Vol. 118: 91–111, 2016
doi: 10.3354/dao02947

DISEASES OF AQUATIC ORGANISMS
Dis Aquat Org

Published February 25

Recommended reporting standards for test accuracy studies of infectious diseases of finfish, amphibians, molluscs and crustaceans: the STRADAS-aquatic checklist

Ian A. Gardner[1,*], Richard J. Whittington[2], Charles G. B. Caraguel[3], Paul Hick[2], Nicholas J. G. Moody[4], Serge Corbeil[4], Kyle A. Garver[5], Janet V. Warg[6], Isabelle Arzul[7], Maureen K. Purcell[8], Mark St. J. Crane[4], Thomas B. Waltzek[9], Niels J. Olesen[10], Alicia Gallardo Lagno[11]

[1]Atlantic Veterinary College, University of Prince Edward Island, 550 University Avenue, Charlottetown, PEI C1A 4P3, Canada
[2]Faculty of Veterinary Science, University of Sydney, 425 Werombi Rd, Camden, NSW 2570, Australia
[3]School of Animal & Veterinary Sciences, Roseworthy Campus, University of Adelaide, Roseworthy, SA 5371, Australia
[4]CSIRO Australian Animal Health Laboratory, Geelong, VIC 3220, Australia
[5]Fisheries and Oceans Canada, Pacific Biological Station, Nanaimo, BC V9T 6N7, Canada
[6]Diagnostic Virology Laboratory, National Veterinary Services Laboratories, VS, APHIS, USDA, Ames, IA 50010, USA
[7]IFREMER SG2M-LGPMM, Laboratory of Genetics and Pathology of Marine Molluscs, 17390 La Tremblade, France
[8]US Geological Survey, Western Fisheries Research Center, 6505 Northeast 65th Street, Seattle, WA 98115, USA
[9]Department of Infectious Diseases and Pathology, University of Florida, Gainesville, FL 32611, USA
[10]National Veterinary Institute, Technical University of Denmark, Frederiksberg C, Denmark
[11]Jefa Unidad de Salud Animal Servicio Nacional de Pesca y Acuicultura Calle, Victoria 2832, Chile

ABSTRACT: Complete and transparent reporting of key elements of diagnostic accuracy studies for infectious diseases in cultured and wild aquatic animals benefits end-users of these tests, enabling the rational design of surveillance programs, the assessment of test results from clinical cases and comparisons of diagnostic test performance. Based on deficiencies in the Standards for Reporting of Diagnostic Accuracy (STARD) guidelines identified in a prior finfish study (Gardner et al. 2014), we adapted the Standards for Reporting of Animal Diagnostic Accuracy Studies — paratuberculosis (STRADAS-paraTB) checklist of 25 reporting items to increase their relevance to finfish, amphibians, molluscs, and crustaceans and provided examples and explanations for each item. The checklist, known as STRADAS-aquatic, was developed and refined by an expert group of 14 transdisciplinary scientists with experience in test evaluation studies using field and experimental samples, in operation of reference laboratories for aquatic animal pathogens, and in development of international aquatic animal health policy. The main changes to the STRADAS-paraTB checklist were to nomenclature related to the species, the addition of guidelines for experimental challenge studies, and the designation of some items as relevant only to experimental studies and ante-mortem tests. We believe that adoption of these guidelines will improve reporting of primary studies of test accuracy for aquatic animal diseases and facilitate assessment of their fitness-for-purpose. Given the importance of diagnostic tests to underpin the Sanitary and Phytosanitary agreement of the World Trade Organization, the principles outlined in this paper should be applied to other World Organisation for Animal Health (OIE)-relevant species.

KEY WORDS: Reporting standards · Sensitivity · Specificity · Finfish · Amphibians · Molluscs · Crustaceans · STRADAS-paraTB · Diagnostic validation

*Corresponding author: iagardner@upei.ca

INTRODUCTION

Estimation of indices of test accuracy, such as diagnostic sensitivity and specificity, is an important component of the evaluation process for tests used for detection of infectious diseases in aquatic and terrestrial animals. The World Organisation for Animal Health (OIE) *Manual of Diagnostic Tests for Aquatic Animals 2015* (OIE 2015a) describes a 4-stage validation pathway to assess a test's fitness-for-purpose: (1) analytical characteristics, (2) diagnostic sensitivity and specificity, (3) reproducibility among laboratories, and (4) program implementation. Diagnostic sensitivity and specificity estimates are considered essential for appropriate interpretation of test results for presumptive diagnosis, confirmation of clinical disease, and targeted surveillance for disease freedom to support animal trade, to list a few purposes. These parameters are important to consider when comparing different tests for the same disease. The OIE Aquatic Manual pertains to finfish, amphibians, molluscs and crustaceans; however, some classes of pathogens move between taxonomic groups, so the principles outlined in this paper should be applied more generally. For example, some ranaviruses (e.g. Frog virus 3), which are listed by the OIE as a cause of disease in amphibians, may also cause disease in finfish and reptiles (Waltzek et al. 2014). The same tests may be applied in species from all of these groups even though the diagnostic characteristics may be different.

Regardless of the study design chosen for estimation of diagnostic sensitivity and specificity, many health-related journals recommend application of the Standards for Reporting of Diagnostic Accuracy (STARD) statement (Bossuyt et al. 2003a,b) to enhance clear and transparent peer-reviewed reporting of pertinent information from a study. First published in 2013, STARD is based on recommendations of scientists and editors and is now endorsed by more than 200 biomedical journals. STARD does not prescribe design elements but has 25 checklist items that specify key information that should be reported. Because of different purposes of testing in human and animal health, STARD was modified for paratuberculosis in ruminants to account for different terminology and epidemiological units, and these guidelines were named the Standards for Reporting of Animal Diagnostic Accuracy Studies — paratuberculosis (STRADAS-paraTB) (Gardner et al. 2011). Recently, we used the STARD checklist to evaluate the quality of reporting in finfish studies and found highly variable reporting of its 25 items, a lack of guidance for researchers reporting use of experimental challenge studies to obtain sensitivity and specificity estimates, and 2 items that were minimally relevant to finfish studies (Gardner et al. 2014).

The purpose of the present study was to modify the STRADAS-paraTB guidelines, which was authored by 2 of us (I.A.G. and R.J.W.), to increase its relevance to test accuracy studies in aquatic animals (e.g. finfish, amphibians, molluscs and crustaceans) and to provide examples and explanations/elaborations for each of the 25 checklist items. In a PubMed search in October 2015, we did not find any published test accuracy studies involving aquatic animals that explicitly mentioned or followed the STARD or STRADAS-paraTB reporting recommendations. Reasons could include lack of awareness of or perceived lack of relevance of these guidelines. Therefore, our motivation was that increasing the relevance of these checklist items for aquatic animals would ultimately enhance adoption of reporting guidelines for test accuracy studies in aquatic animals regardless of species.

METHODS AND PROCESSES

The initial checklist known as STRADAS-aquatic was developed by 3 of us (Gardner, Caraguel, and Whittington) and was subsequently expanded and refined by the expert panel. Potential experts were identified based on finfish papers reviewed in Gardner et al. (2014) and on knowledge of test evaluation papers in amphibians, molluscs, and crustaceans. Inclusion criteria were that experts must have either published a paper on test accuracy, were the head or member of a European Union (EU) or OIE reference laboratory for aquatic animal pathogens, or were involved in formulation of policy regarding trade in aquatic animals.

Of 11 additional experts contacted by one of us (Gardner), all agreed to participate. Contributors self-classified themselves as molecular biologists (Purcell), molecular virologists (Arzul, Corbeil, Garver, Moody, and Waltzek), aquatic animal health scientists/researchers or epidemiologists (Caraguel, Crane, Gardner, Hick, Olesen and Whittington), or laboratory diagnosticians (Warg). Of the 14 authors, 4 (Crane, Gardner, Purcell, and Waltzek) are associate editors or editorial board members of journals publishing aquatic animal heath papers, 4 (Caraguel, Garver, Olesen, and Purcell) authored a paper that was evaluated by Gardner et al. (2014), and 5 others (Crane, Hick, Moody, Waltzek and Whittington) had authored a manuscript on test accuracy. Six (Arzul,

Crane, Moody, Olesen, Purcell, and Whittington) are heads or members of OIE or EU reference laboratories for aquatic animal diseases, and one of us (Gallardo Lagno) is a member of the OIE Aquatic Animal Health Standards Commission. Collectively, authors had experience with species in all 4 taxonomic groups and with terrestrial animals.

Experts were required to contribute examples and elaborations based on their own experiences and/or knowledge of published literature and critically review the manuscript including suggesting improved wording in the examples used. The panel operated remotely by email and telephone and as opportunities arose, the lead author had in-person discussions with coauthors at conferences. No major discrepancies in opinions as to content were evident during the process although there was discussion as to the most appropriate item for some considerations. During the writing process, reference to design issues was minimized to maintain focus on reporting.

THE STRADAS-aquatic CHECKLIST

Of the 25 STRADAS-paraTB checklist items, the following adaptations were made: (1) substantial changes to 5 items (Items 3, 4, 6, 23, and 25) to incorporate use of experimental challenge studies; (2) modifications to 2 items (Items 17, 20) for antemortem test application only and to a single item (Item 14) for field study application only; (3) substitution of the word 'herd' or 'flock' with 'population' in several items (Items 2, 3, 4, 11, 15, and 16); and (4) minor wording changes to 10 items (Items 2, 5, 7, 10, 12, 13–16, and 19) (see bolded adaptations in Table 1).

We present examples from published papers which are intended to represent best practices, but in some cases improvements are suggested. Although many examples are from PCR evaluation studies, we emphasize that the principles apply equally well to traditionally used tests such as virus, bacterial, and parasite isolation, to gross and histopathological examination of organs and tissues, to tests that detect serological responses and tests based on new technology such as *in situ* hybridisation using RNA probes and multiplex assays using reagents conjugated to fluorescent beads.

As in Gardner et al. (2011), we use square brackets for any wording additions that improve readability of an example and for spelling out acronyms. We use the term test under evaluation (TUE) to designate any test for which diagnostic sensitivity and specificity are estimated. Definitions and terms used in the manuscript are in the Supplement (www.int-res.com/articles/suppl/d118p091_supp.pdf).

The structure of this manuscript reflects that of the STARD checklist and a research article: Title/Abstract/Keywords; Introduction; Materials and methods; Results; and Discussion.

Title/Abstract/Keywords

Item 1: Identify the article as a study of diagnostic accuracy (recommend MeSH [medical subject headings] terms 'sensitivity and specificity').

Examples:

Kent et al. (2013) entitled their paper 'Sensitivity and Specificity of Histology for Diagnoses of Four Common Pathogens and Detection of Nontarget Pathogens in Adult Chinook Salmon (*Oncorhynchus tshawytscha*) in Fresh Water' and included estimates of sensitivity and specificity in their abstract. In contrast, Garver et al. (2011, p. 95) did not include sensitivity and specificity in their title but rather in their abstract: 'Test performance characteristics evaluated on experimentally infected Atlantic salmon *Salmo salar* L. revealed a diagnostic sensitivity (DSe) ≥ 93% and specificity (DSp) = 100%.' Both papers were identified when diagnostic sensitivity and specificity were used as search terms.

