detail to allow replication	**10.Description of the tests under evaluation,** in sufficient detail to allow replication, **and/or cite references**	Yes
	10b	Reference standard, in sufficient detail to allow replication		
	11	Rationale for choosing the reference standard (if alternatives exist)	Rationale for choosing the **tests under evaluation in relation to their purpose**	Yes
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	**12.** rationale for test positivity cut-offs or result categories of the **tests under evaluation**, distinguishing pre-specified from exploratory	Yes
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory		
	13a	Whether clinical information and reference standard results were available to the performers or readers of the index test	**13.** Whether clinical information was available to the performers or readers of the **tests under evaluation**	No
	13b	Whether clinical information and index test results were available to the assessors of the reference standard		
Analysis	14	Methods for estimating or comparing measures of diagnostic accuracy	**14a. BLCM model** for estimating measures of diagnostic accuracy	Yes
			14b. Definition and rationale of prior information and sensitivity analysis	Yes
	15	How indeterminate index test or reference standard results were handled	How indeterminate results **of the tests under evaluation** were handled	No
	16	How missing data on the index test and reference standard were handled	How missing data of the tests under evaluation were handled	No
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	As is	No
	18	Intended sample size and how it was determined	As is	No
Results				
Participants	19	Flow of participants, using a diagram	As is	
	20	Baseline demographic and clinical characteristics of participants	As is	No
	21a	Distribution of severity of disease in those with the target condition	**21. Not applicable: the distribution of the targeted conditions is unknown, hence the use of BLCM**	No
	21b	Distribution of alternative diagnoses in those without the target condition		

Table 1 (*Continued*)

SECTION & TOPIC	ITEM	STARD2015	STARD-BLCM	Elaboration
	22	Time interval and any clinical interventions between index test and reference standard	Time interval and any clinical interventions between **the tests under evaluation**	Yes
Test results	23	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	Cross tabulation of the **tests'** results (or **for continuous tests results** their distribution **by infection stage**)	No
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	Estimates of diagnostic accuracy **under alternative prior specification** and their precision (such as 95% **credible/probability** intervals)	Yes
	25	Report any adverse events from performing the index test or the reference standard	Report any adverse events from performing the **of the tests under evaluation**	No
Discussion	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	As is	No
	27	Implications for practice, including the intended use and clinical role of the index test	Implications for practice, including the intended use and clinical role of the **tests under evaluation in relevant settings (clinical, research, surveillance etc.)**	Yes
Other information				
	28	Registration number and name of registry	As is	No
	29	Where the full study protocol can be accessed	As is	No
	30	Sources of funding and other support; role of funders	As is	No

Table 2
Summary of disagreement (D) between two reviewers (IG and SN) on the adequacy of the papers to fulfill the specific reporting items.

Paper	Item no.									
	4	6	11	10	12	14a	14b	22	24	27
Angelidou et al. (2014)			D						D	
Florou et al. (2009)	D	D	D						D	D
Fosgate et al. (2009)	D	D	D				D			D
Mercier et al. (2009)		D	D				D			D
Norton et al. (2010)		D	D				D		D	D
Stringer et al. (2013)		D	D				D			
Weber et al. (2009)		D	D							

on prevalence estimation, 6 were on methodological development, and 1 was on tuberculosis.

3.2. Evaluation of the selected papers based on the modified STARD statement

Seven of the 9 papers were reviewed further, while the following 2 were excluded: Alinovi et al. (2009) used maximum likelihood estimation of sensitivity and specificity rather than Bayesian methods, and another paper (Scott et al., 2010) lacked clarity about which samples were used in the different studies described in the publication. The seven papers were reviewed on whether the modified items (i.e., items 1, 4, 6, 10, 11, 12, 14, 22, 24 and 27) were adequately described or not. Subsequent to identification of disagreements between the reviewers (IG and SN), in each case the primary reasons for discrepant assessments were discussed (e.g., whether differences were due to a strict versus liberal interpretation of the criteria or whether there were more fundamental differences in their classification). Table 2 lists the items where disagreements occurred for the 7 selected papers. The items for which major disagreements occurred are described below to give insight into the primary reasons for discrepancies, which often pertained to different expectations about the level of detail required to meet the reporting goals of each item.

Item 4 (Study objectives and hypotheses): the study purpose was vaguely or implicitly defined. The purpose should link to the rationale, justification, and setting of the study.

Item 6 (Eligibility criteria): description of the source population with a complete description of inclusion and exclusion criteria also constitutes a challenge (e.g., incomplete or non-explicit descriptions, or lack of clarity).

Item 11 (Rationale for choosing the reference standard): the STARD guidelines require the rationale for using the selected reference standard to be described, but there is no reference standard in a latent class analysis. Item 11 in STARD-BLCM addresses the rationale for choosing the tests. The authors disagreed about whether the rationale was effectively provided for studies involving established tests that are accepted to have a diagnostic potential. However, they agreed that the rationale for a "new" test under evaluation should be provided.

Items 14 (Methods for estimating or comparing measures of diagnostic accuracy): this item was modified to explicitly describe reporting of the BLCM, the incorporation of prior information, and the results under alternative priors (see Section 4.3.3, Items 14a, 14b and Section 4.4.2, Item 24).

Item 27 (Implications for practice, including the intended use and clinical role of the index test): the assessment of the clinical utility of the tests under evaluation is subjective since it depends on issues such as cost, time to obtain test results, and the strengths and weaknesses of alternate tests. Often these broader considerations are not described by authors and the subjective perspective of peer reviewers is superimposed on perspectives of the authors. This led to disagreements particularly since the standard reporting requirements are not clear.

Variation in the author's assessments of the different reporting items serves as a clear example that the current "standards" under STARD2015 may not be interpreted in the same way in the case of BLCMs. We acknowledge that standardized reporting may result in the loss of the essentials, namely the justification – purpose – context – target-condition – rationale – utility complex (Fig. 1). The purpose is given by the justification for the evaluation, but it also provides an indication of the intended use. It does so while the target condition is defined from the purpose. Based on the target condition, we must find tests that at least in theory can detect the target condition, and hence we describe the rationale. The context, however, may deem one test more relevant than another, and therefore the different purposes can result in different relevant tests. All these factors need to be considered when the utility of the test is assessed, and consequently all are relevant to the interpretation. Care should be taken to describe and interpret the results for an appropriate use that relates to these items and this should be explicitly described in the paper.

