Rev. Sci. Tech. Off. Int. Epiz., 2021, **40** (1), 287-298

Examples of proper reporting for evaluation (Stage 2 validation) of diagnostic tests for diseases listed by the World Organisation for Animal Health

P. Kostoulas [(1) *], I.A. Gardner [(2)], M.C. Elschner [(3)], M.J. Denwood [(4)], E. Meletis [(1)] & S.S. Nielsen [(4)]

(1) Laboratory of Epidemiology, Faculty of Public Health, University of Thessaly, Mavromichali Street, Karditsa, 43100, Greece
(2) Atlantic Veterinary College, University of Prince Edward Island, 550 University Avenue, Charlottetown, Prince Edward Island, C1A 4P3, Canada
(3) Friedrich-Loeffler-Institut (FLI), Federal Research Institute for Animal Health, Institute of Bacterial Infections and Zoonoses, Naumburger Strasse 96a, 07743 Jena, Germany
(4) Department of Veterinary and Animal Sciences, University of Copenhagen, Grønnegårdsvej 8, 1870 Frederiksberg C, Denmark
* Corresponding author: pkost@vet.uth.gr

Summary

Reporting and design standards are key indicators of the quality of diagnostic accuracy (validation) studies but, with the exception of aquatic animal diseases and paratuberculosis in ruminants, there is limited guidance for designing these studies in animals. There is, therefore, a need for generic guidelines that are based on disease characteristics, such as mode of transmission, latent period and pathogenesis. Comprehensive, clear and transparent reporting of primary test accuracy studies for diseases listed by the World Organisation for Animal Health (OIE) has value for the end users of diagnostic tests and, ultimately, for decision-makers, who require systematic reviews and meta-analysis of multiple tests for specified diseases and testing purposes. The recent publication of reporting standards for Bayesian latent class models, to analyse test accuracy data from naturally occurring disease events, fills an important gap as these methods are being increasingly used for OIE-listed diseases. Adherence to design and reporting standards, as well as to guidelines, helps to ensure that research funding for test validation studies is used appropriately and that the strengths and limitations of single tests or test combinations are made clear to test users. The authors provide a review of key points that are often overlooked or misinterpreted in test validation studies, as well as two concrete examples of good practice for use as a reference point for future studies.

Keywords

Aquatic animals – Bayesian latent class model – Diagnostic sensitivity – Diagnostic specificity – Infectious diseases – Terrestrial animals – Test accuracy – Test validation – Validation pathway – Wild mammals – World Organisation for Animal Health guidelines.

Introduction

The World Organisation for Animal Health (OIE) Validation Pathway, described in the OIE *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals* (*Terrestrial Manual*), provides a four-stage template to assess the fitness of an assay for an intended purpose. Such validation includes estimates of the analytical and diagnostic performance characteristics of the test. Stage 2 of the Pathway, 'Diagnostic performance of the assay', describes diagnostic accuracy studies (DAS), which enable the estimation of diagnostic sensitivity (Se) and specificity (Sp), in addition to likelihood ratios (LRs). These studies are both time-consuming and resource-intensive (1). In principle, they should be guided by careful planning that minimises important biases (2), tests sufficient representative specimens for the target analyte of interest (e.g. genomic material or serum antibodies) and

https://doi.org/10.20506/rst.40.1.3225

the specified testing purpose, and generates results that are applicable to multiple countries, wherever possible. Ideally, laboratories conducting the testing should be operating under a quality management system (3, 4) and be certified by national or international standard-setting organisations.

Funding for DAS in animals during naturally occurring disease events is often limited unless the disease is zoonotic and, in many cases, DAS are based on existing samples (obtained from an experimental challenge study on disease pathogenesis or sample repositories with specimens of known infection status). Large-scale, collaborative, multi-centre validation studies using the same set of specimens are infrequently carried out for animal diseases. Furthermore, experimental studies or estimates from historical data may not be representative of the current field situation (e.g. pathogenesis and epidemiology). Therefore, DAS should be based on samples that represent the target population, rather than on samples naively extrapolated from settings and sampling schemes that may not be relevant.

Within this context, guidelines on the proper design and reporting of DAS are essential to ensure explicit descriptions of the information required to assess the relevance of the diagnostic accuracy estimates. Key points include the explicit description of: *(a)* the index test(s), to allow replication; and *(b)* its (their) intended use and clinical role, as well as implications for practice.

The objective of this paper is to describe the current status of reporting and design standards, explain guidelines that can be used for OIE-listed animal diseases and present concrete examples of the proper implementation of key elements of those guidelines, to encourage prospective authors/users to adhere to them during the planning, conduct and reporting of DAS.