Explanation:

The MEDLINE/PubMed database uses a controlled vocabulary thesaurus termed MeSH (www.nlm.nih.gov/bsd/disted/meshtutorial/introduction) to allow hierarchical searching of the biomedical literature including diagnostic accuracy studies. Use of the MeSH terms 'sensitivity' and 'specificity' is recommended in the title, keywords or abstract to increase the likelihood of retrieval of relevant studies in a bibliographic search. However, studies that use the terms in their analytical rather than diagnostic context may also be retrieved and require exclusion because they lack estimates of diagnostic sensitivity and specificity. The term 'validation', which is commonly used by OIE for test accuracy studies, has a nearest match of 'validation studies' in MeSH. This term describes works in which the reliability and relevance of a procedure for a specific purpose are established, and thus it is relevant to more than diagnostic sensitivity and specificity studies. Inclusion of additional databases (e.g. CAB Abstracts, www.cabi.org/publishing-products/online-information-resources/cab-abstracts/) in the search may be needed to retrieve all relevant articles because PubMed may

Table 1. Standards for Reporting of Animal Diagnostic Accuracy Studies (STRADAS-aquatic) checklist of items for reporting in diagnostic test accuracy studies for finfish, crustaceans, and molluscs based on the STARD (www.stard-statement.org) and STRADAS paratuberculosis (STRADAS-paraTB) (Gardner et al. 2011) checklists. TUE: test under evaluation. Modification of text from the STRADAS-paraTB checklist is in **bold**

Section and topic	Item	Description of item
TITLE/ ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH [MeSH: medical subject headings] heading 'sensitivity and specificity').
INTRODUCTION	2	State the **intended purpose of the TUE and reasons for test accuracy assessment,** such as estimation of diagnostic accuracy or comparison of accuracy between tests in a specified matrix (specimen type) for a defined purpose at the animal or **population** level.
MATERIALS AND METHODS Animals and **populations**	3a	**For field studies,** describe **the study population including other susceptible species around the target population**. Describe setting and locations where data were collected for all relevant levels of the study sample (animals and **populations**), **detailing** inclusion and exclusion criteria.
	3b	**For experimental studies, describe source, life stage, and health history of aquatic animals and specifically indicate prior infection status for the pathogen(s) of interest, including diagnostic testing in study animals and/or source population.**
	4a	**For field studies,** describe selection of animals and **populations**. Describe sample selection methods (random, convenience, etc.) within each level of the sampling hierarchy (e.g. regions, **sites, cages/net-pens, tanks, or ponds**), including exclusion criteria, **and** number of study animals and **populations.**
	4b	**For experimental studies, describe (1) design (e.g. number of treatment and control groups), randomization process, numbers of replicates (number of housing units and animals per housing unit), duration of experiment including start date, and challenge conditions (e.g. challenge strain and passage level for the organism(s), dose, exposure route), and animal use and care committee approval, (2) sampling (time post-challenge that samples were harvested including numbers at each time), and (3) husbandry and environmental conditions (e.g. housing type, acclimation time, water source and relevant physical and chemical characteristics, feeding regimen, handling and care).**
	5	Describe **specimen collection**: Describe the collection, **target organs including gross lesions (if sampled),** specimen size **and number sampled**, transportation, handling **(laboratory scientist and/or field collection)** and storage **(laboratory and/or field) methods and times for** specimens prior to the performance of the test under evaluation (TUE) and the reference standard.
	6	Describe study design: **For field studies,** was data collection planned before the TUE and reference standard were performed (prospective study) or after (retrospective study)? **For experimental studies, were archived samples included? Describe details of storage and retrieval techniques and times.**
Test methods	7	Describe the reference standard **(if used)** and its rationale.
	8	Describe technical specifications of materials and methods involved including how and when measurements were taken, and/or cite references for TUE and reference standards. Specify quality control samples for TUE and reference standard and specimen/analytical unit size of tested samples.
	9	Describe the outcome measure and rationale for the cutoffs and/or categories of the results of the TUE and reference standard.
	10	Describe the name, location, and qualifications of the laboratory, including the number, training, and expertise of persons executing the TUE and reference standard. **Specifically indicate if the laboratory or analyst(s) is involved in any internal or external assessment program (e.g. proficiency testing).**
	11	Describe whether or not the readers of the TUE and reference standard were blind (masked) to the results of the other test and describe any individual or **population**-level information available to the readers.
Statistical methods	12	Describe methods for calculating and comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty **in the estimates** (e.g. 95% confidence **or probability** intervals).
	13	Describe methods for **estimating** test repeatability and reproducibility, if done.
RESULTS	14	**For field studies,** report when study was done, including **start** and end dates.
Animals and **populations**	15	Report demographic and other biologically relevant characteristics of the study sample at the individual **level** (e.g. age, sex, **ploidy species, genotype if known, weight,** and risk factors) and at the **population** levels (e.g. production system, **water quality, water temperature, and water salinity).**

(Table continued on next page)

Table 1 (continued)

Section and topic	Item	Description of item
	16	Report the number of animals and **populations** satisfying the **inclusion** criteria that did or did not undergo the TUE and/or the reference standard; describe why animals and **populations** failed to receive either test.
Test results	17	Report time interval between collection of samples for the TUE and the reference standard, and interventions administered between **for samples collected ante-mortem**.
	18	Report distribution of severity of disease or stage of infection (define criteria) and other relevant diagnoses or treatments in animals in the study sample.
	19	Report a cross tabulation of the results of the TUE (including indeterminate and missing results) by the results of the reference standard. For continuous results, **report** the distribution of the test results by the results of the reference standard.
	20	Report any adverse events from performing the TUE or the reference standard **for samples collected ante-mortem**.
Estimates	21	Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).
	22	Report how indeterminate results, missing responses, and outlier values of the TUE and the reference standard were handled.
	23	Report estimates of variability of diagnostic accuracy between relevant subpopulations, **operators/**readers, **host factors, agent factors,** or testing sites, if done. **For challenge studies, report temporal variation in sensitivity estimates compared with days post-challenge.**
	24	Report estimates of test repeatability and reproducibility, if done.
DISCUSSION	25	Discuss **the fitness for the stated purpose**, and the utility of the TUE in various settings (clinical, research, surveillance, etc.) in the context of the currently available tests. **For challenge studies, critically evaluate the relevance of the experimental challenge study to naturally occurring infection/disease and explain why the latter source of samples was not used.**

include only a subset of aquatic animal journals publishing studies of diagnostic accuracy and other aquatic topics. For example, *Journal of Fish Pathology* and the *Bulletin of European Association of Fish Pathologists* are not indexed in PubMed.

Introduction

Item 2: State the intended purpose of the TUE and reasons for test accuracy assessment, such as estimation of diagnostic accuracy or comparison of accuracy between tests in a specified matrix (specimen type) for a defined purpose at the animal or population level.

Example:

The aim of our study was to develop molecular tools for a specific, rapid and sensitive diagnostic of bonamiosis in the European flat oyster *Ostrea edulis* … We compared the sensitivity of these new diagnostic assays [species-specific conventional PCR, real-time PCR, and multiplex PCR] with 2 OIE listed procedures (www.oie.int/international-standard-setting/aquatic-manual/access-online/), viz, the standard histological procedure and the PCR-RFLP assay, assuming that none of the compared procedures could be considered as a 'gold standard' (Ramilo et al. 2013, p. 150).

In the introduction, the authors also indicate that sensitive and specific methods are needed to carry out early detection of *Bonamia* spp. with the aim of preventing infection of healthy animals and dispersion of the agent to non-affected areas, implying that was their study purpose.

Explanation:

The concept of 'fitness for an intended purpose' is fundamental to the validation pathway recommended in the OIE *Manual of Diagnostic Tests for Aquatic Animals 2015* (OIE 2015a), including evaluation in the species for which the test is intended. Because choice of study populations and animals will vary with purpose, clear specification of the study aims and an intended purpose is essential. The objective statement by Ramilo et al. (2013) above could have been improved by an explicit description of test purpose and reference to specimen type, namely gills and gonads, because sensitivity will vary with tissue predilection and load of the target analyte. Ideally, the objective statement should also specify the epidemiological unit(s) of interest because test results for aquatic food animals are typically interpreted at a population rather than at the individual animal level.

Applicable OIE-recognized purposes should also be reported wherever possible. For example, in a

study of the OIE-listed amphibian disease chytridiomycosis (*Batrachochytrium dendrobatidis*), Hyatt et al. (2007) evaluated the diagnostic accuracy, repeatability and reproducibility of histopathology, histochemistry, and real-time TaqMan PCR as well as various sampling protocols (toe clipping, water baths and filters, and swabs). They stated that uses for these assays and protocols, as per the quantification results, are purposes '… analogous to those defined by the OIE [5 specific functions listed in detail]…for assays intended for the detection and for minimising the translocation of pathogens via the international movement of livestock and associated commodities …' (Hyatt et al. 2007, p. 176)

Materials and methods

Animals and populations

Item 3a: For field studies, describe the study population including other susceptible species around the target population. Describe setting and locations where data were collected for all relevant levels of the study population (animals and populations), detailing inclusion and exclusion criteria.

Example:

> Three Atlantic salmon seawater sites were studied [on the west coast of Norway]. Two sites were included at the start of their PD [pancreas disease] outbreak (sites 1 and 2). A third site was included at the time of smolt seawater transfer because of a high probability of contracting PD based on its location in close proximity to existing outbreaks (site 3). This third site was also included in a cohort study, where the smolt groups tested negative for SAV [salmonid alphavirus] in the freshwater phase. All sites were situated in a region considered endemic for PD although with no recent history of clinical PD … Detailed site, stock and outbreak information [including fallow period, time of seawater transfer, number of fish on farm at seatransfer, PD diagnosis, fish weight, cumulative PD-related mortality, sampling period and time of last sample collection] can be found in Table 1 … Each site had 3 cages selected for inclusion in the study, preferably cages evenly distributed within the site and containing smolt from different smolt producers (Jansen et al. 2010, p. 725).

Explanation:

External generalizability of estimates of diagnostic sensitivity and specificity to related species and other geographically distinct populations is in part dependent on the study design, sampling methods (Item 4a), knowledge of management and housing practices (wild stocks vs. cultured animals), concurrent diseases, and animal demographics in the sampled and source populations. Relevant environmental parameters at study sites (e.g. temperature and salinity) should be reported for field studies and for tank experiments (see Item 4b). Disclosure of these parameters as well as disease-specific variables where available, such as pathogen load, tissue distribution, and prevalence within the population, allow for assessment of generalizability. Start and end dates of the study should also be reported in the 'Materials and methods' or 'Results' sections of the paper (see Item 14).

Item 3b: For experimental studies, describe source, life stage, and health history of aquatic animals and specifically indicate prior infection status for the pathogen(s) of interest, including diagnostic testing in study animals and/or source population.