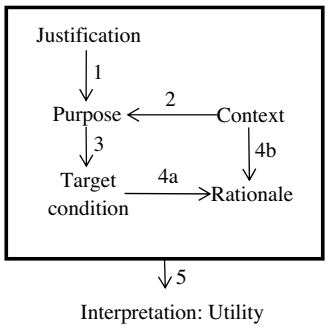

Fig. 1. The justification – purpose – context – target-condition – rationale – utility complex. Testing can for example be justified by a production loss. The purpose would therefore typically aim to reduce this loss, and the target condition might be any condition where the loss is present or can be predicted to occur. The context may dictate the relevance of a specific target condition, while other target conditions can simultaneously be of interest, or because the prevalence is really high or really low. The rationale for testing thus needs to be evaluated in the context of such preconditions. Ultimately, the utility of the result is maximized if the interpretation and subsequent decision following a test-result are seen in light of this complex.

The review in the case of MAP – an infection where BLCMs have been broadly applied – and the identified discrepancies in the opinions of the reviewing authors demonstrates clearly the need for the proposed modifications.

4. STARD-BLCM: Standards for the Reporting of Diagnostic accuracy studies that use Bayesian Latent Class Models

In this section, we present each item that was modified. The STARD-BLCM reporting criteria apply to any setting where a perfect reference test is not available. The STARD2015 items that were not modified are not discussed. Some items of STARD2015 also cover the needs under BLCM and we only changed the text to address the absence of a reference test without any further elaboration. A summary of the proposed modifications is in Table 1. With the exception of the items relating to the incorporation of prior information the proposed modifications pertain beyond Bayesian estimation (e.g., maximum likelihood). The split of items 10, 12, 13 and 21 of STARD2015 into two subsections was not preserved as it referred separately to the reference (which is absent in BLCMs) and the index test. On the other hand, item 14 was split into subsections to accommodate the reporting of prior information under BLCMs.

Each item is followed by an example of proper reporting and further elaboration. The examples were identified through a literature search for BLCMs in the fields of medicine and veterinary medicine. Initially, a list of potential examples was identified and circulated among the authors. For each example the authors checked whether the reporting with respect to the item under consideration was adequate. Examples that were not found adequate were replaced and re-circulated among the authors. This process was repeated until all examples were unanimously approved.

4.1. Title or abstract

Item 1: Identification as a study of diagnostic accuracy, using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC) and Bayesian latent class models.

Example

"The objective of this study was to determine the sensitivity and specificity of an ELISA for detection of *C. burnetii* antibodies in milk and blood samples, using latent class models in a Bayesian analysis" (Paul et al., 2013).

Elaboration

Electronic databases are nowadays the primary resources for identification of relevant studies. Authors should ensure that appropriate MeSH terms are included in at least one of the 3 designated locations (title, abstract or keywords) to promote the retrieval of their work from Medline searches. Currently, "Diagnostic accuracy" is not a MeSH term as opposed to "sensitivity" and "specificity" (Bossuyt et al., 2003). The latter and "Bayesian analysis", which also is a valid MeSH term, if included in the title, abstract or keywords, facilitates identification of studies for the estimation of the sensitivity and specificity of diagnostic tests by the use of Bayesian analysis. In the example above, it is evident that animals with *C. burnetii* antibodies are the target, not those with *C. burnetii* infection. It could have been useful to also include the target population in the specification of the objective by adding, for example, "in cattle >2 years of age".

4.2. Introduction

Item 4: Study objectives and hypotheses, such as estimation of diagnostic accuracy of the tests for a defined purpose through BLCM.

Example

"The objective of this study was to use Bayesian LCMs to analyze existing data from a cohort of patients presenting to hospital with suspected dengue infection. We estimated the accuracy of three rapid diagnostic tests (Panbio NS1, IgM and IgG cassette tests), our reference assay for dengue infection, and the combination of all three rapid tests when used at clinical presentation." (Pan-ngum et al., 2013).

Elaboration

The diagnostic accuracy of the tests under evaluation depends on the purpose of testing (Gardner et al., 2011). The OIE endorses the use of tests to (1) demonstrate freedom from infection, (2) demonstrate freedom from infection or agent in individual animals or products for trade purposes, (3) eradicate infection, (4) confirm a diagnosis of clinical cases, (5) estimate prevalence of infection to facilitate risk analysis, and (6) determine immune status in individual animals or populations. Other test purposes may also be relevant in different circumstances, such as prognosis. An additional reason for explicitly specifying the conditions of the validation study is to set up the framework for the identification of the relevant studies/sources that can be used for elicitation/selection of prior information on the parameters of interest (i.e., Se and Sp of the tests) during Bayesian analysis (see also item 14b).

4.3. Materials and methods

4.3.1. Participants

Item 6: Eligibility criteria and description of the source population.

Example

"Between October 2006 and December 2009, all children aged 0.5–10 years admitted to Modilon Hospital, the provincial hospital to which the majority of children with severe illness are referred, were assessed for recruitment to an observational study of severe pediatric illness. Inclusion criteria included i) impaired consciousness (Blantyre Coma Score (BCS) $\leq$4, or $\leq$2 at 0.5, 1 or 6 h after correction of hypoglycemia, a seizure or parenteral anticonvulsant therapy, respectively), ii) prostration (inability to sit/stand unaided), iii) multiple seizures, iv) hyperlactatemia (blood lactate >5.0 mmol/L), v) severe anemia (hemoglobin <50 g/L), vi) dark urine, vii) hypoglycemia (blood glucose $\leq$2.2 mmol/L), viii) jaundice, ix) respiratory distress (deep breathing, inter-costal indrawing, sub-costal recession, persistent alar flaring, tracheal tug,

and/or respiratory rate >60/min), x) persistent vomiting, xi) abnormal bleeding, and/or xii) signs of shock. These criteria reflect the World Health Organization (WHO) definition of severe malarial illness Metabolic acidosis was defined as a plasma bicarbonate ≤12.2 mmol/L" (Manning et al., 2012).