Reporting standards and guidance for authors

Complete and transparent reporting of DAS, including a candid discussion of the strengths and limitations of the study design, is essential to ensure that test users have a clear understanding of how the test is likely to perform in target animal populations. For instance, the target condition of the index test(s) must be explicitly described (i.e. antigen or antibody detection) to enable a critical assessment of the relevance of the estimated Se/Sp for different surveillance programmes and target populations. An explicit description of the target population is necessary to assess whether diagnostic accuracy estimates are relevant and useful for other populations. The **STA**ndards for **R**eporting of **D**iagnostic accuracy (STARD)–2015 checklist of 30 items (Table I) was an important step towards completeness and transparency of DAS in medicine (5, 6, 7). The 30 items are categorised into sections (title, abstract, methods, results,

Table I
Checklist of items listed by the Standards for Reporting of Diagnostic Accuracy 2015 (STARD–2015) and the STARD – Bayesian latent class (mixture) models (BLCM) (references for both)
Modifications of the STARD–BLCM are in **bold**

Section & topic	Item	STARD–2015	STARD–BLCM
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 area under the curve (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 (BCLM)**
Abstract	2	Structured summary of study design, methods, results and conclusions (for specific guidance, see STARD for Abstracts) (5)	Unmodified
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** [(a)]
	4	Study objectives and hypotheses	Study objectives and hypotheses, **such as estimation of diagnostic accuracy of the tests for a defined purpose through BLCMs**
Methods			
Study Design	5	Whether data collection was planned and the reference standard performed before the index test (prospective study) or after (retrospective study)	Whether data collection was planned before the **tests** were performed (prospective study) or after (retrospective study)
Participants	6	Eligibility criteria	Eligibility criteria **and description of the source population**
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in the registry)	Unmodified

Table I (cont.)

Section & topic	Item	STARD–2015	STARD–BLCM
	8	Where and when potentially eligible participants were identified (setting, location and dates)	Unmodified
	9	Whether participants formed a consecutive, random or convenience series	Unmodified
Test methods	10 (a)	Index test, in sufficient detail to allow replication	**Description of the tests under evaluation,** in sufficient detail to allow replication, **and/or cite references**
	10 (b)	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**
	12 (a)	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory [i.e. whether these cut-offs or result categories were defined prior to the study or after collecting the data (6)]	Definition of and rationale for test positivity cut-offs or result categories of the **tests under evaluation**, distinguishing pre-specified from exploratory
	12 (b)	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory [i.e. whether these cut-offs or result categories were defined prior to the study or after collecting the data (6)]	
	13 (a)	Whether clinical information and reference standard results were available to the performers or readers of the index test	Whether clinical information was available to the performers or readers of the **tests under evaluation**
	13 (b)	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	**14 (a): BLCM model** for estimating measures of diagnostic accuracy
			14 (b): Definition and rationale of prior information and sensitivity analysis
	15	How indeterminate index test or reference standard results were handled	How indeterminate results **of the tests under evaluation** were handled
	16	How missing data for the index test and reference standard were handled	How missing data **for the tests under evaluation** were handled
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory [i.e. whether these analyses were defined prior to the study or after collecting the data (6)]	Unmodified
	18	Intended sample size and how it was determined	Unmodified
Results			
Participants	19	Flow of participants, using a diagram	Unmodified
	20	Baseline demographic and clinical characteristics of participants	Unmodified
	21 (a)	Distribution of severity of disease in those with the target condition	**Not applicable: the distribution of the targeted conditions is unknown, hence the use of BLCM**
	21 (b)	Distribution of alternative diagnoses in those without the target condition	
	22	Time interval and any clinical interventions between the index test and reference standard	Time interval and any clinical interventions between **the tests under evaluation**
Test results	23	Cross-tabulation of the index test results (or their distribution) by the results of the reference standard	Cross-tabulation of the **test** results (or, **for continuous test results,** their distribution **by infection stage)**
	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% **credibility/probability** intervals)
	25	Reports of any adverse events from performing the index test or reference standard	Report any adverse events from performing the **tests under evaluation**
Discussion	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	Unmodified
	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.)**

Table I (cont.)

Section & topic	Item	STARD–2015	STARD–BLCM
Other information			
	28	Registration number and name of registry	Unmodified
	29	Where the full study protocol can be accessed	Unmodified
	30	Sources of funding and other support; role of funders	Unmodified

a) In STARD–2015 the terms 'index (first) test' and 'reference standard' are used but the terms 'tests under evaluation' and 'candidate tests' are more generic and more suitable for a Bayesian latent class (mixture) model

discussion and other information) that follow the structure of a manuscript. This first checklist, STARD–2015, was written for test accuracy studies in human patients but adaptations of STARD have been made for paratuberculosis (8) and infectious diseases of aquatic animals (9), to deal with aspects of terminology, populations as epidemiological units, and the use of experimental challenge studies for some diseases. The STARD–2015 checklist was recently modified to address the use of Bayesian latent class (mixture) models (BLCMs) for DAS that are not based on a perfect reference test but take account of the imperfect diagnostic test performance and data misclassification, and simultaneously estimate the diagnostic Se and Sp of all tests under evaluation (10). The STARD–BLCM checklist applies equally to all latent class models for DAS – whether these are implemented within a Bayesian or frequentist framework – but, in practice, the majority of such studies use a Bayesian method of fitting the models. Therefore, STARD–BLCM also covers additional reporting requirements specific to Bayesian methods, e.g. the inclusion of prior information.