Example:

> Spawning fall Chinook salmon to provide SPF [specific-pathogen-free] progeny for this research were obtained from Strawberry Creek, Wisconsin in October 2004 … Five families were selected on the basis of negative kidney tissue results by ELISA [screening] testing [for *Renibacterium salmoninarum*], negative or borderline positive ovarian fluid results by ELISA testing, and negative or borderline positive ovarian fluid results by MF-FAT [membrane filtration-fluorescent antibody technique] testing (≤1 bacteria in 150 microscopic fields). In November 2004, eyed eggs from the selected families were transferred from Wild Rose Hatchery, Wisconsin to the Western Fisheries Research Center in Seattle, Washington. The fish were hatched and reared in sand-filtered, UV-treated Lake Washington water for 2 years prior to challenge. (Elliott et al. 2013, p. 787)

Explanation:

Estimates of test performance in experimental challenge studies may be influenced by prior infection history, concurrent disease and host genetics. Consideration of prior exposure to pathogens is even more important when the TUE specifically addresses prior infection status (e.g. serological tests for antibodies). Hence, the genetic background of the animals (selected or wild type), the health history of animals (e.g. recorded outbreaks in wild or farmed source populations) that are used in experiments should be reported, as should results of testing a subsample of animals (e.g. by histology and PCR) to confirm freedom of infection by the agent under investigation as in the preceding example.

Item 4a: For field studies, describe selection of animals and populations. Describe sample selection methods (random, convenience, etc.) within each level of the sampling hierarchy (e.g. regions, sites, cages/net-pens, tanks, or ponds), including exclusion criteria and number of study animals and populations.

Example:

Sample collection was conducted in September 2003. Samples were taken from 3 BC Atlantic salmon farms [shown in Fig. 1 in their paper] that represented different prevalence populations: (1) a farm undergoing an IHNV [infectious hematopoietic necrosis virus] epizootic (fish with high IHNV prevalence; average weight = 1.7 kg), (2) a farm that had recently experienced an epizootic (fish with low IHNV prevalence; average weight = 5.0 kg), and (3) a farm that never had experienced an outbreak (fish with no IHNV; average weight = 1.5 kg). From farm 1, 50 fresh mortalities or moribund fish were sampled from 6 affected cages via cage removal by a mortality diver and an uplift system. From farm 2, 50 fish from 9 affected cages were taken by the mortality diver. From farm 3, 50 healthy fish from 1 pen were seined at the IHNV-negative site [see Table 2 in their paper] (McClure et al. 2008, p. 13).

Explanation:

The method of sampling should be described in sufficient detail to allow replication of the approach in other studies and ensure consistency in sampling from pen to pen, site to site etc., thereby allowing readers to generalize results to other populations. The example from McClure et al. (2008) indicates a targeted approach in Farms 1 and 2 to obtain samples from clinically affected fish and a convenience (non-random) sample of healthy fish at Farm 3 for estimation of specificity.

Sampling of moribund and clinically affected animals that fit the case definition is essential to provide positive controls as part of a test accuracy study. Gross pathology, specific morbidity signs and histopathological analysis of clinically affected animals are required to obtain a diagnosis and should be reported as pertaining to the purpose of testing. When another pathogen-specific test is available (e.g. conventional PCR), besides the test being validated, it increases the likelihood of accurate diagnosis. Likewise, the same testing protocol should be performed and reported for animals from disease-free areas that were used as negative controls.

Item 4b: For experimental studies, describe (1) design (e.g. number of treatment and control groups), randomization process, numbers of replicates (number of housing units and animals per housing unit), duration of experiment including start date, and challenge conditions (e.g. challenge strain and passage level for the organism(s), dose, exposure route), and animal use and care committee approval, (2) sampling (time post-challenge that samples were harvested including numbers at each time), and (3) husbandry and environmental conditions (e.g. housing type, acclimation time, water source and relevant physical and chemical characteristics, feeding regimen, handling and care).

Examples:

Newly metamorphosed, lab-bred *Ambystoma tigrinum nebulosum* from 2 different clutches were housed individually in plastic containers in 946 ml of water before the experiment. Each was fed 2 crickets, twice a week, and had its water changed weekly. Animals were randomly assigned to 2 groups: infected and control. Each group included metamorphosed animals from each clutch. A total of 68 animals, 34 from each clutch, were in the experimentally infected group, and 18 were in the uninfected control group. At 5 times (2, 5, 8, 12, and 15 days), 14 infected animals (7 from each clutch) and 4 uninfected control animals (2 from each clutch) were sampled to assess the diagnostic performance of the PCR test (Greer & Collins 2007, p. 526–527).

To generate a population of VHSV [viral hemorrhagic septicemia virus] infected fish, 63 Atlantic salmon smolts (mean weight 223 g per fish) were i.p. [intraperitoneally] injected with 1×10^4 pfu [plaque forming units] $fish^{-1}$ of VHSV isolate 99-292 (genotype IVa). For a population of non-infected fish, 50 Atlantic salmon were left unhandled and maintained in a separate tank. Fish were held in a 750 l tank with 8°C seawater. At 3, 5, 7, 10, and 11 d post challenge, 10 fish from each of the virus-challenged and negative control tanks were killed with an overdose of tricaine methane sulphonate (MS_222) (Garver et al. 2011, p. 102).

Explanation:

The nature of samples generated during experimental infection will be different from samples obtained during application of the TUE for field testing. Husbandry under experimental conditions will not be the same as in the wild or commercial aquaculture. Factors including stocking density, water source and relevant physical and chemical characteristics, stress levels, health status including concurrent infection, and quality and quantity of feed are likely to influence the prevalence of infection and the quantity of pathogen present. By describing the conditions under which the experimental samples were generated, differences in these factors which might affect diagnostic characteristics of the test can be accounted for. The challenge conditions, particularly the dose and strain of pathogen, should be reported to allow readers to evaluate how well the experimental model reflects natural exposure under field conditions. The examples by Greer & Collins (2007) and Garver et al. (2011) include most of the necessary elements with the exception of Animal Ethics Committee approval of the experiment design. Research studies with amphibians and fish, but typically not molluscs and crustaceans, must be approved by the Animal Ethics Committee in most research institu-

tions and should be reported as done in Purcell et al. (2013), with the possible addition of approved protocol numbers (see Hyatt et al. 2007).

Item 5: Describe specimen collection. Describe the collection, target organs including gross lesions (if sampled), specimen size and number sampled, transportation, handling (laboratory scientist and/or field collection) and storage (laboratory and/or field) methods and times for specimens prior to the performance of the test under evaluation (TUE) and the reference standard.

Example:

From each of the 20 [euthanized] fish, heart and mid-kidney tissue in RNAlater® (Ambion) for virological examination by Rt RT-PCR [real-time reverse transcriptase PCR] and heparinized blood samples for detection of antibodies against SAV [salmonid alphavirus] were collected. Additionally, heart and mid-kidney tissues were collected in viral transport medium [Eagle's minimum essential medium, pH 7.6, supplemented with 10% newborn bovine serum and 100 µg/mL gentamicin] for virus isolation in cell culture (except for site 3 at slaughter). From 10 fish, including the moribund fish, tissue samples were collected in 10% neutral buffered formalin for histopathological examination, and consisted of heart, pyloric caeca with pancreas, muscle, gill, liver, kidney and spleen; with the exception of the last samples from sites 1 and 3 when only heart, pyloric caeca with pancreas and muscle were collected. Samples were shipped on ice with overnight delivery or, if necessary, refrigerated overnight prior to overnight shipping to the NVI [Norwegian Veterinary Institute]. All samples, except for formalin fixed tissues, were stored at −80°C until analysis was performed (Jansen et al. 2010, p. 725–726).

Explanation:

Specimen handling, transportation, and storage may affect the sensitivity of some tests (e.g. virus, bacteria, and parasite isolation) more than others (e.g. quantitative PCR), but relevant information should be reported regardless of test type. The description in the example could have been improved by indicating whether gross lesions, if present in heart and muscle, were sampled and how soon testing was done after storage, as the latter may affect the probability of virus isolation from infected tissues. Flow charts are useful to show sampling schemes, as in Patil et al. (2008), who also included sampling times in their text description.

Item 6: Describe study design. For field studies, was data collection planned before the TUE and reference standard were performed (prospective study) or after (retrospective study)? For experimental studies, were archived samples included? Describe details of storage and retrieval techniques and times.

Example:

Gustafson et al. (2008) described their infectious salmon anemia (ISA) surveillance program used to generate data for evaluation of 2 diagnostic tests (conventional RT-PCR and indirect fluorescent antibody, IFAT) as 'cross-sectional'. In this study, 10 461 samples were collected from 2002 to 2005 from Atlantic salmon farms in Maine, USA, as part of a surveillance program and submitted for parallel testing for ISA virus (ISAV) by IFAT and RT-PCR.

Explanation:

Prospective and cross-sectional designs, using standardized procedures for sample collection, transportation and handling, usually allow for better quality and consistency of samples with more detailed descriptive data of populations and animals sampled than would normally occur in a retrospective study using repository samples. The example by Gustafson et al. (2008) meets those criteria as do the examples in Items 3b and 4b, where the authors used prospectively generated experimental samples for validation of the TUE. The infection trials described in Items 3b and 4b were conducted for the purpose of generating material of known history of pathogen exposure to facilitate evaluation of test accuracy. In contrast, Warg et al. (2014b) described the use of historical experimental infection trials in specific-pathogen-free Pacific herring *Clupea pallasi* to generate kidney and spleen samples for comparison of the sensitivity of 2 real-time PCR assays for VHSV genotype IVa. Storage methods (including preservation technique, tissue or sample type, temperature, repeated thaw-freeze samples, to name a few) should be reported as in Warg et al. (2014b) because deterioration of samples may lead to false-negative test results (also see Item 5).

Test methods

Item 7: Describe the reference standard (if used) and its rationale.

Example:

Histological examination was the standard reference test [for abalone herpesvirus (AbHV)] … in conjunction with epidemiological information about reference populations, and expert advice was used as the presumptive test to determine the true status of the samples. [Results of qPCR were compared with results of histological classification of neural tissues for the presence or absence of ganglioneuritis.] To minimise the effects of an imperfect reference standard, other information, such as epidemiological evidence of AbHV infection, was taken into consideration during the present study (e.g. prevalence and mortality were

included in the selection of the sites where abalone were sampled) (Corbeil et al. 2010, p. 8–9).

Explanation:

Authors should provide a justification for their choice of reference standard (RS) for the specified testing purpose based on whether the test is recommended in the OIE *Manual of Diagnostic Tests for Aquatic Animals 2015* (OIE 2015b), test cost, rapidity of results, laboratory considerations, etc. Ideally the RS should have published estimates of diagnostic sensitivity and specificity (see the Supplement for more discussion of RS). Comparison of a TUE with an imperfect RS may limit the ability of authors to demonstrate the superiority of new technology for detection of infectious diseases (Limmathurotsakul et al. 2012). For example, virus, parasite, and bacterial isolation are often used as RS for evaluation of the accuracy of PCR assays because false-positive isolation results rarely occur. If PCRs have greater diagnostic sensitivity (and comparable diagnostic specificity) to organism isolation, it is never possible to demonstrate this with traditional statistical approaches because an imperfect RS constrains the estimates of a TUE (e.g. PCR) to be less than 100%. The use of latent class analysis (LCA) methods (also see Item 13 and the Supplement) does not require specification of a RS and can be used to estimate sensitivity and specificity of all TUE subject to certain assumptions (Branscum et al. 2005).