Elaboration

Accuracy of the tests under evaluation varies between different populations mirroring the potential variation in the distribution of infection stages, strains and immune responses of the host within each population. For example, accuracy estimates from populations of high prevalence may not be relevant to populations of low prevalence because increased prevalence of infection can be associated with a greater burden of clinically affected individuals. As a result, tests may have increased Se when applied in high prevalence populations (Johnson et al., 2009) while Sp estimates are less affected by variations in the prevalence of infection. Hence, the accuracy of diagnostic tests cannot be extrapolated unconditionally to other populations that are not relevant to the source population (Greiner and Gardner, 2000). An explicit description of the source population is required. Furthermore, demographics of relevance to the target condition should preferably be included (e.g., for chronic infections, the age-distribution can be highly relevant).

4.3.2. Test methods

Item 10: Description of the tests under evaluation, in sufficient detail to allow replication, and/or cite references.

Example

"In all six studies included in our analysis, the HI [hemagglutination inhibition] test has been performed according to the protocol prescribed by the OIE Diagnostic Manual (i.e. using four hemagglutinin units of virus antigen and 1% chicken erythrocytes, diluted in PBS), and titres ≥1:16 were considered positive" (Comin et al., 2013).

Elaboration

When novel diagnostic methods are under consideration an explicit description of the test method is required. However, if the technical specifications have been previously published, this information can be referenced.

Item 11: Rationale for choosing the tests under evaluation in relation to their purpose.

Example

"We used the commercial indirect absorbed test ID Screen® (IDvet, Grabels, France) for detection of MAP specific antibodies and the AdiavetTM ParaTB realtime PCR (Adiagene, Paris, France) for detection of the *IS*900, both tests done according to the instructions of the manufacturers. Antibodies were detected because they are considered to be an indicator of progression of MAP infection, and the realtime PCR was used to detect animals excreting detectable amounts of MAP (Nielsen, 2014). The cut-offs recommended by the manufacturers were also used, but non-specific reactions were expected for both tests. Non-specific antibody reactions were considered possible due to occurrence of non-MAP mycobacterial cross-reactions, and non-specific detection of MAP based was considered, because *IS*900 like elements may occur in other mycobacteria and MAP may be excreted from animals where an infection is not established (Cousins et al., 1999)" (Osterstock et al., 2007).

Elaboration

The choice of the tests under evaluation and their biological principles should be discussed in connection with their intended use. This is done in order to set the background of their use (i) in relation to the targeted condition and (ii) to provide the necessary biological information relevant to the correlation between test results. Hence, a justification of the valid use of the tests under the specific assumptions underlying the Bayesian latent class modeling approach is required (for details see item 14a). In the example above, the reasoning for using an antibody test is given. Finally, caution is required in the selection of the evaluated tests so as to select the best possible predictor for the intended use while at the same time conditional dependence should be minimized by avoiding biologically similar tests (see item 14α2). For example, when the targeted condition is "infection" as is often the case in prevalence estimation surveys, tests that aim to identify all forms of infection should be preferred and their accuracy in identifying infection is of interest. On the other hand, the targeted condition of interest can be restricted to individuals that can transmit the infection, thus excluding cases where the infection is present but latent, when the focus lies in a disease control scheme aiming to avoid transmission (Nielsen and Toft, 2008).

Item 12: Definition of and rationale for test positivity cut-offs or result categories of the tests under evaluation, distinguishing pre-specified from exploratory.

Example

"Regarding the PathoProof™ Mastitis PCR Assay (Thermo Fisher Scientific, Vantaa, Finland); the thermal cycling protocol involved 40 cycles for the reaction. We used the recommended cut-off with Ct values ≤37 defined as positive (Bexiga et al., 2011)" (Mahmmod et al., 2013).

Elaboration

Test results on a continuous scale are often dichotomized based on a cut-off value. Continuous outcomes may also be categorized into more than two groups. Ideally, the actual continuous test outcomes should be analyzed since categorization of a continuous response (Thurmond et al., 2002), like the milk-ELISA for detection of antibodies against MAP, results in loss of valuable information. This is because the information conveyed in the test result is reduced to considering all positive results as equal, no matter how near or far they are from the cutoff. Hence, potential associations between the continuous outcome and risk factors or productivity indices, is attenuated or lost (Toft et al., 2005).

Bayesian models have been proposed to discriminate between non-infected and infected individuals based on continuous responses from one or two correlated tests (e.g., Choi et al., 2006a,b). In addition, methods have been developed that specifically allow for 2 infection stages (Jafarzadeh et al., 2010), and flexible nonparametric models have been developed that do not rely on potentially restrictive parametric assumptions (e.g., Branscum et al., 2005, 2015; Erkanli et al., 2006).

The loss in the discriminatory power due to lack of a perfect reference test can be minimized by the use of data where the true infection status for some individuals is determined by an additional conditionally-independent test (Branscum et al., 2008; Kostoulas et al., 2013). Moreover, a diagnostic interpretation approach that uses the continuous test outcomes to determine distributions by infection status and different risk profiles has also been proposed (Toft et al., 2005).

If the results of the tests under evaluation are to be dichotomized then optimal selection of cut-offs must take into account the assumed prevalence in the target population and the relative consequences of false-positive and false-negative test results (Greiner et al., 2000).

4.3.3. Analysis

Item 14a: BLCM model for estimating measures of diagnostic accuracy.

A complete specification of the BLCM, including a rationale for the selected prior distributions used in the primary and sensitivity analysis, is required. The appropriate modeling structure for esti-

mation of the diagnostic accuracy is implicitly determined by the underlying biological principles, intended use and the targeted conditions of the tests. Taking these into consideration, we describe the key elements that should all be present in the description of the BLCM and the underlying assumptions that must be explicitly stated.

14a.1. Example 1 – definition of infection

"In this case, when establishing a herd-level diagnosis of *M. bovis* from BTM samples, it is necessary to discuss the latent state, as the PCR test is limited to detection of free *M. bovis* DNA in a milk sample whereas the ELISA has the capability to detect antibodies directed against *M. bovis* secreted in the milk. This implies that the latent herd-level infection state in this study is defined by a situation with simultaneous presence of both target analytes in a BTM sample. These two events are not guaranteed to reflect the same latent state at animal-level, e.g. it is not known whether arthritis leads to *M. bovis* DNA in the milk. However, a BTM sample can be representative for the herd-level *M. bovis* infection status because it represents a combination of different animal-level infections" (Nielsen et al., 2015).