Several authors have previously recommended that veterinary journal editors, reviewers and authors should adhere to reporting standards for DAS studies and submit checklists with page numbers indicating where each item is addressed. Experience in human health indicates that poorly designed studies are often incompletely reported and the same is probably true for animal diseases. Incomplete reporting can be corrected by authors, whereas many design flaws (e.g. spectrum-of-disease and selection bias) cannot. That is why reporting standards should ideally be consulted during the design of DAS, from the beginning. In this way, common pitfalls that will eventually lead to poor reporting can be avoided. The Meridian Network website, hosted by Iowa State University (11), lists documents on reporting standards for diagnostic accuracy studies and has downloadable templates.

The authors provide two examples of DAS below, which they believe to have reported their methods and findings appropriately. The first example was based on a perfect reference test and used the STARD guidelines, and the second used latent class models to simultaneously estimate the diagnostic accuracy of all tests, as well as the STARD–BLCM checklist. The authors' intention is to provide concrete examples of good practice as a reference point for future studies.

Example of good practice adhering to STARD–2015: a test accuracy comparison study of seven serological assays for glanders

Glanders, a contagious and predominantly chronic infection of horses, mules and donkeys, is caused by the Gram-negative bacterium *Burkholderia mallei*. Serological testing is required for the international movement of equids or their trade because latently infected animals may introduce *B. mallei* infection into previously non-infected countries. Hence, high Se is an important prerequisite for any diagnostic test or combination of tests. Elschner *et al.* (12) reported estimates of Se, Sp and LR for seven serological tests, based on samples from 254 equids with glanders and 3,000 that were glanders-free.

The STARD–2015 checklist was used to guide preparation of the study report. In the following sections, the authors provide extracts from the paper (shown indented and in quotation marks) as an example of the reporting of some key items in the checklist. Annotations to clarify the meaning of the text are shown in square brackets. Some checklist items were either not applicable (Items 21 (b), 28 and 29) or not routinely used for OIE-listed animal diseases (Items 6, 9 and 19). The study did not require ethical approval as all blood samples were collected for diagnostic purposes or official monitoring studies. For brevity, the authors do not provide a description or example for each checklist item. Detailed explanations with examples of appropriate reporting for each of the STARD–2015 items, including comments, are available (6), not only for glanders but also for other animal diseases, such as paratuberculosis (8) and aquatic animal diseases (9).

The Introduction section

The **Introduction section** of STARD–2015 has two topics: Item 3 (scientific and clinical background) and Item 4 (study objectives and hypotheses). Both are important to establish the context and purpose of the research.

Item 3: Scientific and clinical background, including the intended use and clinical role of the index test

> 'CFT [complement fixation test] is the World Organisation for Animal Health (OIE) prescribed serodiagnostic method for international trade purposes and is also recommended for surveillance investigations. This test is known to have high sensitivity, but it gives a considerable number of false-positive results. These false-positive results cause unnecessary restrictions on international trade of animals… and result in financial losses for owners and the horse industry. Recently, new serological tests have been developed to overcome the disadvantages of the CFT… The test is difficult to standardise, as there are no standardised sera available. Furthermore, the CFT requires experienced operators, ongoing training and quality management systems to be implemented in the laboratory. The performance of the test is demanding, and the results depend on the antigen used and methods, such as incubation conditions.'

This overview gives the reader a background to the specific challenges involving the CFT test, and how these are relevant to the clinical use of the test.

Item 4: Study objectives and hypotheses

> '… this prospective study compared the diagnostic accuracy (sensitivity and specificity) of the WB [Western blot] technique, five indirect ELISAs [enzyme-linked immunosorbent assays] (iELISA) and the OIE-prescribed CFT for the serological detection of *B. mallei* antibodies in equids.'

This is a clearly defined and succinctly explained hypothesis, in which the verb 'compare' clarifies that the tests will be contrasted, and the population given, 'equids', highlights the epidemiological setting.

The Methods section

The **Methods section** has four topics (study design, participants, test methods and analysis), each with multiple items. Examples for Items 10 (a): index tests; 10 (b): the reference standard; and 14: methods for estimating and comparing measures of diagnostic accuracy; are provided below. (The sources of the materials have not been included in this extract.)