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

Example:

During diagnostic validation of the VHSV [viral hemorrhagic septicemia] RT-qPCR [reverse transcriptase, real-time PCR], various quality controls were employed to monitor reaction efficiencies and ensure scientific integrity. At each stage of the VHSV RT-qPCR assay (i.e. cDNA synthesis and qPCR) at least one positive and one negative control was included … For cDNA synthesis, the positive control consisted of 1 µg of RNA extracted from Atlantic salmon kidney spiked with VHSV IVb whereas the negative control was DEPC [diethyl pyrocarbonate treated] water only (no RNA). Finally, for the VHSV-qPCR portion of the assay, a low and high positive control was included such that one reaction contained VHSV-IVb positive cDNA at 5×10^6 copies μl^{-1} (high) and another contained 2.5×10^2 copies μl^{-1} (low). These reactions were expected to generate C_t [cycle threshold] values of 20.8 and 35.9, respectively. Negative qPCR controls were included with each run … (Garver et al. 2011, p. 102–103).

Explanation:

Tissue processing and sample preparation methods must be described in detail because laboratories often adopt and modify methods to suit their specific needs, including availability of instruments and reagents. Depending on the method used, virus yield from tissue samples may vary greatly, and there could be substantial variation in sensitivities with subsequent contamination (Whittington & Steiner 1993, Hick et al. 2010, Rimmer et al. 2012). Likewise, diagnostic sensitivity of PCR-based assays may be affected by the choice of PCR reagents (Elliott et al. 2013, Jonstrup et al. 2013). For quality control purposes and to provide unambiguous definitions of positive and negative test results, both negative and positive controls should be included in each of the test runs. To show the needed information in this item, a description of the quality control samples used in a real-time RT-PCR validation study is presented in the example above (Garver et al. 2011) and in Greer & Collins (2007). Authors should also specify whether or not internal and/or artificial controls are included in PCR assays to identify false-negative or false-positive test results (Snow et al. 2009, Purcell et al. 2014), as well as the protocol for retesting such samples.

Samples collected for the TUE and RS need to be matched carefully, even if collected at the same time. Consider the example of nervous necrosis virus (NNV) infection, where brain and eye are the target organs for viral replication, and young/small fish are primarily affected. In subclinical NNV infection, it is possible for one retina to be positive and the other negative when evaluated using the RS (Hick et al. 2011). As the brain is small and there are 2 eyes per fish, authors should report how the sample was collected and divided for comparison of 2 different tests on tissues, such as PCR and histopathology. It is important that these specifics are reported because of potential impacts on test agreement for diseases where pathogen distribution in tissues may not be uniform or pathogen load may be low.

Item 9: Describe the outcome measure and rationale for the cutoffs and/or categories of the results of the TUE and reference standard.

Example:

Samples were considered negative for AbHV [abalone herpesvirus] when the C_T value was >35.8. Conservatively, samples with a C_T value <35.0 were considered clearly positive. Thus, based on this conservative estimate, there is an indeterminate range (C_T 35.0 to 35.8) and samples that yielded C_T values within this range were retested (Corbeil et al. 2010, p. 3–4).

Explanation:

Choice of a cutoff (threshold) value to designate test results as positive or negative will affect diagnostic sensitivity and specificity estimates and should consider test purpose, prevalence of infection, costs of false-positive or false-negative test results, and whether the epidemiological unit is the individual animal or the population. Authors should provide analytical or epidemiological justifications for their choices of cutoff values. This might be a simple statement such as 'mean + 2SD of known negative reference samples' or 'at the manufacturer's recommended value' when referring to an ELISA test kit. Considerations for PCR are more complex (Caraguel et al. 2011) and include use of a quantitative standard to define a positive result on a run-to-run basis (Hick & Whittington 2010).

In some situations, including the example above, 2 cutoff values might be used to define 3 categories of test results (positive, inconclusive, or negative) with the inconclusive category reflecting measurement uncertainty. Samples in the inconclusive range may be retested with the same or additional tests depending on the testing purpose and epidemiological unit of interest. For example, when the Corbeil et al. (2010) study was done, there was only a single TaqMan PCR based on Open Reading Frame (ORF) 49 but additional tests, TaqMan ORF 66 and ORF 77 PCR, as well as a conventional PCR are now used to obtain a product to sequence to confirm AbHV infection. Criteria to assess whether results of retested samples in the indeterminate range are positive, negative, or still have an inconclusive result should be specified in the report. Regardless of the choice of cutoff value, interpretation criteria for technical replicates, especially for those with discordant results, should be reported as in Purcell et al. (2013) with the retest result, if appropriate. For clarity, Purcell et al. (2013) could have reported the number of suspect samples and the final interpretation of their retest.

Item 10: Describe the name, location, and qualifications of the laboratory, including the number, training, and expertise of persons executing the TUE and reference standard. Specifically indicate if the laboratory or analyst(s) is involved in any internal or external assessment program (e.g. proficiency testing).

Example:

> Laboratories ('Labs') participating in this comparison had different levels of high-throughput or rRT-PCR [real-time reverse transcription PCR] testing experience (Table 2 in Warg et al. 2014a) ranging from extensive experience (high) to limited experience (recently trained). In Labs A, D, and E, personnel were experienced with high-throughput testing and with conducting rRT-PCR. In Labs B and H, personnel were experienced with large numbers of samples and with conventional RT-PCR, but were recently trained to perform real-time assays. Labs C and G have dual roles as both diagnostic and research laboratories with some experience with both high-throughput testing and rRT-PCR. Lab F also has dual function being involved in both research and diagnostics; in this laboratory, the technicians were recently trained in high-throughput testing and to perform real-time assays (Warg et al. 2014b, p. 18).

Explanation:

Animal health laboratories developing and evaluating test accuracy for OIE-listed diseases are expected to have a quality management system (QMS) which addresses technical, managerial, and operational elements of testing and interpretation of test results (OIE 2015b). Typically, the QMS will adhere to standards such as ISO/IEC17025:2005 (www.iso.org/iso/catalogue_detail.htm?csnumber=39883). Testing laboratories should report whether they have an accredited QMS for the laboratory and ISO 17025 or similar accreditation for the TUE and RS, as appropriate. Reporting of participation in proficiency testing (see the Supplement) may increase reader confidence in skills of personnel in the laboratory performing tests.

Although difficult to obtain for retrospective studies where details of RS testing may not be available, reporting of the number, training, and expertise of personnel carrying out the TUE and RS allows readers to evaluate the experience of the research or diagnostic team. Diagnostic accuracy of a TUE and/or RS may be influenced by personnel factors such as operator skill in extraction, processing, testing and reading of samples, especially if the interpretation of results has a subjective component (e.g. visual examination of tissue smears). The study in the above example (Warg et al. 2014b) provides the expertise of personnel as well as the names of the laboratories, distinguishing between diagnostic and research laboratories, but lacks specification of the actual accreditation. The number of individuals performing tests is elaborated in another section of their manuscript.

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

Example:

> Fish [challenged and unchallenged with viral hemorrhagic septicemia virus] … were individually bagged and labeled using a computer-generated random

number so that the treatment group was unknown to the laboratory testers (Garver et al. 2011, p. 102).

Explanation:

Although blinding is not necessary in routine diagnostic testing, testers participating in diagnostic accuracy studies should be blinded (masked) to either the results of the TUE or RS, or both, as applicable to the testing situation. Blinding avoids both potential bias in the interpretation of test results and overly optimistic estimates of diagnostic sensitivity and specificity or the area under the receiver-operating curve in diagnostic test evaluations. Procedures for blinding should be described as in the Garver et al. (2011) example or as in Thébault et al. (2005). If the joint results of the TUE and RS (positive–positive, positive–negative, negative–positive, and negative) are reviewed by testing laboratories and retesting of samples with discrepant (discordant or non-agreeing) results is done using a third test, estimates of sensitivity and specificity will change and may be inflated (Hadgu 1999). If this approach is used, sensitivity and specificity based on the original test data and the retest data should both be reported for transparency.

Statistical methods

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

Example:

As neither of the tests used in this study was perfect, estimation of the sensitivity (Se) and specificity (Sp) of the RRT-PCR [real-time reverse transcriptase PCR], VI [virus isolation] and HP [histopathology] was carried out by employing a Bayesian formulation of the latent class approach … Two models were used: one assuming conditional independence (CID) between tests given infection status, and the other allowing full conditional dependence (COC model) between RRT-PCR and VI tests given infection status … The CID and COC models were compared employing the deviance information criteria (DIC) … Using Bayesian posterior probabilities (POPR) calculated as the proportion of MC [Monte-Carlo] samples for which the hypotheses were true, we tested the one-sided hypotheses that the Se and Sp of RRT-PCR were better than those of VI and HP, respectively (Abayneh et al. 2010, p. 530).

Explanation:

Latent class analysis (LCA), as described in the example, is recognized as an appropriate method by OIE when samples of unknown status are used in diagnostic accuracy studies (OIE 2015a). To our knowledge, specific guidelines for reporting LCA studies of diagnostic accuracy have not been published, but the underlying assumptions of the model (e.g. conditional independence, see the Supplement) should be described and informative priors must be justified, if used in Bayesian analyses. The latent class (e.g. infected, infectious, or diseased) should be explicitly described by the authors because the target condition for which diagnostic sensitivity and specificity are being estimated is not implicit in the analysis. Each analysis should include a description of criteria to assess model convergence, and a sensitivity analysis should be done comparing informative versus non-informative priors if the former are used. For complex models involving covariates and/or a hierarchical structure, there should be a consideration of model identifiability (see the Supplement and Jones et al. 2010). Code for analyses should be provided in an appendix or supplemental file, or referenced. Abayneh et al. (2010) provides a good example of reporting many of the needed elements.

When traditional statistical approaches are used rather than LCA, uncertainty in estimates will depend on the number of positive and negative reference samples used in a study. These may be limited and difficult to access depending on the disease situation in a testing country. For example, ostreid herpesvirus-1 (OsHV-1) is endemic in oysters in France, and it is very difficult to find negative animals, and these may need to be obtained from countries free of infection. Hence, diagnostic sensitivity or specificity estimates in a particular study may be more or less precise depending on the sample sizes used for estimation of these parameters. Confidence intervals (CI) and their calculation methods (e.g. exact binomial, normal approximation, score method of Wilson) should be reported so that readers are fully informed about the uncertainty in estimates. Corbeil et al. (2010) provide a thorough description of statistical methods based on the assumption of a perfect RS, namely histopathology, including the software used for the analysis.

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

Example:

45 apparently healthy fish were from 3 exposed cages … 35 apparently healthy fish were from an infected cage … and 20 dead or moribund fish were from ISA [infectious salmon anemia] virus clinically affected cages … From each fish, kidney samples were collected aseptically in replicates … coded with a random identification number to blind laboratory

operators and to avoid test review bias … From each salmon, duplicate samples were sent on dry ice to the reference laboratory (lab A) to estimate the repeatability, and single samples were transported on dry ice to 2 other laboratories (labs B and C) to estimate the reproducibility … Each of the participating laboratories agreed to test for the presence and absence of ISAV using the same RT-PCR protocol provided by the reference laboratory (lab A) … (Caraguel et al. 2009, p. 11).