Elaboration

A description of the target condition of interest (e.g. infection, disease, infectiousness etc.) is required. LCMs, in conjunction with what the tests actually detect (e.g. organisms or immune responses to organisms), create their own definition of the latent infection status. Thus, a definition/interpretation of the latent infection status under consideration from a biological and health perspective is crucial so as to effectively communicate to the broader, non-expert, audience the contextual meaning of prevalence, Se and Sp. Essentially, this involves an explicit description of the target condition. A flow diagram can effectively illustrate how the tests under evaluation relate to the target condition, and can be particularly helpful when different tests may be targeting different latent variables, which are in turn inter-related. An example of use of such a diagram for validation of diagnostic tests for *Chlamydia trachomatis* can be found in Dendukuri et al. (2009).

14a.2. Example 2 – conditional dependence of tests.

"Since both tests measure the same biological principle (IgG antibody) we assumed that both tests were correlated to some extent. Indeed, the conditional dependence between these two tests had been already demonstrated when they were applied on pigs (Gardner et al., 2000). Thus, we used the conditional dependence model for two tests, two populations (Branscum et al., 2005), which allowed us to estimate the Se and Sp correlations (γSe and γSp) between both tests" (Mainar-Jaime and Barberán, 2007).

Elaboration

Dependence must be addressed in any BLCM analysis. For example, the commonly used two-test, two-population model with conditional dependence between tests lacks identifiability and Bayesian methods provide a natural approach if additional information is available for prior specification. Tests based on different biological principles might not be correlated to any substantial extent and, hence, are expected to be conditionally independent (Branscum et al., 2005). Conditional independence captures the notion that, conditional on disease status, knowledge of results of the first test does not modify our expectation about the outcome of the second test. For example, "infection" is usually a process that is biologically distinct from "immune response to the infection", and thus presence of pathogen and presence of antibodies may be regarded as conditionally independent.

A useful explanation of conditional dependence can be found in Gardner et al. (2000). The presence and extent of the dependence among the tests under evaluation determines the types of Bayesian analyses that are appropriate. Adjusting for dependence among tests is needed because estimates of diagnostic accuracy can be biased upwards if dependence is ignored. However, for tests with Se and Sp values close to 1, the impact of ignoring dependence may be negligible (Georgiadis et al., 2003).

There is a tradeoff between incorporating test dependence in a LCM and the resulting loss of identifiability (i.e. the degrees of freedom is smaller than the number of parameters to be estimated from the model). Omitting dependence terms that have a small effect on the final point and interval estimates may be the best choice of action. Models that allow for conditional dependencies among multiple tests are well established and should be adopted when tests are suspected to be conditionally dependent. Subsequently, estimates under these models and models that assume independence must be compared to assess whether the incorporated dependencies lead to noticeably different results. In case of appreciable differences, results from all models should be presented for transparency.

14a.3. Example 3 – constant accuracy of tests across populations.

"In order to verify that the test properties (Se, Sp) were constant throughout all populations, we repeated the analysis with exclusion of each of the 6 herds, one at a time. This did not cause substantial changes to the estimates, which supports that the assumption was not violated" (Mahmmod et al., 2013).

Elaboration

Constant accuracy is a crucial assumption in most latent class analyses that model cross-classified test results. This assumption implies that a homogeneous distribution of the various infection stages exists among the different populations. If not, Se and Sp may vary among subpopulations thus violating the assumption of constant Se and Sp. If there are differences in Se and Sp across subpopulations, the differences should be modeled appropriately. This is accomplished by allowing test accuracy parameters to vary across subpopulations (Bermingham et al., 2015) and by incorporating additional prior information, if possible, that would be needed to mitigate the lack of identifiability (Johnson et al., 2009). For example, with two correlated tests and two populations with differing test accuracy across populations, there would be 10 parameters and only 6 degrees of freedom. If pooled samples are used, then Se and Sp depend on the pool size and models that assume different sensitivities in samples with different pool sizes should be used (Dhand et al., 2010).

14a.4. Example 4 – distinct difference in the prevalence of the target condition across subpopulations.

"The test performance of the ELISA was evaluated using Bayesian latent class analysis based on two dependent tests and two populations as described by Branscum et al. (2005). For evaluation purposes, a subset of samples from 2011 (n = 242) were grouped into subpopulation 1 [wild boars aged $\leq$12 months (n = 143)] and subpopulation 2 [wild boars aged >12 months (n = 99)]. Based on numerous reports of an age-dependent *T. gondii* prevalence in a wide range of host species, the prevalence (π1) in subpopulation 1 was expected to be lower compared to the prevalence (π2) of subpopulation 2" (Wallander et al., 2015).

Elaboration

BLCMs applied to data from two tests require distinct differences in the prevalence of infection among sampled subpopulations. This may be challenging as the prevalence of infection is often not known in advance. The greater the difference in the prevalence between the subpopulations, the greater the precision that is achieved for the estimated Se, Sp and prevalence of infection (Toft et al., 2005). However, while large differences in prevalences between subpopulations may possess the above-mentioned favorable property in terms of precision of estimation, a large difference in the prevalence among subpopulations may be due to different

mixtures of infection stages in these subpopulations. If this is the case, the assumption of constant test accuracy is violated (Greiner and Gardner, 2000). However, distinct differences in prevalence do not necessarily correspond to differences in the distribution of the infections stages.

14a.5. Example 5 – illustration of relation between evaluated tests and the latent infection status.

"We hypothesize that they can be classified into two types of tests measuring different latent variables, as illustrated in Fig. 1. We hypothesize that the LCR and PCR tests measure the DNA latent variable, which is in turn a proxy for the true disease status . . . We hypothesize that the DNAP and culture tests, on the other hand, measure the disease latent variable. Though the DNAP test measures DNA it does so at a much higher organism load. Thus, in practical terms, we considered DNAP closer to culture. The model in Fig. 1 implies that LCR and PCR are conditionally dependent within truly diseased and truly non-diseased latent classes." (Dendukuri et al., 2009).