Item 10 (a): Index test, in sufficient detail to allow replication

> 'CFT was performed as described in the OIE manual. Briefly, serum samples were diluted 1:5 in CFT buffer, inactivated, and two fold dilutions of them were mixed with Malleus CFT antigen and 5 complement haemolytic units – 50% of guinea pig complement. Sera, complement and antigen were mixed in the plates and incubated overnight at 4°C. A 2% suspension of sensitized sheep red blood cells was added and plates were incubated for 45 minutes at 37°C and then centrifuged for 5 minutes at 600 g. Samples with 100% hemolysis in a dilution of 1:5 were categorized as negative, samples showing 25–75% hemolysis in a dilution of 1:5 were classified as suspicious and samples showing 100% inhibition of hemolysis in a dilution of 1:5 were classified as positive.'

In this case, it is sufficient to give a brief overview of the method and provide a reference for the full details (OIE *Manual*). If an external reference had not been available, then more details may have been required.

Item 10 (b): Reference standard, in sufficient detail to allow replication

'**Samples from glanderous animals**

> True positive serum samples (n = 254) were collected from equids [226, 23 and 5 sera from horses, mules and donkeys, respectively]… in which the infection was confirmed by clinical signs (n = 112), and by *B. mallei* isolation or molecular detection of *B. mallei* by real-time PCR [polymerase chain reaction] as recommended in the OIE manual… (n = 142). The animals were considered to be "clinically positive" (n = 112) on the basis of the presence of at least one clinical sign consistent with glanders, and the fact that they were detected during a culture-confirmed glanders outbreak in this population including close contact to infected animals.'
>
> Country information was provided but 'no further demographic data [other than species] were available'.

The text gives a justification of the target condition. Additionally, the population investigated is well characterised. Depending on the pathogenesis and target population, further information may be required, e.g. with regards to age distribution, if age is a risk factor.

Item 14: Methods for estimating or comparing measures of diagnostic accuracy

> 'Ninety-five percent confidence intervals for sensitivity, specificity and likelihood ratios were based on standard formulas [reference given] and done using MedCalc version 13.1.0.0. Sensitivity and specificity covariances for pairs of tests (a measure of conditional dependence or correlation) were calculated... assuming the true infection status … was known with certainty…'

Details of the methods, including the specific references and software used, are given.

The Results section

The **Results section** includes two topics (participants and test results), both relevant to the study because of its hybrid design, using known non-infected animals and animals with the target condition, as defined by the reference standard of culture or real-time polymerase chain reaction (PCR) detection of *B. mallei*.

Item 23: Cross tabulation of the index test results (or their distribution) by the results of the reference standard

> 'The frequency of combinations of the 4 serologic test results (positive/negative) in infected and non-infected populations is shown in Table 2 [see Table 2 in (12)].'

Cross-tabulation gives the reader an overview of the results and is extremely informative.

Item 24: Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)

> 'The CFT was the most sensitive test in this study (98.0%)… The highest sensitivity of pairs of tests in parallel (99.6%) was achieved by use of CFT and any of the other 3 tests (WB [Western blotting], IDVet– [a commercial test] or Hcp1– [haemolysin-coregulated protein 1] ELISA). The remaining 3 combinations all had combined Se <99%. Use of test pairs in series yielded values of 99.9% or 100% for all 6 combinations. Positive sensitivity and specificity covariances between tests decrease the gain in sensitivity for use of 2 tests in parallel… and gain in specificity when 2 tests are used for serial interpretation… respectively. Estimates obtained in the Bayesian analysis [median and 95% probability intervals] were similar to those obtained in the traditional frequentist analysis based on known infection status (Table 4 [see Table 4 in (12)].
>
> 'Table 5 [see Table 5 in (12)] shows the likelihood ratio positive and negative values of the four most promising tests (CFT, WB, IDVet and Hcp1–ELISA) as well as the positive and negative predictive values for scenarios of 0.1% and 3% prevalence, which were considered to be typical and worst cases for prevalence in equids. Based on the assumption that the true infection of equids was known, the prevalence in the study sample was 7.9% (254/3213) which is substantially above estimates in naturally-infected populations.'

Although the text gives only median estimates, 95% probability/confidence intervals (PCI) are given in Table 4 [see Table 5 in (12)], so this does fulfil the requirements. The additional information of the negative and positive predictive values at different prevalence scenarios is useful and clinically relevant information, although only applicable in a population with the given prevalence.

The Discussion section

The **Discussion section** has two items: 26 (study limitations) and 27 (implications for practice).