Explanation:

Description of a reproducibility study should include the number of replicates used for each factor investigated (e.g. samples, runs, operators, batches, laboratories), the range of analytical activity covered by the selected samples, the nature of the aliquots (crude, homogenised, extracted, diluted, spiked, 'plasmid'), whether operator(s) were blinded, any transformation of test results (ratio, categorization), and the chosen analytical approach. Details about the source and analytical activity of the selected specimen should be provided to assess how well they covered the assay's operating range and their fitness for the intended purpose (OIE 2015a). The process to aliquot and label the samples should be reported to assess if replicates were processed using identical preparation steps to those used for a routine test sample (e.g. including extraction or dilution steps). Data from proficiency testing panels but not ring trials (see the Supplement) might be suitable for preliminary reproducibility estimates subject to panel design. Therefore, the panels should be reported in sufficient detail for readers to judge the appropriateness of the samples. If the repeatability and reproducibility of the TUE have been estimated and reported previously, the relevant citation should be provided.

Analytical approaches and, therefore, reporting details differ for qualitative (binary, ordinal) and quantitative (continuous) test outcomes. Regardless, the precision of continuous results should be reported using the scale (e.g. transformation, categorization, standardization, or truncation) that will be used for the intended purpose of the assay, and appropriate tests of agreement for binary or ordinal test results should be reported (e.g. kappa with 95% CI). For a review of methods for estimating and reporting agreement for continuous test outcomes, see for example, Barnhart et al. (2007), Garver et al. (2011), and OIE (2015c). Caraguel et al. (2011) is a good example of reporting of more complex modelling methods to investigate the impact of sample or operational factors on test precision.

Results

Animals and populations

Item 14: For field studies, report when study was done, including start and end dates.

Example:

Site 2. Pancreas disease [attributable to salmonid alphavirus] was diagnosed in cage X in January 2006, 4 months after seawater transfer. The first study samples (sampling number 0) were collected in February. The majority of the histopathological-positive fish showed stage 2 lesions (chronic PD, Table 4) except one fish in cage Z showing stage 1 lesions (acute PD, Table 4). At the second sampling (sampling number 1) nearly 5 months later, only stage 2 histopathological lesions (chronic PD, Table 4) were found. At the third sampling (sampling number 2), as well as at slaughter (sampling number 3), 14 months after the initial sampling, only stage 3 lesions were detected (late/regenerative PD, Table 4). The onset of the main mortality during the outbreak occurred in May, when the seawater temperature was 10°C, with a second, smaller mortality peak also observed in November (Fig. 1b) (Jansen et al. 2010, p. 730).

Explanation:

Dates of the study and details of recruitment of populations can be described either in the 'Materials and methods' or 'Results' section, depending on author preference. For longitudinal studies of aquatic animal populations, reporting of study dates is important since mortality and morbidity events are very often seasonal, depending on rapid fluctuations in water temperatures, light exposure, salinity, and quality, and may vary with host factors such as age and size of animals. For infectious diseases, sampling of clinically affected animals will typically yield higher estimates of diagnostic sensitivity for organism detection tests; hence, reporting of clinical status and longitudinal mortality data is important. Jansen et al. (2010) provide detailed descriptions of disease progression at all 3 sites (including Site 2 as in the example) and present the longitudinal mortality and sampling data. Their Tables 2 & 3 clearly present dates of sampling, including sampling numbers (relative to time of disease diagnosis), results for the TUEs (including virus isolation), and the clinical status of tested fish.

Item 15: Report demographic and other biologically relevant characteristics of the study sample at the individual level (e.g. age, sex, ploidy, species, genotype if known, weight, and risk factors) and at the population level (e.g. production system, water quality, water temperature, and water salinity).

Example:

EHNV [epizootic haematopoietic necrosis virus]-infected samples were obtained from experimentally exposed redfin perch (*Perca fluviatilis*), Murray–Darling rainbowfish (*Melanotaenia fluviatilis*), eastern mosquitofish (*Gambusia holbrooki*), freshwater catfish (*Tandanus tandanus*), Macquarie perch (*Macquaria australasica*) and silver perch (*Bidyanus bidyanus*) from trials in which fish were exposed to 10^2–10^3 $TCID^{50}$/ml by bath exposure at 18–24°C [as described in detail in Becker et al. 2013]. Naturally infected redfin perch were obtained from different waterways in New South Wales and the Australian Capital Territory in Australia. Populations included both sexes and a variety of ages … The panel of samples included specimens of both sexes and different ages and species: redfin perch, river blackfish (*Gadopsis marmoratus*), golden perch (*Macquaria ambigua*), trout cod (*Maccullochella macquariensis*), freshwater catfish, Macquarie perch and rainbow trout (Jaramillo et al. 2012, p. 187–188).

Explanation:

In most studies based on cross-sectional sampling in a single fish species, the reported animal demographic information is limited (e.g. mean weight of fish in populations, see McClure et al. 2008). For multiple fish species, age/size and gender information, location of sample collection, and number of samples for each species should be reported. The text description and Table 3 of Jaramillo et al. (2012) provide a good example of the necessary elements. Their data were also subclassified by exposure history to EHNV (deliberately exposed, naturally exposed, probably exposed, and not known to be exposed). Warg et al. (2014b) is another example, listing in addition other known viral infections in the sampled fish (see their Table 1).

Biologically relevant summary information should be provided for all aquatic animal species because disease prevalence and intensity of pathogen exposure may depend on age, season, and location. A basic description is given in the Martenot et al. (2010) study comparing a new PCR (TUE) for OsHV-1 in Pacific oysters with a PCR reference test that was hampered by low analytical sensitivity. The demographic, temporal, and biological characteristics pertaining to the samples enable a reader familiar with OsHV-1 to determine that the samples were from hosts of a susceptible age from an endemically infected region and were collected at a time when the disease was active. Therefore, viral load in the 'positive' samples was probably representative of a typical diagnostic scenario and thereby supporting conclusions about relative sensitivity. Had the comparison been conducted at a time when infection was sub-clinical, the RS may have performed poorly due to low viral load.

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

Example:

Samples (kidney, spleen or ovarian fluid) were collected from fish at 5 hatcheries for this study and each facility was coded numerically (1, 2, 3, 4 or 5) … Samples were either plated immediately or transported on ice to the University of Idaho within 24 h of collection where they were then plated for bacterial culture, except for samples from Hatcheries 4 and 5 where logistical constraints precluded plating … For sensitivity and specificity estimates for MF-FAT [membrane filtration fluorescent antibody test], ELISA and nested PCR, analysis was limited to fish that had test results for all assays resulting in 187 fish sampled from Hatcheries 1, 2 and 3 [see their Tables 2 & 3] (Long et al. 2012, p. 409–411).

Explanation:

Studies may be undertaken using populations of aquatic animals of different infection status (infected or non-infected) as defined by the RS to obtain estimates of diagnostic sensitivity and specificity. Geographic origin may be used as a criterion for classification of populations rather than the RS. In the case of valuable stock (e.g. barramundi broodstock of high commercial value) where the RS for detection of nervous necrosis virus requires destruction of animals, assumptions may need to be made about infection status instead of applying the RS in the evaluation of new antibody-detection ELISA tests.

Test results

Item 17: Report time interval between collection of samples for the TUE and the reference standard and any interventions administered between for samples collected ante-mortem.

Example:

… sampling program on 2 farms … was initiated after one of the farms had experienced a severe outbreak of VHS which was diagnosed only one week prior to our first sampling for this study … Samplings were carried out on the 24th of March, April 30th and May 21st, 2003 … The fish were anesthetised with benzocaine and 1–2 ml of blood collected by puncture of the caudal vein using vacutainers. After blood sampling fish were euthanized, dissected and pathological signs were noted (Schyth et al. 2012, p. 594).

Explanation:

The 11 test accuracy studies in Gardner et al. (2014) were based on lethal sampling for tissue collection (post-mortem sampling). If samples are only collected post-mortem and then tested by the TUE and RS, this item is not relevant. However, in longitudinal studies evaluating tests for infectious diseases, the TUE may be initially used and then additional samples collected at post-mortem. The primary reason for reporting this item is the possibility of disease progression bias if the time interval between the TUE and the RS is biologically significant.

The importance of the time lag depends on the pathogenesis of the disease of interest and is, therefore, pathogen-specific. For example, in a hypothetical study of rectal swab culture for *Yersinia ruckeri* (enteric redmouth), an interval of weeks or months between swab collection and post mortem examination to obtain tissues for histological examination may not be significant due to the chronic nature of these infections and the persistent carrier state. In contrast, a hypothetical validation study of a serum neutralization (SN) assay for antibodies against VHS virus would need to consider stage of infection. In the convalescent stage, discordance between virus isolation results from tissues and SN test results might be expected because, over time, virus may be cleared from the host due to the immune response, but antibody titres may decline at a slower rate. Although not a true validation study, Schyth et al. (2012) covered this issue well. Similar concerns would apply to molluscs even though antibodies are not present. Attempts are being made to correlate gene expression changes and serum protein responses with disease status, and data suggest that molluscs are able to clear viral infections over periods of weeks to months (Paul-Pont et al. 2013).

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

Example:

A total of 400 farmed Atlantic salmon were collected from 4 distinct populations with different infection prevalences representing a range of infection stages according to clinical and historical information … The 4 study populations included: (i) Pop I (near-zero prevalence population) with 100 apparently healthy fish from 2 non-exposed cages (50 fish each) at distinct non-infected sites but from a region historically infected; (ii) Pop II (low prevalence population) with 130 apparently healthy fish from 3 exposed cages (20, 50 and 60 fish, respectively) at distinct sites declared infected; (iii) Pop III (moderate prevalence population) with 70 apparently healthy fish from 2 infected cages (20 and 50 fish, respectively) after an ISA outbreak at distinct sites; and (iv) Pop IV (high prevalence population) with 100 fish including a mixture of apparently healthy, dead and moribund fish from 4 infected cages (10, 14, 31 and 45 fish, respectively) during an ISA outbreak at distinct sites (Caraguel et al. 2012, p. 166).

Explanation:

Diagnostic sensitivity and specificity are population parameters that can vary according to the distribution of factors that influence stage of disease and prevalence of infection (Greiner & Gardner 2000). Infection prevalence can also vary with environmental conditions, host species, and pathogen load, but the complex relationship between these factors is poorly documented for most infectious diseases of animals.

Diagnostic sensitivity is expected to be higher when a greater quantity of the target analyte such as live virus is present in individual animals, typically during the clinical phase of infection compared with the recovery phase when animals may be apparently healthy. If the spectrum of disease severity in the study population is not representative of the spectrum of disease in the target population, the study estimates of diagnostic sensitivity and specificity may be invalid. Thus, it is important to report the former to assess a test's fitness for intended use in a target population. In fish aquaculture, the severity scale for clinical signs is often limited to 'apparently healthy', 'runt', 'moribund', or 'dead or mortality' and is used typically as a proxy for disease stages. For shellfish, the classification may only be moribund/dead versus healthy. The severity of disease in the study population is then often reported using the proportion of animals in these clinical categories (Nérette et al. 2008, Caraguel et al. 2012). Fig. 1 in Jansen et al. (2010) is also a good example because it combines histopathological findings, pathogen load, serological status and clinical signs to describe the spectrum of disease stages in the study population. Caraguel et al. (2009) provide a novel example of presentation of test results across multiple populations. Finally, concomitant disease(s) or interventions, such as vaccination or antibiotic treatment, in a study population may impact test accuracy (i.e. more false-positive or false-negative results) and should also be reported. Ideally, test accuracy estimates should be obtained for each set of relevant environmental and host conditions, but the cost to get this information might be prohibitive.