Elaboration

Tests that are based on different biological principle may target different latent variables. These latent variables may in turn be measures of the latent disease status. If this is the case, appropriate models that adjust for the fact that distinct subsets of multiple tests are measuring distinct latent variables, should be employed (Dendukuri et al., 2009). In the abovementioned example the authors used a diagram to illustrate the relation between the evaluated tests and latent variables. From this presentation it was clear that two of the tests – LCR and PCR – measure the same underlying latent variable (i.e. DNA status), which is different from the latent variable of interest (*Chlamydia trachomatis* infection status) that was directly measured by the other two tests, the DNAP and culture. The diagrammatic presentation and associated explanation also made it clear that conditional dependence between the LCR and PCR tests existed and had to be modeled.

Item 14b: Definition and rationale of prior information and sensitivity analysis.

Example 1

"In consultation with a panel of experts from the McGill Centre for Tropical Diseases, we determined equally tailed 95 percent probability intervals (i.e., 2.5% in each tail) for the sensitivity and specificity of each test (see Table 5). These were derived from a review of the relevant literature and clinical opinion. The particular beta prior density for each test parameter was selected by matching the center of the range with the mean of the beta distribution . . . and matching the standard deviation of the beta distribution . . . with one quarter of the total range." (Joseph et al., 1995).

Example 2

Here, we provide an example of sound prior specification for the estimation of Se and Sp of an ELISA that detects antibodies against MAP and we discuss why this process should result in different prior choices for dairy cattle, sheep or goats.

A detailed review of the studies on the evaluation of antemortem tests for paratuberculosis in dairy cattle, sheep and goats was given by Nielsen and Toft (2008). Not all of these studies could serve as a source for specifying priors in either sheep or goats. Differences exist between cows, sheep and goats in the immune response (Corpa et al., 2000; Florou et al., 2009; Kostoulas et al., 2006a,b; Perez et al., 1999; Verna et al., 2007), the strain distribution and, hence, the infection stages and associated test performance. The naïve extrapolation of Se and Sp estimates among cattle, sheep and goats would result in a very poor, maybe even deceptive, prior specification. Estimates that are specific to sheep or to goats must be used for prior specification separately in sheep and in goats, respectively. Ideally, studies that adjust for all latent infection stages and do not lead to bias in the estimation of one or more parameters (e.g., Se and/or Sp) would be ideal (Kostoulas et al., 2006b; Mercier et al., 2009).

Elaboration

A major advantage of Bayesian analysis is the incorporation of substantive scientific (prior) information in the estimation process, which constitutes an efficient way of combining existing knowledge with the data at hand. Yet, by definition, it introduces subjectivity into the estimation process, and mis-specifying prior information can transform this major advantage into a drawback (Rothman et al., 2008).

The relative influence of the prior and data on posterior estimates depends on the prior precision and the strength of the data. For a given identifiable model structure, large datasets have a predominant influence on posterior distributions, while the same informative prior would have greater influence if the sample size were small. The impact of priors is always crucial in the case of non-identifiable models (Johnson et al., 2001; Jones et al., 2010). There is always lack of identifiability when the degrees of freedom is less than the number of unknown parameters in the model, and even when there are the same number of parameters and degrees of freedom, the model may still lack identifiability (Jones et al., 2010).

Prior elicitation is based on relevant and scientifically justifiable information that is obtained from data that are unrelated to the current data and/or based on the opinion of experts. If past data are used for prior specification, brief information about the data, including reference to published studies or reports and methodology for generating priors should be specified (Dhand et al., 2010). If expert opinion is used to generate priors, credentials of the experts providing the priors should be provided, in addition to the approach used for eliciting priors (Dhand et al., 2013). The names and experience of experts should be described for the benefit of readers unfamiliar with the disease of interest. Diffuse or standard reference priors (Christensen et al., 2010) could be used in identifiable models if both of the above approaches are not feasible or are deemed unreliable.

A complete Bayesian analysis involves the assessment of the impact of alternative prior information, including use of diffuse priors, on the final estimates. Regardless of the type of prior used, sensitivity analyses should be conducted and reported by changing input values by, for example, ±10–20% to evaluate the impact of priors on model outputs. Particular attention should be given to sensitivity analyses for input values about which the authors have low confidence. Sensitivity analysis demonstrates the possible dependence of inferences on the specified priors. When findings are appreciably different, we recommend discussion and/or presentation of the posterior inferences (e.g. median and 95% probability/credible intervals) under alternative priors (see item 24).

4.4. Results

4.4.1. Participants

Item 22: Time interval and any clinical interventions between the tests under evaluation.

Example

"Milk sampling was done by the farmers as part of the milk control scheme and the blood samples were taken by the herd veterinarians. The time interval between the milk and blood sampling was minimized and ranged from zero to three days with median one" (Paul et al., 2013).

Elaboration

Target variable bias is a serious issue that could arise when LCMs are applied in the analysis of data from acute infections. In the case of chronic and persistent infections, the pathogen, different types of biomarkers (i.e. antigens) and immune responses are generally

expected to be detectable throughout most of the course of infection. Hence, the latent class of infection is unambiguous regardless of the type of test used (Branscum et al., 2005). Depending on the type of the tests under consideration this may not hold in other occasions. For example, in acute infections the time period during which antigens and antibodies are both detectable is narrow. If BLCMs are based on the cross classified results of antigen- and antibody-based tests the infection status under consideration is limited to the time period during which both antigens and antibodies exist. The Se estimate for each test would be the proportion of individuals correctly classified as positive among the individuals that have detectable levels of both antigens and antibodies. Hence, caution is required when applying LCMs in such cases.

4.4.2. Test results

Item 24: Estimates of diagnostic accuracy under alternative prior specification and their precision (such as 95% credible/probability intervals).

Example

"Table 3 shows the posterior medians obtained with the different models used... Under the independence model when prior information was used for all parameters (model 1) posterior medians were... The results for the same model under the conditional dependence assumption (Table 3, model 5) were very similar with all estimates within ±1% of those found in model 1... Sensitivity analyses showed that estimates did not depend heavily on the priors as they varied slightly when non-informative priors were used... (Table 3)" (Mainar-Jaime et al., 2008).