Item 26: Study limitations, including sources of potential bias, statistical uncertainty, and generalisability

> 'Our study therefore had to use a hybrid design in which cases were selected based on clinical presentation and positive PCR or culture results. Non-infected samples were collected from countries officially free of infection. This design may have led to some bias in estimates as glanders-positive cases [reference samples with target condition] may have been the most severely affected in source populations. However, because sample inclusion was not based on positive results by any serologic tests, we believe this bias was not substantial and was unlikely to impact the sensitivity and specificity rankings of the tests… The cut offs for the ELISAs and the WB were based on those recommended by the developer or manufacturer of the tests. It is obvious that other results will be obtained with different cut off values. Whether a cut off value shift is sufficient for the needs of OIE to improve the test properties of the ELISAs has to be proven by testing more positive [reference] samples and under consideration of the important prevalence-

> independent test characteristics of LR– [likelihood ratio] and LR+.'

This allows the reader to appraise how biases might have affected results and possible modifications of the candidate tests.

Item 27: Implications for practice, including the intended use and clinical role of the index test

> 'Considering both sensitivity and specificity [95% confidence limits are in parentheses], the WB (96.9%, 99.4%), the Hcp1–ELISA (95.3%, 99.6%) and the IDVet–ELISA (92.5%, 99.5%) should be further developed to meet OIE demands in the near future. These three tests were available for the study in very different stages of development. The WB, as performed at FLI OIE reference laboratory, was based on a semi-purified antigen and reading of WB results needs experienced operators. The in-house production of this test is expensive and thus this test is not appropriate for screening large numbers of samples in surveillance programs. Hence, it is suitable as a confirmatory test. The Hcp1–ELISA was available as a "semi-ready" non-commercialized format. The ELISA plates had to be coated with antigens, blocked and washed before the test could be used. If this ELISA would be produced commercially, this test could be a promising tool for the future. The IDVet–ELISA was provided as a ready-to-use commercialized kit.'

A few conclusive remarks summarising the findings and application of the tests in practice are given. This allows extrapolation of the results and provides useful information to decision-makers.

Example of good practice adhering to STARD–BLCM: accuracy of three assays to detect *Mycobacterium bovis* in blood and milk samples from Egyptian dairy cows

Bovine tuberculosis (bTB) is a chronic zoonotic disease caused by *Mycobacterium bovis* with a worldwide distribution. It is characterised by a long latent infection period and the various infection stages can vary among populations (2). Thus, there is a need for accurate, reliable and cost-effective diagnostics that should be, whenever possible, validated through samples from the target/source population. In this study, Elsohaby *et al.* (13) estimated the Se and Sp of PCR, mycobacterial culture and interferon-gamma (IFN–γ) assays for *M. bovis* detection in blood and milk samples from Egyptian dairy cows. Because of the long latent period and the absence of a perfect reference test (i.e. gold standard), BLCMs are a natural choice for the evaluation of *M. bovis* diagnostics.

Elsohaby *et al.* used a BLCM approach and followed the STARD–BLCM guidelines (13). As in the previous examples of good practice, the current authors provide extracts from this work (shown indented and in quotation marks) as an example of how to report key items from the STARD–BLCM checklist (Table I) (10). These items relate to key points in which STARD–BLCM differs from STARD–2015. Those key points are:

- the definition of the infection status (target condition) under the BLCM approach
- the assumptions underlying the BLCM model and a biological justification for these assumptions
- the incorporation of prior information and justification for its use, as well as the validation of model assumptions
- the proper reporting of results under alternative prior specifications (i.e. sensitivity analysis).

Readers may refer to the STARD–BLCM publication (10) for examples and descriptions of each of the STARD–BLCM items.

The Introduction section

The **Introduction section** of STARD–BLCM has two topics: Item 3 (scientific and clinical background, including the intended use and clinical role of the tests under evaluation) and Item 4 (study objectives and hypotheses, such as estimation of diagnostic accuracy of the tests for a defined purpose through BLCM). Both are important but Item 4 is relevant to the BLCM framework.

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

> '… to the authors' knowledge, the evaluation of mycobacterial culture, PCR and IFN–γ performance for the detection of *M. bovis* in cattle blood and milk using latent class models has not been performed in Egypt, where farm management practices, control strategies and disease burden may be different from developed countries. Therefore, the aim of the present study was to estimate the Se and Sp of PCR, mycobacterial culture and IFN–γ assays for the

> detection of *M. bovis* in blood and milk from dairy cows in Egyptian herds within a Bayesian framework. As a secondary objective, the true within-herd prevalence of *M. bovis* infection in Egyptian dairy herds was estimated.'

A clear rationale is given for the study, including the primary and secondary objectives.

The Methods section

The **Methods section** has four topics (study design, participants, test methods and analysis), each of which has multiple items. Here, the current authors provide examples for items related to key elements of the latent class approach and the Bayesian estimation process, specifically for Items 11: the rationale for choosing the tests under evaluation in relation to their purpose; 14 (a): a BLCM model for estimating measures of diagnostic accuracy; and 14 (b): definition and rationale of prior information and sensitivity analysis.