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

Example:

For continuous test results, an interactive dot diagram as shown in our Fig. 1 (Fig. 4 reproduced here from Corbeil et al. 2010) can be used to illustrate the best separation (minimal false-negative and false-positive results) between the positive and negative groups of a TaqMan PCR for abalone herpesvirus. A cut-off C_T value of 35.8 (horizontal solid line) was used and histopathology was the reference standard (0 = negative, 1 = positive).

Explanation:

Presentation of results of the TUE compared with the RS in 2 × 2 tables, sometimes categorized by population-level exposure status, is common practice in many test evaluation studies (see for example Corbeil et al. 2010, Garver et al. 2011, Jaramillo et al. 2012). However, few authors present analysis of data in a continuous form (exceptions include True et al. 2009, Corbeil et al. 2010, Purcell et al. 2013). Corbeil et al. (2010) also presented the receiver operating characteristic (ROC) curve with its 95% CI for their TaqMan PCR (see their Fig. 3). Where valid inconclusive results of the TUE occur, these can be presented in 3 × 2 tables as described in Shinkins et al. (2013).

When more than 2 tests are included in the test evaluation study, transparent reporting is more challenging. One solution is to report the frequencies of the joint test results classified as positive or negative for all possible combinations of results. In Table 1 of their paper, Abayneh et al. (2010) show the cross-tabulated results of 3 tests (histopathology, real-time RT-PCR, and virus isolation) in 2 subpopulations of fish. This format allows collapsing of data across tests and populations, if necessary, and can also be modified to account for missing results due to insufficient or poor quality samples or selective testing to reduce study costs. For example, if virus isolation were attempted on a random sample of 83 (50%) rather than all 166 samples in the Abayneh et al. (2010) study, 8 additional data rows would be needed to account for the failure to test all samples. This should be reported as described in Item 16.

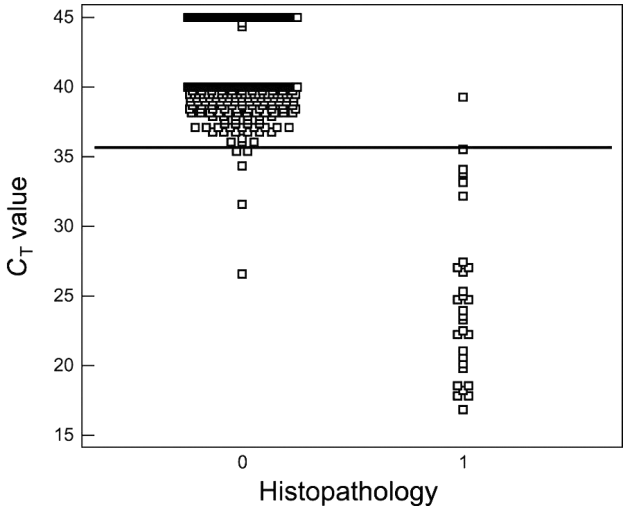

Fig. 1. Interactive dot diagram showing the distribution of TaqMan PCR results for abalone herpes virus compared with histopathology as the reference standard (0 = negative, 1 = positive). Horizontal line at a cycle threshold (C_T) of 35.8 was the cutoff for test interpretation. Reproduced from Corbeil et al. (2010; their Fig. 4)

Another solution is to create a comprehensive flow chart of sample numbers, sampling protocols, and results of the TUE versus RS to report transparently not only comparative results between assays but also possible time-varying (see Item 17), sample-dependent (see Item 16), and inadequate RS effects that may present a potential bias to interpretation of TUE results. For example, consider a modified version of the sampling method flow chart (see Item 7) from Patil et al. (2008) (Fig. 2). Here the lag time between the TUEs and reference standard testing for white spot syndrome virus (WSSV) detection, missing samples (for 2-step PCR), and different tissue sampling may be biases (see also Item 22) the authors could have addressed when discussing and interpreting the benefits of immunoblot testing on pleopod samples compared with other testing and sampling method combinations.

Item 20: Report any adverse events from performing the TUE or the reference standard for samples collected ante-mortem.

Explanation:

Similar to Item 17, if samples are only collected and tested post-mortem, this item is not relevant. However, if blood samples, gill snips, mucus scrapings, fin clips or tissue biopsies are taken for ante-mortem diagnosis, and especially if repeated over time as in an experimental challenge study, there may be harms (e.g. mortality, morbidity, and reduced growth) attributable to sampling and handling. In field settings in finfish, environmental (e.g. high temperature, low dissolved oxygen) and physiological stressors (e.g. spawning in adult migrating salmon) may increase the risk of adverse outcomes. The negative effects of collection of samples for diagnostic testing should be reported, even if individual fish are of low financial value. Table 1 of Elliott et al. (2015) showed mortality associated with 5 candidate non-lethal sampling methods for *Renibacterium salmoninarum* in Chinook salmon compared with anaesthesia only and no treatment,

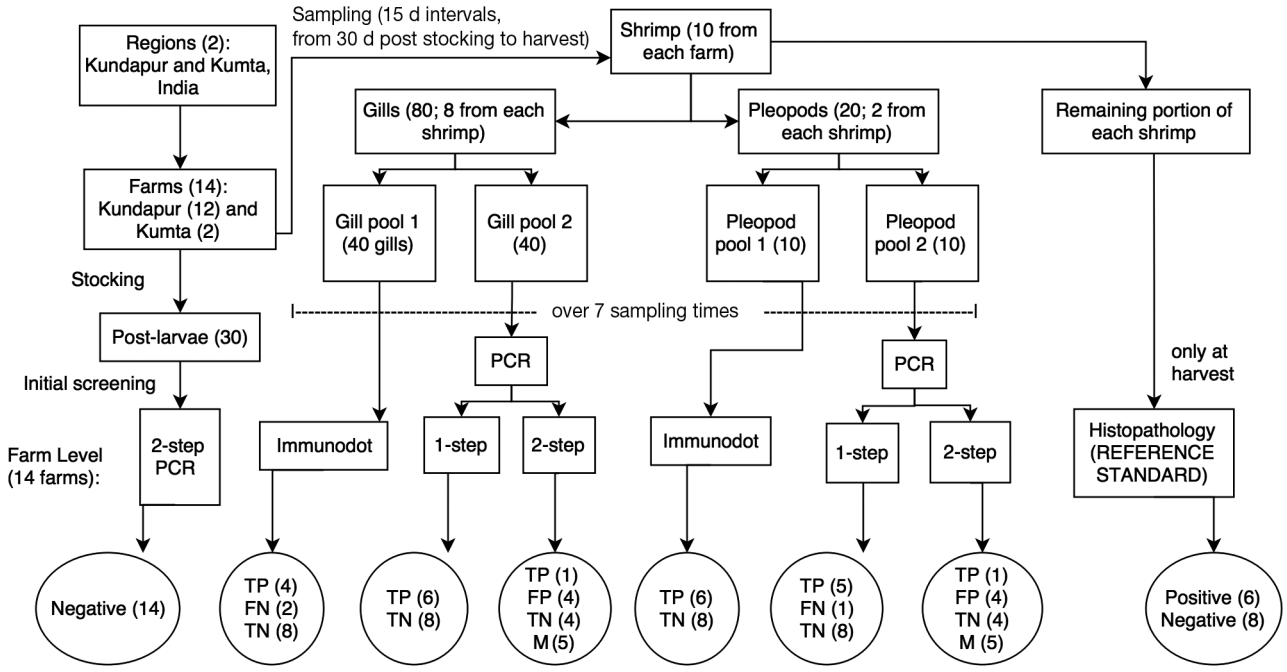

Fig. 2. Flow chart modified from Patil et al. (2008; their Fig. 1) to include results of tests under evaluation (TUEs) relative to the reference standard (see Item 19). TP: true positive (i.e. TUE positive and reference standard positive); FP: false positive (i.e. TUE negative and reference standard positive); FN: false negative; TN: true negative

thereby allowing readers to assess negative consequences of the non-lethal sampling.

Negative effects of repeated sample collection may be site-specific, systemic, or both and could impact test validation under 2 conditions. The first instance is if there was a delay between collection for the TUE and the RS as described in Item 17. However, negative consequences are important only if animals are lost from the study or if the consequences differentially affect either the TUE or RS. For example, a single blood collection from the caudal vein in finfish can lead to spinal cord damage with locomotor dysfunction. These detrimental effects may directly impact the TUE if longitudinal sampling approaches are used in individuals. Second, collection of a sample for Test A may damage tissues to be evaluated using Test B. Comparisons of skin scrapings with histopathology would be an obvious example, which would generally force the use of different sites on the host for each test.

Estimates

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

Example:

> The highest diagnostic sensitivity and specificity were 96.7 (95% CI: 82.7 to 99.4) and 99.7 (95% CI: 99.3 to 99.9), respectively, at a threshold cycle (C_T) value of 35.8 … [and] The AUC [area under the curve] of the AbHV [abalone herpesvirus] TaqMan assay was 0.998. Hence, this assay can be considered a highly accurate test (Corbeil et al. 2010, p. 1 & 7).

Explanation:

Reporting of 95% confidence intervals is standard practice in the majority of diagnostic accuracy studies in aquatic animals when reference samples are used or results of a TUE are compared with a RS that is assumed to be perfect. If a LCA analysis in a Bayesian framework is used for parameter estimation using samples from populations of unknown status, the mean (or median) sensitivity and specificity should be reported with 95% probability (credibility) intervals. These are typically derived from the output of software programs such as WinBUGS (Lunn et al. 2000) that allow ready implementation of Bayesian methods. Table 2 in Abayneh at al. (2010) provides a good example of reporting of these methods. In another well-reported study, Thébault et al. (2005) compared use of TAGS (an internet-based program to implement LCA with maximum likelihood methods, available at https://rpouillot.shinyapps.io/TAGS/), Bayesian LCA models, and traditional RS analysis to estimate diagnostic sensitivity and specificity for *Marteilia refringens* in oysters. The authors reported estimates by all 3 methods in Table 6 of their paper and discussed the comparative findings

and strengths and weaknesses of the different statistical approaches.

Item 22: Report how indeterminate results, missing responses, and outlier values of the TUE and the reference standard were handled.

Example:

> Electrophoresis gels were examined carefully, and PCR was repeated on samples where a very weak intensity band at the expected size was observed initially. If the second PCR result was positive again, the final result was positive; if not, it became negative (Caraguel et al. 2009, p. 12).

Explanation:

Inconclusive test results (see definitions in the Supplement and Shinkins et al. 2013) may be caused by sample (e.g. poor quality or low quantity), animal (e.g. concomitant disorder or therapy), or analytical factors (e.g. measurement uncertainty near the cutoff). Historically, many terms have been used somewhat interchangeably for inconclusive results including uninterpretable, intermediate, indeterminate, uninformative, suspicious, and suspect. Shinkins et al. (2013) have proposed that inconclusive results should be categorized as either invalid or valid. Invalid results include missing (e.g. insufficient or poor quality sample) and uninterpretable results (e.g. overgrowth of a culture by a non-target organism). These 2 types of results can be treated identically, and the test can be repeated if the event occurs independently of the presence or absence of disease. The frequency of these results should be reported by the RS, if used, and possible effects on the diagnostic sensitivity and specificity of TUE should be explored, and their practical implications (e.g. cost of retesting and diagnostic utility) should also be discussed.

