The following supplement accompanies the article

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

Ian A. Gardner*, Richard J. Whittington, Charles G. B. Caraguel, Paul Hick, Nicholas J. G. Moody, Serge Corbeil, Kyle A. Garver, Janet V. Warg , Isabelle Arzul, Maureen K. Purcell, Mark St. J. Crane, Thomas B. Waltzek, Niels J. Olesen, Alicia Gallardo Lagno

*Corresponding author: iagardner@upei.ca

Diseases of Aquatic Organisms 118: 91–111 (2016)

Supplement. Definitions and terms, modified from STRADAS-paraTB (Gardner et al., 2011)

Animal-level (diagnostic) sensitivity: Probability that an animal with the target condition (e.g. infection or disease) will yield a positive test result with the test of interest.

Animal-level (diagnostic) specificity: Probability that an animal without the target condition (e.g. infection or disease) will yield a negative test result with the test of interest.

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

Conditional dependence: A relationship between 2 or more tests in which the sensitivity (or specificity) of the second and subsequent tests depends on whether the first test result is positive or negative, conditional on infection status (see Gardner et al. 2000 for formal probabilistic description of the relationship).

Conditional independence: A relationship between 2 or more tests in which the sensitivity (or specificity) of the second and subsequent tests does *not* depend on whether the first test result is positive or negative, conditional on infection status.

Identifiability: A statistical model lacks identifiability or is described as non-identifiable when the data distribution can be generated by more than a single set of parameter values. Historically, it was believed that, if the degrees of freedom (df) in the data met or exceeded the number of parameters in a fitted model, the model was identifiable, but this may not always be true (Jones et al. 2010). Assuming conditional independence between the tests, a 2-test, 2-population Bayesian model is identifiable when estimating sensitivity, specificity and prevalence from a multinomial sampling design, because the data have 6 df and there are 6 parameters in the model (2 sensitivities, 2 specificities and 2 prevalences) to be estimated. Conditional dependence adds 2 additional parameters (sensitivity and specificity covariances), which results in lack of identifiability unless prior information about at least 2 parameters is included in the model.

Inconclusive test result: There are 2 types of inconclusive test results. *Invalid* inconclusive results include uninterpretable outcomes or missing values (see definition of ‘uninterpretable test result’). *Valid* inconclusive results (sometimes also termed indeterminate, intermediate, suspicious, or suspect) include a test outcome that cannot be classified as either negative or positive. This classification is based on the use of 2 cutoff values which define 3 categories of test results (positive, indeterminate, and negative), where the indeterminate range can reflect measurement uncertainty. For more details, see Shinkins et al. (2013)

Infected (non-infected) animal: An animal that has (does not have) the target organism in its tissues or organs. The terms 'infected' and 'carrier' can be used interchangeably.

Infected (non-infected) population: A population that has at least one (zero) infected animal in the epidemiological unit (e.g. net-pen, site, zone, etc.).

Latent class analysis (LCA): A statistical method which allows estimation of the diagnostic sensitivity and specificity of tests without designation of a perfect reference standard, as required in Item 9. LCA methods use maximum likelihood or Bayesian approaches, with the latter allowing the investigator to incorporate relevant prior information about the sensitivity and specificity of one or more tests or about the prevalence of infection in the tested populations. With large sample size, maximum likelihood and Bayesian methods yield similar inferences. LCA allows demonstration that a novel test has superior accuracy over an existing test (e.g. PCR compared with virus isolation), while this is not possible using classical statistical methods based on a perfect reference standard. For further description of models, inferences, and examples of WinBUGS code for running Bayesian models, see Branscum et al. (2005).

Populations and study samples: The *target* population is the population to which the test accuracy estimates might be extrapolated. The *source* population is the population from which the study sample is selected. The *study* population (sometimes termed the study sample) is the sample of sites, net-pens, cages, individual animals, etc., which are included in the diagnostic accuracy study.

Population-level sensitivity (specificity): The probability that the population-level test result is positive (negative) in an infected (non-infected) population. The relationship between individual-animal and population-level estimates of diagnostic sensitivity and specificity when results of multiple samples are interpreted is described in Martin et al. (1992).

Population-level test: A population-level test is any testing strategy to classify the status (e.g. infected or non-infected) of the population. In the case of salmon aquaculture, the designated population might be a net-pen (cage), a site with multiple net-pens, or a zone based on multiple sites with hydrological connectivity. Analogous scenarios exist for molluscs, amphibians and crustaceans. Population-level tests can be based on individual specimens collected ante- or post-mortem, or tissue and organ specimens tested in artificially created pools. A population-level test result requires designation of the number or percentage of positive test results (threshold) required for classification of the population as positive. The threshold might be a single positive sample, but where multiple individual specimens are used, this threshold may exceed one because of the need for higher population-level specificity.

Prevalence: The proportion of positive test results in randomly sampled animals is an estimate of the apparent (test) prevalence of the target condition (e.g. infection) in the source population. The true prevalence of the target condition in a population can be estimated from apparent prevalence by adjusting for diagnostic test sensitivity and specificity.

Proficiency test: Proficiency testing (PT) is the determination of a laboratory's accuracy of analytical measurements by testing specimens (sometimes termed check sample/test panels) of undisclosed analyte content but with an external standard of quality. Participation in external PT enables the laboratory to benchmark the reliability of results of individual technicians compared with those from other participating laboratories, including the laboratory distributing samples, and when done annually with its own past performance. Typically, each laboratory will use a standardized operating protocol developed and validated to detect a specific analyte. Participation in PT schemes is a requirement for accredited laboratories and provides an independent assessment of the testing methods used and the level of staff competence.