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

> 'The infection status (target condition) targeted by PCR and mycobacterial culture constitutes a blood/ milk sample containing either the live *M. bovis* organism or its debris at any concentration level. However, the IFN–γ assay targets IFN–γ production due to stimulation of lymphocytes in blood samples.'

This text gives a clear and helpful rationale for the biological basis of the applied diagnostic tests and specifies the target condition of each test.

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

> 'In constructing BLCMs, three assumptions are key… Firstly, the target population should comprise two or more subpopulations with different prevalences. In this regard, the 11 dairy herds construed to have different within-herd prevalences (potentially attributable to variations in herd management practices [reference given]) constituted the specific subpopulations. Secondly, the Se and Sp of the index tests should remain constant across the subpopulations. To evaluate this assumption, the specific subpopulations were sequentially eliminated from the models followed by a re-estimation of the Se and Sp of the tests. Thirdly, the tests are presumed to be conditionally independent, given the disease status. For our situation, the PCR and mycobacterial culture tests were assumed to be conditionally dependent… but independent from IFN–γ. Specifically, two conditional covariances between pairs of the Se and Sp of PCR and culture were defined… The covariances were tested for departures from zero (zero covariance denoting conditional independence) using a Bayesian P-value.'

The text gives a clear description and justification for the structure of the model used, which is interpretable by readers who are not necessarily familiar with how these models are coded.

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

> 'With prior information available on PCR and culture… estimates were used to specify the beta prior distributions for the two tests… As for the remainder of the parameters, uninformative beta distributions were used… Furthermore, for each of the models containing prior information, separate models without informative priors were built and the relative goodness of fit for the different model specifications compared using the Deviance Information Criterion (DIC) – model preferability being based on the smallness of the DIC value.'

The text gives a clear derivation and justification of the priors used for Se and Sp, and states that minimally informative priors were used for other parameters.

The Results section

The **Results section** includes two topics (participants and test results). Here, the current authors present Item 24 (estimates of diagnostic accuracy under alternative prior specification and their precision), which is relevant to BLCMs.

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

> 'The cross-tabulated counts of the blood and milk tests are displayed in Table 1, Table 2, respectively [see tables in (13)]. Of note, following the comparisons between models (for both blood and milk samples) with and without prior information, the models without informative priors had better fit to the data (DICs [deviance information criteria] = 118; 43 for blood and milk samples respectively) compared to informative models (DICs = 153; 86 for blood and milk samples respectively). Consequently, the test characteristics and within-herd prevalences were derived based on the uninformative prior models.

> ‘The estimates of the true within-herd prevalences of *M. bovis* in the 11 dairy herds along with the Se and Sp of PCR, culture and IFN–γ in blood and PCR and culture in milk samples are shown in Table 3, Table 4, respectively [see tables in (13)]. Importantly, in both samples, PCR and culture showed statistically significant conditional dependence. As for blood samples, the IFN–γ test registered a higher Se [0.97 (95% PCI: 0.87–1.0)] than PCR [0.68 (95% PCI: 0.53–0.95)] and culture [0.22 (95% PCI: 0.13–0.37)]. However, the Sp estimates of PCR [0.98 (95% PCI: 0.95–1.00)], culture [0.99 (95% PCI: 0.98–1.00)] and IFN–γ [0.97 (95% PCI: 0.88–1.00)] were comparable.’

Results based on the final selected model are presented (median and 95% probability intervals).

The Discussion section

The **Discussion section** has two items. The current authors present Item 27, which relates to BLCM and the interpretation of the Se/Sp estimates.

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

> ‘The growing number of bTB infected dairy herds in Egypt calls for accurate diagnosis of *M. bovis* at the cow level to facilitate cost-effective bTB eradication since whole herd-culling is not economically sustainable. In a significant effort to improve the bTB eradication in Egyptian dairy herds, the GOVs [General Organizations of Veterinary Services] are implementing a routine SCT [single cervical tuberculin test] and slaughter surveillance system… Previous studies in Egypt rely on the classical diagnostic evaluation using mycobacterial culture as the reference test to estimate the diagnostic test characteristics… This approach is associated with many reported bias[es]… Se estimates vary with the stage of infection, and animals with advanced *M. bovis* infection are more likely to be culture-positive compared to early-stage of infection… Therefore, using culture as a reference test for validation of bTB diagnostic tests… would be impractical…’

The text shows how the results of the current study will affect (improve) surveillance programmes aimed at bTB eradication in the region.