The frequency of both types of inconclusive results is rarely reported in aquaculture studies. If common, these results will impact the usefulness of the assay in routine use. Therefore, it is important to report the frequency of inconclusive test results and how they were managed. Inconclusive samples may be retested using the same test or a confirmatory test. Any additional testing is now part of the overall detection protocol, and the details of the interpretation of each test result combination should be reported. The interpretation of several test results from the same sample can result in very different overall values for diagnostic sensitivity and specificity. For instance, addition of a confirmatory test to the TUE, interpreted in series, is likely to decrease its diagnostic sensitivity and increase its diagnostic specificity (assuming conditional independence between the 2 tests) (Gardner et al. 2000). With real-time PCR technology, it is routine to run duplicates of samples, and contradictory results (e.g. one positive and one negative) often occur when C_T values are close to the cutoff. Interpretation or retesting of contradictory samples should be detailed in the report. For example, in one of our laboratories the original raw material is used, is extracted in duplicate, and all 4 nucleic acid samples are tested.

Item 23: Report estimates of variability of diagnostic accuracy between relevant subpopulations, operators/readers, host factors, agent factors, or testing sites, if done. For challenge studies, report temporal variation in sensitivity estimates compared with days post-challenge.

Example:

Table 3 of Nérette et al. (2008) reported specificity estimates and 95% CI of a PCR for ISAV in 4 populations (apparently healthy fish in a non-outbreak cage on the same infected site or nearest neighboring site, apparently healthy fish in a sick (outbreak) cage; moribund fish in an outbreak cage; and apparently healthy fish from a population assumed to be free of ISA). Tables 5 & 7 of Hyatt et al. (2007) report the diagnostic window or capability to detect infection for the different tests evaluated. For example, they explain in one such case:

> The TaqMan assay following the wash protocol was the most efficient assay, as this technique detected infection in >50% of the infected animals as early as 7 d p.i. More than 25, 50 and 75% of infected animals were detected at Days 7, 14 and 21, respectively (Table 5). Following the wash protocol, this test achieved the highest sensitivity, i.e. 97%, at Day 35 p.i. in this experiment (Table 3). (Hyatt et al. 2007, p. 183)

Explanation:

As described in Item 18, test accuracy can vary with operational (e.g. technician, reader and laboratory) and biological factors (e.g. strain or genotype, disease stage, host profile, and population profile). When investigated, covariate-specific estimates should be reported with 95% CI. These estimates provide a better understanding of the accuracy of the test under variable conditions and potentially facilitate extrapolation of study findings to an external population or different testing purpose. For instance, if a test was evaluated for screening purposes using a mixed population with apparently healthy and moribund fish, one could use the diagnostic sensitivity and specificity estimates of moribund fish only when testing to confirm clinically suspect (moribund) cases. Other study examples that provide covariate-specific estimates for various clinical categories

include Jansen et al. (2010) and, for separate agent genotypes, Gustafson et al. (2008).

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

Example:

> Overall repeatability revealed slightly lower Pa [observed proportion of agreement] than overall reproducibility (0.81 [95% CI 0.75–0.86] and 0.82 [95% CI 0.76–0.88], respectively), although the overlap of CIs provided little evidence of significant difference (Table 2). Tests from pairwise comparisons involving lab C showed serious disagreement with the 2 other laboratories regardless of the sample type (significant McNemar's test). Estimates of k [kappa statistic] ranged from 0.57 to 0.73 and supported Pa results (Table 2) (Caraguel et al. 2009, p. 14).

Explanation:

Estimates of repeatability and reproducibility (see Supplement definitions) should be reported with 95% CI. For continuous outcomes, imprecision is most relevant for test values near the cutoff point and should be reported. The precision of continuous tests can be illustrated graphically across analyte concentration using a boxplot, concordance plot or Bland-and-Altman (see, for example, Corbeil et al. 2010, Garver et al. 2011, and Jonstrup et al. 2013). For binary outcome tests, a novel graphical approach using phylogenetic tree representation was developed to represent clustering (agreement) among test replicates (Caraguel et al. 2009). Changes in repeatability and reproducibility according to relevant sample factors (e.g. tissue homogenisation in Caraguel et al. 2009), host factors (e.g. degree of infection in Caraguel et al. 2011), operator/readers (e.g. level of experience), laboratory factors (e.g. different instruments used in different laboratories; see Hyatt et al. 2007), agent factors (e.g. strain), population factors (e.g. prevalence in Caraguel et al. 2011), or testing location should be reported if investigated.

Discussion

Item 25: Discuss the fitness for the stated purpose and the utility of the TUE in various settings (clinical, research, surveillance, etc.) in the context of the currently available tests. For challenge studies, critically evaluate the relevance of the experimental challenge study to naturally occurring infection/disease and explain why the latter source of samples was not used.

Example:

> The IQ Plus™ WSSV Kit with POCKIT system, a diagnosis assay allowing pond-side detection of WSSV, would help shrimp farmers and local offices to respond to disease outbreaks in an efficient and timely manner. For field users, equipment and accessories required to run the assay (POCKIT™, a mini-centrifuge, pipettes, and pipette tips) are combined into a mobile package (POCKIT™ Xpress) to allow great mobility of the system. Compared to shipping samples to centralized laboratories for WSSV diagnosis, the POCKIT™ assay could significantly lower the costs and shorten the sampling-to-result turn-around time from days to a few hours … Testing by IQ PlusTM WSSV Kit with POCKIT system costs around US$10 per sample, which is relatively inexpensive in comparison to the costs of sending samples to be diagnosed by standard and/or real-time PCR assays at a laboratory, which could cost more than US$50 for each target pathogen plus fees for handling and shipping (Tsai et al. 2014, p. 7).

Explanation:

Interpretation is a final and critical step in diagnostic accuracy studies in the context of fitness for purpose. Ultimately, decisions about test selection for disease or epidemiologic investigations, or how to interpret results for trade, are the end-use. In their abstract and discussion, authors should provide clear statements about the fit and constraints of their findings in the context of target populations and other available tests for a designated purpose. Considerations such as test cost, laboratory capacity, rapidity of results, and technical complexity should also be discussed in this context as they affect test choice. Candid discussions of situations in which test results can or cannot be reliably applied (e.g. populations and decision contexts) will facilitate informed decisions by readers about potential end-use.

Hyatt et al. (2007) provide a good example of this in their study of assessing 3 assays and various sampling techniques for *B. dendrobatidis* in amphibians. The authors directly mention the term 'fitness of purpose' in the first section of their study's discussion, then provide details for each aspect, mentioning which sampling methods would be better in experimental versus field applications. Their Table 13 lists their recommendations (based on their comparative analyses) for an international standard for the OIE for sampling (field, lab, pools, storage) and for using real-time TaqMan PCR for *B. dendrobatidis* in amphibians and how to interpret the results.

Potential adverse effects of tests with imperfect specificity should be explicitly described especially if aquatic animal trade is affected. Tsai et al. (2014, p. 7) stated that:

> WSSV-like sequences, occupying around 20% of the *P. monodon* genome, have been found to be present throughout the shrimp genome in a genome sequencing study … This is a cause for concern that the prob-

ability of false-positive diagnosis of white spot syndrome disease is likely to rise from cross reactivity of PCR primer and probe with target homologs within shrimp genome.

If experimental challenge studies are used for estimation of diagnostic sensitivity, authors should discuss how their experiment mimics natural exposure and disease progression. For example, a design based on cohabitation and waterborne exposure would be appropriate for abalone viral ganglioneuritis because disease progression to mortality is only a few days, and hence, it is difficult to harvest live wild abalone at different stages of infection to investigate the pathogenicity of the etiological agent. Therefore, experimental infection of healthy animals is necessary to study the interaction(s) between host and pathogen (S. Corbeil and K. A. Garver pers. obs.) as well as temporal changes in pathogen distribution in tissues and associated histopathological changes. The Jonstrup et al. (2013) study described the use of samples from fish after an acute challenge experiment to validate a test for surveillance purposes. In that study, quantification of the viral pathogen by both the TUE and the RS was used to demonstrate the presence of a range of viral quantities, including samples close to the TUE's limit of detection. Therefore, the necessary positive samples for validation were obtained for a scenario where attainment of field samples for the combination of species and pathogen was not possible because, in a surveillance to demonstrate freedom setting, negative samples are all that are available locally.

CONCLUSIONS

Design and reporting quality are interrelated aspects of test accuracy studies, but evaluation of studies in public health suggests that even well-designed studies may fail to be reported in sufficient detail to allow end-users to critically assess utility of the TUE. In a prior review of reporting quality of the published finfish papers (Gardner et al. 2014), deficiencies were identified in many items. The goal of the present study was to develop a checklist tool to improve reporting of test accuracy studies in cultured and wild aquatic animals in the context of fitness for purpose. To support their use, we provide real examples that show most, if not all, of the necessary elements for each of the 25 checklist items. We did not prioritize items because we believe that all are important and interrelated and therefore require explicit description in manuscripts.

Funding for diagnostic validation studies for aquatic pathogens is very limited in most countries. Hence, simple changes in reporting will benefit the aquatic animal health community and help to ensure that findings are useful for decision support in real-world applications. Reporting guidelines offer additional benefits to authors, reviewers and journal editors by providing a structured and consistent approach for manuscript preparation and peer-review. As an initial step to improve reporting quality, we recommend that authors follow the STRADAS-aquatic guidelines and submit the checklist as supplementary materials with their papers noting the page and line number where each item is addressed in a manuscript. Finally, we reiterate that the guidelines and checklist are generic and are applicable to all aquatic animals. The guidelines should also help researchers make more informed decisions about design given knowledge of key information to be reported.

Acknowledgements. This research was undertaken, in part, with funding from the Canada Excellence Research Chairs Program. We thank Emilie Laurin for technical assistance and an anonymous reviewer whose comments led to an improved manuscript.