Elaboration

Estimates of the diagnostic accuracy for the tests under evaluation using alternative priors must be reported: the median and/or the mean and 95% credible/probability intervals. If inconsequential differences in posterior estimates occur across different priors, this should be stated in the text. If substantial differences are observed, output that represents these differences should be placed in a table and should be discussed in the text, possibly alongside of results of the primary analysis. Bayesian estimates are dependent on priors. The extent of this dependence is influenced by the amount and quality of the data, the true Se and Sp values, and the validity of the assumptions underlying the LCM, including the difference in prevalences between subpopulations (Toft et al., 2005) and the structure of the model. Estimates from non-identifiable models will always depend on the specified priors. Presentation of posterior parameter estimates under alternative prior specifications as well as under diffuse priors is the preferred and clear indicator of the impact of prior information on posterior inference. The careful choice and justification of prior information (see also item 13) is critical if posterior estimates drastically change under different priors.

4.5. Discussion

Item 27: Implications for practice, including the intended use and clinical role of the tests under evaluation in various settings (clinical, research, surveillance etc.).

Example

"Using latent class analysis and published data, we estimated the accuracy of the HI test and six other diagnostic assays in detecting AIV [avian influenza virus] antibodies, without making reference to a gold standard. Because the HI test is commonly considered the gold standard for type-specific AIV antibody detection, its performance has rarely been questioned. Compared to the only previous study in which the accuracy of the HI test was estimated for poultry (Yamamoto et al., 2007) we found a similar Se and a much higher Sp, as well as much narrower credible intervals. This comparison might seem unfair, because we included the data of the previous study in our model. However, according to a sensitivity analysis (data not shown), the estimated Se and Sp of the HI test remained basically unvaried when excluding the study from the analysis, suggesting that the data from such study are consistent with those of the other studies" (Comin et al., 2013).

Elaboration

A discussion about the diagnostic accuracy of the tests should take account of the specified settings, which are the target population and the purpose of testing, at the individual or the population level (Gardner et al., 2011). Authors should provide a thorough consideration of their findings, from a biological perspective, and the implications of their results about the distribution of the targeted infection stages in the population under study. Authors could also highlight the differences between their results and comparable estimates from studies that were based on a "perfect reference" test, where differences may be due to the fact that Bayesian analysis adjusts for all latent infection stages as opposed to many reference tests that fail to detect individuals at the early infection stages, despite the fact that they are considered to be perfect. Finally, the importance of prior information on the estimated posteriors should be discussed.

5. Statistical/methodological considerations of Bayesian modeling

The following section provides information on the statistical/methodological presentation of any Bayesian modeling. These are not specific to BLCMs and are given here as a reminder of the essential elements that any proper presentation of Bayesian models should include.

5.1. Model description

All Bayesian models, including LCMs, must be described with proper mathematical notation. Ideally, the corresponding software codes should be made available, preferably with step-by-step explanations so that interested readers can understand and apply the models in their work. In addition, the mathematical description of the model should be accompanied by a clear explanation of the biological phenomenon that is captured by each modeling component (i.e. the description of why conditional dependencies are introduced and what these dependencies mean for tests that are based on the same biological principle).

5.2. Model selection criteria and goodness-of-fit tests

Comparisons between alternative models that are fit to the same data can be based on global measures of comparative model fit, such as the deviance information criterion, Bayes factors or pseudo Bayes factors based on pseudo marginal likelihoods (Geisser and Eddy, 1979; Branscum et al., 2015). These measures of comparative model fit dictate which model best fits the data at hand. They do not indicate whether the best among the competing models has an adequate fit to the data. This must be assessed by the use of goodness-of-fit procedures, which can be based on the comparison between the observed data and the predictions under the model (Gelman et al., 2014). However, there are no formal tests of goodness-of-fit for unidentified models and in a saturated model, where the number of degrees of freedom equals the number of parameters, observed and predicted values are identical.

5.3. Convergence diagnostics

BLCMs are often fit using Markov-chain Monte Carlo (MCMC) simulation from a joint posterior distribution. Convergence diagnostics of MCMC samples are not foolproof and a combination of

them plus visual inspection of the posteriors should be performed. For instance, the Gelman and Rubin diagnostic can indicate that convergence has occurred for a bimodal posterior distribution. An extensive discussion about convergence of an MCMC chain can be found in Toft et al. (2005).

5.4. Software

The software used in the analysis must be explicitly stated and appropriately cited.

5.5. Appendix

We are in favor of making the software codes corresponding to the models presented available, preferably with detailed explanation.

6. Conclusion

In contrast to studies designed to evaluate the accuracy of tests when the disease status is known, latent class methods are characterized by the absence of a reference test, the incorporation of properly justified prior information, the need for explicit definition of the condition that the tests under evaluation are targeting, and the complexity of the statistical models. The reporting requirements for these key elements are addressed herein and are termed under the modified STARD-BLCM guidelines.

Conflict of interest

The authors declare that they have nothing to declare.

Acknowledgement

We thank the Canada Excellence Research Chairs (CERC) program for funding the contribution of one of us (IG).

References

Angelidou, E., Kostoulas, P., Leontides, L., 2014. Bayesian validation of a serum and milk ELISA for antibodies against *Mycobacterium avium* subspecies paratuberculosis in Greek dairy goats across lactation. J. Dairy Sci. 97 (2), 819–828.

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.

Bermingham, M.L., Handel, I.G., Glass, E.J., Woolliams, J.A., de Clare Bronsvoort, B.M., McBride, S.H., Skuce, R.A., Allen, R.A., McDowell, S.W.J., Bishop, S.C., 2015. Hui and Walter's latent-class model extended to estimate diagnostic test properties from surveillance data: a latent model for latent data. Sci. Rep. 5.

Bexiga, R., Koskinen, M.T., Holopainen, J., Carneiro, C., Pereira, H., Ellis, K.A., Vilela, C.L., 2011. Diagnosis of intramammary infection in samples yielding negative results or minor pathogens in conventional bacterial culturing. J. Dairy Res. 78, 49–55.

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., Lijmer, J.G., 2003. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann. Intern. Med. 138, W1–12.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L., Lijmer, J.G., Moher, D., Rennie, D., de Vet, H.C.W., Kressel, H.Y., Rifai, N., Golub, R.M., Altman, D.G., Hooft, L., Korevaar, D.A., Cohen, J.F., For the STARD Group, 2015. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Radiology 277 (3), 826–832.

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

Branscum, A.J., Johnson, W.O., Hanson, T.E., Gardner, I.A., 2008. Bayesian semiparametric ROC curve estimation and disease diagnosis. Stat. Med. 27, 2474–2496.