Conclusions

The STARD–2015 and STARD–BLCM guidelines for use in the presence and absence of a reference test, respectively, have resulted in the improved reporting of DAS. Adherence to reporting standards for OIE-listed diseases affects the conclusions of a DAS and, potentially, the usefulness of the test. Transparent reporting of key elements facilitates the comparability of results across different settings and ultimately enables end users and policy-makers to make sound choices as to which tests should be used for a defined purpose.

■

Exemples de notifications appropriées se rapportant aux études d’évaluation de tests diagnostiques (étape 2 de la validation) pour les maladies listées par l’Organisation mondiale de la santé animale

P. Kostoulas, I.A. Gardner, M.C. Elschner, M.J. Denwood, E. Meletis & S.S. Nielsen

Résumé

Les normes de notification et de conception sont des indicateurs essentiels de la qualité des études de validation des tests destinées à déterminer leur exactitude diagnostique ; or, en dehors des maladies des animaux aquatiques et de la paratuberculose chez les ruminants, il n’existe guère de lignes directrices pour concevoir ce type d’études pour les tests utilisés en santé animale. À la connaissance des auteurs, il n’existe pas non plus de normes de conception applicables aux études de validation en santé humaine. Par conséquent, il conviendrait de disposer de lignes directrices génériques fondées sur les caractéristiques des maladies telles que leurs modalités de transmission, leur période de latence et leur pathogénie. Une notification complète, claire et transparente des études d'exactitude des tests primaires pour les maladies listées par l'Organisation mondiale de la santé animale (OIE) serait une aide précieuse pour les utilisateurs finaux des tests de diagnostic, mais aussi pour les responsables de l'élaboration des politiques, dont les décisions reposent sur des examens et des méta-analyses systématiques couvrant un grand nombre de tests pour certaines maladies ou pour certains usages d'un test. La publication récente des normes de notification applicables aux modèles bayésiens à classes latentes pour analyser les données de performance d'un test à partir de foyers naturels de maladie comble une lacune importante dans la mesure où ces méthodes sont de plus en plus utilisées pour les maladies listées par l'OIE. L'adhésion à des normes de conception et de notification ainsi qu'à des lignes directrices en la matière permettra de garantir que les fonds alloués aux études de validation des tests sont bien utilisés et que les atouts et les limitations de certains tests individuels ou associations de tests sont clairement perçus par les utilisateurs. Les auteurs passent en revue certains points essentiels qui sont souvent ignorés ou mal interprétés lors des études de validation des tests et proposent deux exemples concrets de bonnes pratiques qui pourront servir de références pour les études à venir.

Mots-clés
Animaux aquatiques – Animaux terrestres – Exactitude d'un test – Lignes directrices de l'Organisation mondiale de la santé animale – Maladies infectieuses – Mammifères sauvages – Modèles bayésiens à classes latentes – Parcours de validation – Sensibilité diagnostique – Spécificité diagnostique – Validation d'un test.

■

Ejemplos de comunicación adecuada de la evaluación (validación de fase 2) de pruebas de diagnóstico de enfermedades incluidas en la lista de la Organización Mundial de Sanidad Animal

P. Kostoulas, I.A. Gardner, M.C. Elschner, M.J. Denwood, E. Meletis & S.S. Nielsen

Resumen

Las normas de comunicación y diseño son indicadores básicos de la calidad de los estudios de validación encaminados a determinar la exactitud de diagnóstico pero, con la excepción de las enfermedades de los animales acuáticos y la paratuberculosis en rumiantes, hay escasas directrices que se apliquen al diseño de esos estudios en animales. Además, hasta donde saben los autores, en el ámbito de la salud humana no hay normas de diseño. De ahí la necesidad de directrices genéricas que estén basadas en las características de las enfermedades, como modo de transmisión, período de latencia o patogénesis. La comunicación exhaustiva, clara y transparente de estudios primarios sobre la exactitud de pruebas de diagnóstico de enfermedades incluidas en la lista de la Organización Mundial de Sanidad Animal (OIE) reviste utilidad no solo para los usuarios finales de la prueba, sino también, en última instancia, para los órganos decisorios, que necesitan metaanálisis y estudios sistemáticos de múltiples pruebas que se apliquen a una u otra enfermedad y sirvan para una u otra finalidad. La reciente publicación de normas de comunicación de modelos bayesianos de clases latentes para analizar los datos de exactitud de pruebas a partir de episodios de enfermedad de origen natural viene a colmar una importante laguna, en la medida en que estos métodos se aplican cada vez más al diagnóstico de enfermedades incluidas en la lista de la OIE. El cumplimiento de las normas de diseño y comunicación, y también de las directrices, ayuda a garantizar que los fondos de investigación destinados a estudios de validación de pruebas sean utilizados debidamente y que el usuario final de una prueba reciba información clara sobre los puntos fuertes y las limitaciones de una prueba o combinación de pruebas. Los autores pasan revista a los principales aspectos que se suelen pasar por alto o malinterpretar en los estudios de validación de pruebas y ofrecen dos ejemplos concretos de buenas prácticas que se pueden utilizar como referencia en futuros estudios.