Random sample: A set of animals drawn from a source population using a formal selection process in which each animal has a known, non-zero probability of inclusion in the sample.

Reference standard/test (RS): A RS is usually considered to be the most accurate test or combination of tests for correct classification of a sample result for infected (diseased) and non-infected (non-diseased) animals. For OIE-listed diseases, a RS is sometimes interpreted as either a recommended or standard method in the disease-specific chapters of the OIE (2015a) *Manual of Diagnostic Tests for Aquatic Animals 2015* (www.oie.int/international-standard-setting/aquatic-manual/). The American Fisheries Society Blue Book also provides guidance about appropriate comparator tests for certain pathogens of finfish and shellfish (AFS-FHS 2014). The term 'gold standard' (which implies perfect sensitivity and specificity of the reference standard) has been extensively used in test accuracy studies of animals, but this term is misleading and is no longer recommended for use.

Reference samples and populations: Samples of known infection status (e.g. infected or non-infected) based on results of the reference standard and epidemiological information. Reference samples should reflect both the target analyte and matrix (e.g. tissue or organ) in which it is commonly found in populations where the test will be use. Non-infected samples will typically be obtained from populations in which all animals are truly non-infected because they had no possible infection or exposure to the agent (negative reference populations), and infected samples (true positives) will be derived from naturally infected populations with variable infection prevalences or experimentally challenged animals.

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

Reproducibility: A measure of the precision (variability) of test results among laboratories or between laboratories and field-based sites, following the same testing protocol. It is estimated as for repeatability. Estimates of reproducibility allow assessment on how consistently the test might perform under variable laboratory conditions. Generally, among-laboratory variation (reproducibility) in test results will exceed within-laboratory variation (repeatability). For more details of statistical methods for evaluation of reproducibility data, see OIE (2015b).

Ring trial: If proficiency testing (PT) schemes are not available, a ring trial organized by a reference laboratory may be used as an alternative. Ring testing differs from PT because it allows for use of different methods among participating laboratories. Hence, a ring test evaluates the ability of a laboratory to detect or diagnose a particular agent or analyte with their method of choice. Ring testing therefore is not relevant to reproducibility assessment. (http://www.oie.int/fileadmin/Home/eng/Health_standards/aahm/current/1.1.02_VALIDATION.pdf)

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

Specimen size: For organism detection tests, the volume, weight, or dimensions of a sample matrix collected for evaluation with the TUE and/or reference standard. For antibody detection tests, specimen size is usually expressed as a volume. The analytical unit size that is tested by the TUE or RS is typically smaller than the specimen size.

Subclinical infection: An infected animal that appears clinically normal and that does not have overt clinical signs (detectable by the owner, site manager, or veterinarian) compatible with the target condition.

Target analyte or analytical target: Biological marker or substance (e.g. organism, nucleic acid, protein) that the TUE or RS detects as a proxy of the target condition. This differs depending on whether the test is based on direct organism detection (e.g. culture, detection of nucleic acid by PCR, or surface antigen by capture ELISA) or indirect indicators of past or current infection (e.g. serum antibodies).

Target condition (or diagnostic target): Theoretical status of interest (e.g. infected versus non-infected) of individuals or populations of aquatic animals.

Test under evaluation (TUE): A diagnostic method with potential discriminatory ability to classify the animal or population with regard to the target condition (e.g. infected versus non-infected). In STARD (Bossuyt 2003), the TUE is termed the 'index test'. An imperfect RS can also be a TUE if included in a latent class analysis of test results.

Uninterpretable test result: A test result that is not acceptable or invalid for technical reasons and is termed an invalid inconclusive result in Shinkins et al. (2013). Examples include overgrowth of cultures by contaminants that preclude counting of the target organisms, or assay results where the positive and negative controls on a plate, chip, etc. do not yield expected results, thereby necessitating re-running of assays. Criteria for the latter should be described.

LITERATURE CITED

AFS-FHS (American Fisheries Society-Fish Health Section) (2014) Suggested procedures for the detection and identification of certain finfish and shellfish pathogens, 2014 edn. http://afs-fhs.org/bluebook/bluebook-index.php

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA and others (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Chem 49:1–6 PubMed doi:10.1373/49.1.1

Branscum AJ, Gardner IA, Johnson WO (2005) Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. Prev Vet Med 68:145–163 PubMed doi:10.1016/j.prevetmed.2004.12.005

Gardner IA, Stryhn H, Lind P, Collins MT (2000) Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. Prev Vet Med 45:107–122 PubMed doi:10.1016/S0167-5877(00)00119-7

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 PubMed doi:10.1016/j.prevetmed.2011.04.002

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 PubMed doi:10.1111/j.1541-0420.2009.01330.x

Martin S, Shoukri M, Thorburn M (1992) Evaluating the health status of herds based on tests applied to individuals. Prev Vet Med 14:33–44 doi:10.1016/0167-5877(92)90082-Q

OIE (World Organisation for Animal Health) (2015a) Manual of diagnostic tests for aquatic animals 2015. Available at www.oie.int/international-standard-setting/aquatic-manual/access-online/ (accessed February 12, 2016)

OIE (World Organisation for Animal Health) (2015b) Manual of diagnostic tests and vaccines for terrestrial animals 2015. Guideline 3.6.5. Statistical approaches to validation. www.oie.int/fileadmin/Home/eng/Health_standards/tahm/GUIDELINE_3.6.5_STATISTICAL_VALIDATION.pdf. (accessed February 12, 2016)

Shinkins B, Thompson M, Mallett S, Perera R (2013) Diagnostic accuracy studies: how to report and analyse inconclusive test results. BMJ 346:f2778 doi:10.1136/bmj.f2778 PubMed doi:10.1136/bmj.f2778