Branscum, A.J., Johnson, W.O., Hanson, T.E., Baron, A.T., 2015. Flexible regression models for ROC and risk analysis: with or without a gold standard. Stat. Med. 34, 3997–4015.

Broemeling, L.D., 2007. Bayesian Biostatistics and Diagnostic Medicine. Chapman & Hall, Boca Raton, FL, U.S.A.

Choi, Y.K., Johnson, W.O., Thurmond, M.C., 2006a. Diagnosis using predictive probabilities without cut-offs. Stat. Med. 25, 699–717.

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

Christensen, R., Johnson, W.O., Branscum, A.J., Hanson, T.E., 2010. Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians. CRC Press, Boca Raton, Florida.

Collins, J., Huynh, M., 2014. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. Stat. Med. 33, 4141–4169.

Comin, A., Toft, N., Stegeman, A., Klinkenberg, D., Marangon, S., 2013. Serological diagnosis of avian influenza in poultry: is the haemagglutination inhibition test really the 'gold standard'? Influenza Other Resp. Viruses 7, 257–264.

Corpa, J.M., Garrido, J., Marin, J.G., Perez, V., 2000. Classification of lesions observed in natural cases of paratuberculosis in goats. J. Comp. Pathol. 122, 255–265.

Cousins, D.V., Whittington, R., Marsh, I., Masters, A., Evans, R.J., Kluver, P., 1999. Mycobacteria distinct from *Mycobacterium avium* subsp. *paratuberculosis* isolated from the faeces of ruminants possess IS900-like sequences detectable IS900 polymerase chain reaction: implications for diagnosis. Mol. Cell. Probes 13, 431–442.

Dendukuri, N., Hadgu, A., Wang, L., 2009. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. Stat. Med. 28, 441–461.

Dhand, N.K., Johnson, W.O., Toribio, J.A.L., 2010. A Bayesian approach to estimate OJD prevalence from pooled fecal samples of variable pool size. J. Agric. Biol. Environ. Stat. 15, 452–473.

Dhand, N.K., Johnson, W.O., Eppleston, J., Whittington, R.J., Windsor, P.A., 2013. Comparison of pre-and post-vaccination ovine Johne's disease prevalence using a Bayesian approach. Prev. Vet. Med. 111, 81–91.

Enøe, C., Georgiadis, M.P., Johnson, W.O., 2000. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Prev. Vet. Med. 45 (1), 61–81.

Erkanli, A., Sung, M., Jane Costello, E., Angold, A., 2006. Bayesian semiparametric ROC analysis. Stat. Med. 25, 3905–3928.

Florou, M., Leontides, L., Kostoulas, P., Billinis, C., Sofia, M., 2009. Strain-specific sensitivity estimates of *Mycobacterium avium* subsp. *paratuberculosis* culture in Greek sheep and goats. Zoonoses Public Health 56, 49–52.

Fosgate, G.T., Osterstock, J.B., Benjamin, L.A., Dobek, G.L., Roussel, A.J., 2009. Preliminary investigation of a humoral and cell-mediated immunity ratio for diagnosis of paratuberculosis in beef cattle. Prev. Vet. Med. 91, 226–233.

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

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

Geisser, S., Eddy, W.F., 1979. A predictive approach to model selection. J. Am. Stat. Assoc. 74, 153–160.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2014. Bayesian Data Analysis. Chapman & Hall/CRC, Boca Raton, FL, USA.

Georgiadis, M.P., Johnson, W.O., Gardner, I.A., Singh, R., 2003. Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. J. R. Stat. Soc. Ser. C (Appl. Stat.) 52, 63–76.

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

Greiner, M., Pfeiffer, D., Smith, R.D., 2000. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Prev. Vet. Med. 45, 23–41.

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

Jafarzadeh, S.R., Johnson, W.O., Utts, J.M., Gardner, I.A., 2010. Bayesian estimation of the receiver operating characteristic curve for a diagnostic test with a limit of detection in the absence of a gold standard. Stat. Med. 29, 2090–2106.

Johnson, W.O., Gastwirth, J.L., Pearson, L.M., 2001. Screening without a gold standard: the Hui-Walter paradigm revisited. Am. J. Epidemiol. 153, 921–924.

Johnson, W.O., Gardner, I.A., Metoyer, C.N., Branscum, A.J., 2009. On the interpretation of test sensitivity in the two-test two-population problem: assumptions matter. Prev. Vet. Med. 91, 116–121.

Jones, G., Johnson, W.O., Hanson, T.E., Christensen, R., 2010. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. Biometrics 66, 855–863.

Joseph, L., Gyorkos, T.W., Coupal, L., 1995. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am. J. Epidemiol. 141, 263–272.

Kostoulas, P., Leontides, L., Billinis, C., Florou, M., 2006a. Application of a semi-dependent latent model in the Bayesian estimation of the sensitivity and specificity of two faecal culture methods for diagnosis of paratuberculosis in sub-clinically infected Greek dairy sheep and goats. Prev. Vet. Med. 76, 121–134.

Kostoulas, P., Leontides, L., Enoe, C., Billinis, C., Florou, M., Sofia, M., 2006b. 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.