Palabras clave
Animales acuáticos – Animales terrestres – Directrices de la Organización Mundial de Sanidad Animal – Enfermedades infecciosas – Especificidad de diagnóstico – Exactitud de una prueba – Mamíferos salvajes – Modelos bayesianos de clases latentes – Proceso de validación – Sensibilidad de diagnóstico – Validación de pruebas.

References

1. World Organisation for Animal Health (OIE) (2019). – Principles and methods of validation for diagnostic assays for infectious diseases. Chapter 1.1.6. *In* Manual of Diagnostic Tests and Vaccines for Terrestrial Animals, 8th Ed. OIE, Paris, France. Available at: www.oie.int/fileadmin/Home/eng/Health_standards/tahm/1.01.06_VALIDATION.pdf (accessed on 26 October 2020).

2. Greiner M. & Gardner I.A. (2000). – Epidemiologic issues in the validation of veterinary diagnostic tests. *Prev. Vet. Med.*, **45** (1–2), 3–22. doi:10.1016/S0167-5877(00)00114-8.

3. World Organisation for Animal Health (OIE) (2019). – Quality management in veterinary testing laboratories. Chapter 1.1.5. *In* Manual of Diagnostic Tests and Vaccines for Terrestrial Animals, 8th Ed. OIE, Paris, France. Available at: www.oie.int/fileadmin/Home/eng/Health_standards/tahm/1.01.05_QUALITY_MANAGEMENT.pdf (accessed on 26 October 2020).

4. Gardner I.A., Colling A. & Greiner M. (2019). – Design, statistical analysis and reporting standards for test accuracy studies for infectious diseases in animals: progress, challenges and recommendations. *Prev. Vet. Med.*, **162**, 46–55. doi:10.1016/j.prevetmed.2018.10.023.

5. Cohen J.F., Korevaar D.A., Gatsonis C.A., Glasziou P.P., Hooft L., Moher D., Reitsma J.B., de Vet H.C., Bossuyt P.M. & the STARD Group (2017). – STARD for abstracts: essential items for reporting diagnostic accuracy studies in journal or conference abstracts. *Br. Med. J.*, **358**, j3751. doi:10.1136/bmj.j3751.

6. Cohen J.F., Korevaar D.A., Altman D.G., Bruns D.E., Gatsonis C.A., Hooft L., Irwig L., Levine D., Reitsma J.B., de Vet H.C.W. & Bossuyt P.M.M. (2016). – STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*, **6** (11), e012799. doi:10.1136/bmjopen-2016-012799.

7. Bossuyt P.M., Reitsma J.B. […] & Cohen J.F. for the STARD Group (2015). – STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Br. Med. J.,* **351**, h5527. doi:10.1136/bmj.h5527.

8. 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** (1–2), 18–34. doi:10.1016/j.prevetmed.2011.04.002.

9. Gardner I.A., Whittington R. […] & Lagno A.G. (2016). – Recommended reporting standards for test accuracy studies of infectious diseases of finfish, amphibians, shellfish and crustaceans: the STRADAS–aquatic checklist. *Dis. Aquat. Organisms*, **118** (2), 91–111. doi:10.3354/dao02947.

10. Kostoulas P., Nielsen S.S., Branscum A., Johnson W.O., Dendukuri N., Dhand N., Toft N. & Gardner I.A. (2017). – STARD–BLCM: standards for the reporting of diagnostic accuracy studies that use Bayesian latent class models. *Prev. Vet. Med.,* **138**, 37–47. doi:10.1016/j.prevetmed.2017.01.006.

11. Iowa State University (2021). – Menagerie of Reporting Guidelines Involving Animals (MERIDIAN). Iowa State University, Ames, United States of America. Available at: https://meridian.cvm.iastate.edu/diagnostic-tests/ (accessed on 27 January 2021).

12. Elschner M.C., Laroucau K. […] & Neubauer H. (2019). – Evaluation of the comparative accuracy of the complement fixation test, Western blot and five enzyme-linked immunosorbent assays for serodiagnosis of glanders. *PLoS ONE,* **14** (4), e0214963. doi:10.1371/journal.pone.0214963.

13. Elsohaby I., Mahmmod Y.S., Mweu M.M., Ahmed H.A., El-Diasty M.M., Elgedawy A.A., Mahrous E. & El Hofy F.I. (2020). – Accuracy of PCR, mycobacterial culture and interferon-γ assays for detection of *Mycobacterium bovis* in blood and milk samples from Egyptian dairy cows using Bayesian modelling. *Prev. Vet. Med.*, **181**, 105054. doi:10.1016/j.prevetmed.2020.105054.