

Assay Validation Methods

This document contains educational materials developed by Dr. Larry Hammell, and is used with permission. This information was adapted from sections of the 'Applied Aquaculture Epidemiology' course, developed by the Atlantic Veterinary College and Canadian Aquaculture Institute. The USFWS acknowledges and thanks Dr. Hammell for the use of this material and his contributions to quality assurance in applied aquaculture.

Testing a Test – Do we ever really have a “Gold Standard”?

Establishing gold standards for diseased and disease free individuals has been the most difficult dilemma in evaluating diagnostic tests for many diseases. For example, a positive bacterial culture rarely produces false positives (with confirmatory testing), however culture is considered lacking for detecting all the true-positives (e.g., bacteria die after removal from fish, in transport, or are overgrown by competing bacteria or fungus). So, the culture method represents a reasonable “gold standard” when it is positive, but it cannot be used as a gold standard if the test is negative (there may be many true-positive fish that do not culture and are reported as negative).

As diagnosticians, we often rely on the best available method, for example when DNA is detected by a DNA probe or PCR assay, it is generally accepted that this test result cannot be wrong when conducted properly. It may be used as a gold standard against which all other assay comparisons are made. However, if DNA testing does not tell us anything about the infection level (e.g., tests results are either positive or negative), this assay may not meet our objective for an appropriate diagnostic test.

Another option for testing validation is to use several available test methods for a disease of interest, and define the true-positives in a relative manner. The major problem here is labor and costs, and the fact that some tests may be testing for different things (i.e., protein, DNA or viable organisms), and sub-clinical infections are more likely to be missed on one or more of the tests. Identification of fish with sub-clinical infections is most often the intent of screening methods, so detection at this level must be addressed by the test methods.

Often the term “gold standard” is applied inappropriately, for only tests which measure a highly specific component, and have been validated quantitatively would meet the standard. Often, tests are developed, demonstrate usefulness in detecting a target pathogen at some infection level, and become accepted by the scientific community as “valid methods”.

Screening versus Diagnostic Tests

It is important to recognize if the animals being sampled truly represent the disease-positive and disease-free state of the population, otherwise a test may be perceived to be more useful than it is. During validation of a test method, it is critical to test animals that represent the level of infection for which the assay will be applied in the field setting. A screening test for sub-clinical infections will not have the same function as a diagnostic test for clinical disease. Evaluating a new technique in clinically diseased fish may lead to the conclusion that the test is equally effective in detecting low loads of the target pathogen. Be cognizant that there are three groups to consider when representing an aquatic population: infected with gross pathology (clinical signs and/or lesions), infected with no gross pathology, and non infected individuals.

Observer Bias

It is also important that anyone performing a test validation be blinded to the true status of the sample materials. Bias does not occur as a conscious decision to be influenced by previous knowledge; it is a subconscious effort and therefore needs to be controlled for. The extent of agreement between tests, measured by *kappa*, is often lower when “non-blind” methods of evaluation are used (Martin and Bonnett, 1987). Even prior knowledge of the approximate true prevalence can result in subtle adjustments in the interpretation of the test results. Evaluation of a diagnostic test should have a random order between true positives and true negatives. Preferably, the diagnostic laboratory should not be aware when the evaluation of a test is being performed so that personnel will treat the samples as routinely as possible, avoiding increased attention and special treatment that samples would not normally receive during routine testing.

Sample Size

Evaluation of a diagnostic test to determine its sensitivity and specificity requires the estimation of a proportion or likelihood ratios at each test outcome level (Simel, Samsa, and Matcher, 1991). In theory, a representative sample of 100-200 diseased animals and 2000 or more non-diseased animals should give reasonably precise point estimates for sensitivity and specificity, respectively (Martin, 1984).

Accuracy and Precision

Accuracy refers to the ability of the test to give a true indication of the nature and quantity of the substance or object being measured (Martin, 1977). Accuracy can be low without affecting the sensitivity and specificity (see several examples in the Interpreting a Diagnostic Test section. Also, bear in mind that a test may be 100% accurate, but be of little value if it is measuring a meaningless parameter for a disease of interest.

Evaluation of the accuracy of a test usually is performed by the molecular biologist and is often referred to as the “sensitivity” of the test. Since clinical decisions are often based upon the dichotomous values of a test (negative and positive category), the accuracy is of concern only as one of several influences on sensitivity and specificity. Within limits, accuracy is less important than precision in terms of screening tests (Martin, Meek, and Willeberg, 1987).

Precision refers to the ability of the test to give consistent results in repeated determinations in the same sample or animal (Martin, 1977). To evaluate precision, duplicate testing of the same fish tissues should be performed by the same laboratory, and between different laboratory staff. There can be poor repeatability of test results between laboratory staff when standardized procedures are not clearly defined and followed. Repeatability between laboratories can also be tested when the persons performing the testing are blinded to the sample status.

References

Martin, S.W. (1977). The evaluation of tests. *Can Jour Comp Med* 41:19-25.

Martin, S.W. (1984). Estimating disease prevalence and the interpretation of screening test results. *Prev Vet Met* 2: 463-472

Martin, S.W., and B. Bonnett. (1987). Clinical epidemiology. *Can Vet J* 28:318-325

Martin, S.W., A.H. Meek, and P. Willeberg (1987). *Veterinary Epidemiology – Principles and Methods*. Iowa State University Press, Ames, Iowa. 343 p.