Kostoulas, P., Browne, W.J., Nielsen, S.S., Leontides, L., 2013. Bayesian mixture models for partially verified data: age- and stage-specific discriminatory power of an antibody ELISA for paratuberculosis. Prev. Vet. Med. 111, 200–205.
Lunn, D., Spiegelhalter, D., Thomas, A., Best, N., 2009. The BUGS project Evolution, critique and future directions. Stat. Med. 28, 3049–3067.
Mahmmod, Y.S., Toft, N., Katholm, J., Grønbæk, C., Klaas, I.C., 2013. Bayesian estimation of test characteristics of real-time PCR, bacteriological culture and California mastitis test for diagnosis of intramammary infections with *Staphylococcus aureus* in dairy cattle at routine milk recordings. Prev. Vet. Med. 112, 309–317.
Mainar-Jaime, R.C., Barberán, M., 2007. Evaluation of the diagnostic accuracy of the modified agglutination test (MAT) and an indirect ELISA for the detection of serum antibodies against *Toxoplasma gondii* in sheep through Bayesian approaches. Vet. Parasitol. 148, 122–129.
Mainar-Jaime, R.C., Atashparvar, N., Chirino-Trejo, M., 2008. Estimation of the diagnostic accuracy of the invA-gene-based PCR technique and a bacteriological culture for the detection of salmonella spp. in caecal content from slaughtered pigs using bayesian analysis. Zoonoses Public Health 55, 112–118.
Manning, L., Laman, M., Rosanas-Urgell, A., Turlach, B., Aipit, S., Bona, C., Warrell, J., Siba, P., Mueller, I., Davis, T.M., 2012. Rapid antigen detection tests for malaria diagnosis in severely ill Papua New Guinean children: a comparative study using Bayesian latent class models. PLoS One 7, e48701.
Mercier, P., Beaudeau, F., Laroucau, K., Bertin, C., Boschiroli, M.-L., Baudry, C., Seegers, H., Malher, X., 2009. Comparative age-related responses to serological and faecal tests directed to *Mycobacterium avium paratuberculosis* (Map) in French dairy goats. Small Rumin. Res. 87, 50–56.
Nielsen, S.S., Toft, N., 2008. Ante mortem diagnosis of paratuberculosis: a review of accuracies of ELISA, interferon- + ¦ assay and faecal culture techniques. Vet. Microbiol. 129, 217–235.
Nielsen, P.K., Petersen, M.B., Nielsen, L.R., Halasa, T., Toft, N., 2015. Latent class analysis of bulk tank milk PCR and ELISA testing for herd level diagnosis of *Mycoplasma bovis*. Prev. Vet. Med. 121, 338–342.
Nielsen, S.S., 2014. Developments in diagnosis and control of bovine paratuberculosis. CAB Rev. 9 (012), 1–12.
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.
(OIE (World Organisation for Animal Health), 2016. Manual of Diagnostic Tests and Vaccines for Terrestrial Animals. Chapter 1.1.6 – Principles and Methods of Validation of Diagnostic Assays for Infectious Diseases, Available at http://www.oie.int/en/international-standard-setting/terrestrial-manual/access-online/ (Accessed 15 December 2016).
Osterstock, J.B., Fosgate, G.T., Norby, B., Manning, E.J., Collins, M.T., Roussel, A.J., 2007. Contribution of environmental mycobacteria to false-positive serum ELISA results for paratuberculosis. J. Am. Vet. Med. Assoc. 230, 896–901.
Pan-ngum, W., Blacksell, S.D., Lubell, Y., Pukrittayakamee, S., Bailey, M.S., de Silva, H.J., Lalloo, D.G., Day, N.P.J., White, L.J., Limmathurotsakul, D., 2013. Estimating the true accuracy of diagnostic tests for dengue infection using bayesian latent class models. PLoS One 8, e50765.
Paul, S., Toft, N., Agerholm J. r.S. Christoffersen, A.B., Agger, J.F., 2013. Bayesian estimation of sensitivity and specificity of *Coxiella burnetii* antibody ELISA tests in bovine blood and milk. Prev. Vet. Med. 109, 258–263.
Perez, V., Tellechea, J., Corpa, J.M., Gutierrez, M., García-Marin, J.F., 1999. Relation between pathologic findings and cellular immune responses in sheep with naturally acquired paratuberculosis. Am. J. Vet. Res. 60, 123–127.
Rothman, K.J., Greenland, S., Lash, T.L., 2008. Modern Epidemiology. Lippincott Williams & Wilkins.
Scott, M.C., Bannantine, J.P., Kaneko, Y., Branscum, A.J., Whitlock, R.H., Mori, Y., Speer, C.A., Eda, S., 2010. Absorbed EVELISA: a diagnostic test with improved specificity for Johne's disease in cattle. Foodborne Pathog. Dis. 7, 1291–1296.
Stringer, L.A., Jones, G., Jewell, C.P., Noble, A.D., Heuer, C., Wilson, P.R., Johnson, W.O., 2013. Bayesian estimation of the sensitivity and specificity of individual fecal culture and Paralisa to detect *Mycobacterium avium* subspecies *paratuberculosis* infection in young farmed deer. J. Vet. Diagn. Invest. 25, 759–764.
Thurmond, M.C., Johnson, W.O., Muñoz-Zanzi, C.A., Su, C.L., Hietala, S.K., 2002. A method of probability diagnostic assignment that applies Bayes theorem for use in serologic diagnostics, using an example of *Neospora caninum* infection in cattle. Am. J. Vet. Res. 63, 318–325.
Toft, N., Nielsen, S.S., Jorgensen, E., 2005. Continuous-data diagnostic tests for paratuberculosis as a multistage disease. J. Dairy Sci. 88, 3923–3931.
van Smeden, M., Naaktgeboren, C.A., Reitsma, J.B., Moons, K.G., de Groot, J.A., 2013. Latent class models in diagnostic studies when there is no reference standard—a systematic review. Am. J. Epidemiol., kwt286.
Verna, A.E., Garcia-Pariente, C., Muñoz, M., Moreno, O., García-Marin, J.F., Romano, M.I., Paolicchi, F., Perez, V., 2007. Variation in the immunoñpathological responses of lambs after experimental infection with different strains of *Mycobacterium avium* subsp. *paratuberculosis*. Zoonoses Public Health 54, 243–252.
Wallander, C., Frössling, J., Vågsholm, I., Uggla, A., Lunden, A., 2015. *Toxoplasma gondii* seroprevalence in wild boars (Sus scrofa) in Sweden and evaluation of ELISA test performance. Epidemiol. Infect. 143, 1913–1921.
Walter, S.D., Irwig, L.M., 1988. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J. Clin. Epidemiol. 41, 923–937.
Weber, M.F., Verhoeff, J., van Schaik, G., van Maanen, C., 2009. Evaluation of Ziehl-Neelsen stained faecal smear and ELISA as tools for surveillance of clinical paratuberculosis in cattle in the Netherlands. Prev. Vet. Med. 92, 256–266.
Yamamoto, T., Tsutsui, T., Nishiguchi, A., Kobayashi, S., Tsukamoto, K., Saito, T., Mase, M., Okamatsu, M., 2007. Preliminary evaluation of diagnostic tests for avian influenza using the Markov chain Monte Carlo (MCMC) method in an emergency surveillance. J. Vet. Med. Sci. 69, 673–675.