LITERATURE CITED

Abayneh T, Toft N, Mikalsen AB, Brun E, Sandberg M (2010) Evaluation of histopathology, real-time PCR and virus isolation for diagnosis of infectious salmon anaemia in Norwegian salmon using latent class analysis. J Fish Dis 33:529–532

Barnhart HX, Haber MJ, Lin LL (2007) An overview on assessing agreement with continuous measurements. J Biopharm Stat 17:529–569

Becker JA, Tweedle A, Gilligan D, Asmus M, Whittington RJ (2013) Experimental infection of Australian freshwater fish with epizootic haematopoietic necrosis virus (EHNV). J Aquat Anim Health 25:66–76

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA and others (2003a) Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Chem 49:1–6

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA and others (2003b) The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 49:7–18

Branscum AJ, Gardner IA, Johnson WO (2005) Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. Prev Vet Med 68:145–163

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

Caraguel CG, Stryhn H, Gagné N, Dohoo IR, Hammell KL (2011) Selection of a cutoff value for real-time poly-

merase chain reaction results to fit a diagnostic purpose: analytical and epidemiological approaches. J Vet Diagn Invest 23:2–15

Caraguel C, Stryhn H, Gagné N, Dohoo IR, Hammell L (2012) Use of a third class in latent class modelling for the diagnostic evaluation of five infectious salmon anaemia virus detection tests. Prev Vet Med 104:165–173

Corbeil S, Colling A, Williams LM, Wong FY and others (2010) Development and validation of a TaqMan® PCR assay for the Australian abalone herpes-like virus. Dis Aquat Org 92:1–10

Elliott DG, Applegate LJ, Murray AL, Purcell MK, McKibben CL (2013) Bench-top validation testing of selected immunological and molecular *Renibacterium salmoninarum* diagnostic assays by comparison with quantitative bacteriological culture. J Fish Dis 36:779–809

Elliott DG, McKibben CL, Conway CM, Purcell MK, Chase DM, Applegate LJ (2015) Testing of candidate non-lethal sampling methods for detection of *Renibacterium salmoninarum* in juvenile Chinook salmon *Oncorhynchus tshawytscha*. Dis Aquat Org 114:21–43

Gardner IA, Stryhn H, Lind P, Collins MT (2000) Conditional dependence affects the diagnosis and surveillance of animal diseases. Prev Vet Med 45:107–122

Gardner IA, Nielsen SS, Whittington RJ, Collins MT and others (2011) Consensus-based reporting standards for diagnostic test accuracy studies for paratuberculosis in ruminants. Prev Vet Med 101:18–34

Gardner IA, Burnley T, Caraguel C (2014) Improvements are needed in reporting of accuracy studies for diagnostic tests used for detection of finfish pathogens. J Aquat Anim Health 26:203–209

Garver KA, Hawley LM, McClure CA, Schroeder T and others (2011) Development and validation of a reverse transcription quantitative PCR for universal detection of viral hemorrhagic septicemia virus. Dis Aquat Org 95:97–112

Greer AL, Collins JP (2007) Sensitivity of a diagnostic test for amphibian ranavirus varies with sampling protocol. J Wildl Dis 43:525–532

Greiner M, Gardner IA (2000) Epidemiologic issues in the validation of veterinary diagnostic tests. Prev Vet Med 45:3–22

Gustafson L, Ellis S, Bouchard D, Robinson T, Marenghi F, Warg J, Giray C (2008) Estimating diagnostic test accuracy for infectious salmon anaemia virus in Maine, USA. J Fish Dis 31:117–125

Hadgu A (1999) Discrepant analysis: a biased and an unscientific method for estimating test sensitivity and specificity. J Clin Epidemiol 52:1231–1237

Hick P, Whittington RJ (2010) Optimisation and validation of a real-time reverse transcriptase-polymerase chain reaction assay for detection of betanodavirus. J Virol Methods 163:368–377

Hick P, Tweedie A, Whittington RJ (2010) Preparation of fish tissues for optimal detection of betanodavirus. Aquaculture 310:20–26

Hick P, Schipp G, Bosmans J, Humphrey J, Whittington R (2011) Recurrent outbreaks of viral nervous necrosis in intensively cultured barramundi (*Lates calcarifer*) due to horizontal transmission of betanodavirus and recommendations for disease control. Aquaculture 319:41–52

Hyatt AD, Boyle DG, Olsen V, Boyle DB and others (2007) Diagnostic assays and sampling protocols for the detection of *Batrachochytrium dendrobatidis*. Dis Aquat Org 73:175–192

Jansen MD, Wasmuth MA, Olsen AB, Gjerset B and others (2010) Pancreas disease (PD) in sea-reared Atlantic salmon, *Salmo salar* L., in Norway; a prospective, longitudinal study of disease development and agreement between diagnostic test results. J Fish Dis 33:723–736

Jaramillo D, Tweedie A, Becker JA, Hyatt A, Crameri S, Whittington RJ (2012) A validated quantitative polymerase chain reaction assay for the detection of ranaviruses (Family *Iridoviridae*) in fish tissue and cell cultures, using EHNV as a model. Aquaculture 356-357: 186–192

Jones G, Johnson WO, Hanson TE, Christensen R (2010) Identifiability of models for multiple diagnostic testing in the absence of a gold standard. Biometrics 66:855–863

Jonstrup SP, Kahns S, Skall HF, Boutrup TS, Olesen NJ (2013) Development and validation of a novel Taqman-based real-time RT-PCR assay suitable for demonstrating freedom from viral haemorrhagic septicaemia virus. J Fish Dis 36:9–23

Kent ML, Benda S, St-Hilaire S, Schreck CB (2013) Sensitivity and specificity of histology for diagnoses of four common pathogens and detection of nontarget pathogens in adult Chinook salmon (*Oncorhynchus tshawytscha*) in fresh water. J Vet Diagn Invest 25:341–351

Limmathurotsakul D, Turner EL, Wuthiekanun V, Thaipadungpanit J and others (2012) Fool's gold: why imperfect reference tests are undermining the evaluation of novel diagnostics: a reevaluation of 5 diagnostic tests for leptospirosis. Clin Infect Dis 55:322–331

Long A, Polinski MP, Call DR, Cain KD (2012) Validation of diagnostic assays to screen broodstock for *Flavobacterium psychrophilum* infections. J Fish Dis 35:407–419

Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000) WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 10:325–337

Martenot C, Oden E, Travaillé E, Malas JP, Houssin M (2010) Comparison of two real-time PCR methods for detection of ostreid herpesvirus 1 in the Pacific oyster *Crassostrea gigas*. J Virol Methods 170:86–89

McClure C, Saksida S, Karreman G, Constantine J, Robinson J, Traxler G, Hammell L (2008) Evaluation of a reverse transcriptase polymerase chain reaction test and virus isolation on field samples collected for the diagnosis of infectious hematopoietic necrosis virus in cultured Atlantic salmon in British Columbia. J Aquat Anim Health 20:12–18

Nérette P, Stryhn H, Dohoo I, Hammell L (2008) Using pseudogold standards and latent-class analysis in combination to evaluate the accuracy of three diagnostic tests. Prev Vet Med 85:207–225

OIE (World Organisation for Animal Health) (2015a) Manual of diagnostic tests for aquatic animals 2015, Chap 1.1.2. Principles and methods of validation of diagnostic assays for infectious diseases. Available at www.oie.int/international-standard-setting/aquatic-manual/access-online/ (accessed January 15, 2016)

OIE (World Organisation for Animal Health) (2015b) Manual of diagnostic tests for aquatic animals 2015, Chap 1.1.1. Quality management in veterinary testing laboratories. Available at www.oie.int/international-standard-setting/aquatic-manual/access-online/ (accessed October 27, 2015)

OIE (World Organisation for Animal Health) (2015c) Manual of diagnostic tests and vaccines for terrestrial animals 2015. Guideline 3.6.5. Statistical approaches to valida-

tion. www.oie.int/fileadmin/Home/eng/Health_standards/tahm/GUIDELINE_3.6.5_STATISTICAL_VALIDATION.pdf. (accessed October, 2015).

Patil R, Palaksha KJ, Anil TM, Guruchannabasavanna and others (2008) Evaluation of an immunodot test to manage white spot syndrome virus (WSSV) during cultivation of the giant tiger shrimp *Penaeus monodon*. Dis Aquat Org 79:157–161

Paul-Pont I, Dhand NK, Whittington RJ (2013) Influence of husbandry practices on OsHV-1 associated mortality of Pacific oysters *Crassostrea gigas*. Aquaculture 412–413: 202–214

Purcell MK, Thompson RL, Garver KA, Hawley LM and others (2013) Universal reverse-transcriptase real-time PCR for infectious hematopoietic necrosis virus (IHNV). Dis Aquat Org 106:103–115

Purcell MK, Hard JJ, Neely KG, Park LK, Winton JR, Elliott DG (2014) Genetic variation in bacterial kidney disease (BKD) susceptibility in Lake Michigan Chinook salmon and its progenitor population from the Puget Sound. J Aquat Anim Health 26:9–18

Ramilo A, Navas JI, Villalba A, Abollo E (2013) Species-specific diagnostic assays for *Bonamia ostreae* and *B. exitiosa* in European flat oyster *Ostrea edulis*: conventional, real-time and multiplex PCR. Dis Aquat Org 104:149–161

Rimmer AE, Becker JA, Tweedie A, Whittington RJ (2012) Development of a quantitative polymerase chain reaction (qPCR) assay for the detection of dwarf gourami iridovirus (DGIV) and other megalocytiviruses and comparison with the Office International des Epizooties (OIE) reference PCR protocol. Aquaculture 358–359:155–163

Schyth BD, Ariel E, Korsholm H, Olesen NJ (2012) Diagnostic capacity for viral haemorrhagic septicaemia virus (VHSV) infection in rainbow trout (*Oncorhynchus mykiss*) is greatly increased by combining viral isolation with specific antibody detection. Fish Shellfish Immunol 32:593–597

Shinkins B, Thompson M, Mallett S, Perera R (2013) Diagnostic accuracy studies: how to report and analyse inconclusive test results. BMJ 346:f2778

Snow M, McKay P, Matejusova I (2009) Development of a widely applicable positive control strategy to support detection of infectious salmon anaemia virus (ISAV) using Taqman real-time PCR. J Fish Dis 32:151–156

Thébault A, Bergman S, Pouillot R, Le Roux F, Berthe FC (2005) Validation of *in situ* hybridization and histology assays for the detection of the oyster parasite *Marteilia refringens*. Dis Aquat Org 65:9–16

True K, Purcell MK, Foott JS (2009) Development and validation of a quantitative PCR to detect *Parvicapsula minibicornis* and comparison to histologically ranked infection of juvenile Chinook salmon, *Oncorhynchus tshawytscha* (Walbaum), from the Klamath River, USA. J Fish Dis 32:183–192

Tsai YL, Wang HC, Lo CF, Tang-Nelson K and others (2014) Validation of a commercial insulated isothermal PCR-based POCKIT test for rapid and easy detection of white spot syndrome virus infection in *Litopenaeus vannamei*. PLoS ONE 9:e90545

Waltzek TB, Miller DL, Gray MJ, Drecktrah B and others (2014) New disease records for hatchery-reared sturgeon. I. Expansion of frog virus 3 host range into *Scaphirhynchus albus*. Dis Aquat Org 111:219–227

Warg JV, Clement T, Cornwell ER, Cruz A and others (2014a) Detection and surveillance of viral hemorrhagic septicemia virus using real-time RT-PCR. I. Initial comparison of four protocols. Dis Aquat Org 111:1–13

Warg JV, Clement T, Cornwell ER, Cruz A and others (2014b) Detection and surveillance of viral hemorrhagic septicemia virus using real-time RT-PCR. II. Diagnostic evaluation of two protocols. Dis Aquat Org 111:15–22

Whittington RJ, Steiner KA (1993) Epizootic haematopoietic necrosis virus (EHNV): improved ELISA for detection in fish tissues and cell cultures and an efficient method for release of antigen from tissues. J Virol Methods 43: 205–220

Editorial responsibility: Jeff Cowley, Brisbane, Queensland, Australia

Submitted: August 5, 2015; Accepted: November 5, 2015
Proofs received from author(s): February 12, 2016