Epidemiologic issues in the validation of veterinary diagnostic tests

M. Greiner\textsuperscript{a,}\textsuperscript{*}, I.A. Gardner\textsuperscript{b}

\textsuperscript{a}Department of Tropical Veterinary Medicine and Epidemiology, Institute for Parasitology and Tropical Veterinary Medicine, Freie Universität Berlin, Königsweg 57, D-14163 Berlin, Germany

\textsuperscript{b}Department of Medicine and Epidemiology, School of Veterinary Medicine, University of California, Davis, CA 95616-8737, USA

Abstract

In this review, we critically discuss the objectives, methods and limitations of different approaches for the validation of diagnostic tests. We show (based on published data and our own experiences) that estimates for the diagnostic sensitivity and specificity may vary among populations and/or subpopulations of animals, conditional on the distribution of influential covariates. Additional variability in those parameter estimates may be attributable to the sampling strategy. The uncertainty about diagnostic parameters is of concern for the decision-maker in the context of clinical diagnosis or quantitative risk assessment as well as for the epidemiologist who uses test data for prevalence estimation or risk-factor studies. Examples for the calculation of diagnostic parameters are presented together with bias-avoidance strategies. We suggest guidelines for an epidemiologic approach to test validation of veterinary diagnostic tests. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Diagnostic test; Validation; Bias; Sensitivity; Specificity

1. Introduction

In many countries, there is increased interest by animal-health officials in appropriate methods for validation of tests because of trade, regionalization, and budgetary constraints associated with disease surveillance and monitoring. Reliable estimates of the diagnostic sensitivity and specificity are needed in many settings. Veterinary practitioners require sensitivity and specificity estimates to update clinical inferences and
diagnosticians acknowledge consideration of accuracy in test selection according to a quality assurance policy. Planners of surveillance programs need sensitivity and specificity estimates for sample-size calculations; epidemiologists would like to adjust prevalence estimates, odds ratios and other study parameters for misclassification. Decision-makers and risk analysts need those estimates for complete specification of probabilities in their decision-trees and scenario pathways, respectively.

The principles of test validation and interpretation are linked to various statistical concepts including probability theory, regression and discriminant analysis, information theory and neural networks (reviewed by Kazmierczak and Catrou, 1995). Considering the many scenarios in which diagnostic tests are validated and used in medicine, veterinary medicine, social sciences and phytopathology, it is not surprising that the related terminology is inconsistent and that a single standard for test validation in practice has not been agreed to. Laboratory diagnosticians have tended to adopt the terminology and concepts of test validation used in clinical chemistry (e.g. Hartmann et al., 1994; Dybkaer, 1995). The OIE (1996) (see also Jacobson, 1998) has published a laboratory oriented guideline for the process of assay validation in veterinary medicine. The OIE guideline emphasizes that a definite test protocol (as a result of preliminary feasibility studies and efforts for test standardization) must be developed prior to the characterization of assay performance (i.e. estimation of accuracy parameters) and that ongoing efforts are required to maintain accuracy. Principles of standardization, quality assurance and validation of serological enzyme-linked immunosorbent assays were addressed by Wright et al. (1993).

The performance of a diagnostic test is described by two independent measures: precision and accuracy. Precision refers to the closeness of agreement among repeated measurements of the same sample under prescribed conditions and accuracy to the closeness of agreement between the result of a measurement and the true value of the measured (Dybkaer, 1995). A definition of accuracy that is more related to diagnostic tests is provided in Appendix A. We use the term accuracy as a synonym for validity, although some authors prefer the latter term. Some features (for example, costs and riskiness of the test) are not related to the measurement process but may be highly relevant to test selection. The latter issues are beyond the scope of our paper. We shall expand on the accuracy aspect (stage 3 of the OIE guideline) and we focus attention on the uncertainty (in the parameter estimates) which is attributable to the sampling process.

Many important diagnostic tests in veterinary medicine are non-serological (for example, skin hypersensitivity tests or the direct detection of pathogens using microbiological methods). Following an idea proposed by the information-theorists (e.g. Somoza and Mossman, 1992), we may regard any device that reduces the uncertainty about the state of disease as a diagnostic test. Relevant states also include infected/non-infected, immunologically protected/non-protected, and pregnant/non-pregnant. In our review, we consider diagnostic tests in their broadest context and do not differentiate between screening and diagnostic tests — notations used in medicine for case detection in asymptomatic and confirmation in symptomatic individuals, respectively. We assume that the diagnostic test follows a standard operating protocol (i.e. the test execution is documented in detail), that the procedure is stable within a
certain tolerance and that preliminary estimates of sensitivity and specificity are available.

Existing guidelines for validation of veterinary diagnostic tests are not very explicit as far as epidemiologic considerations are concerned. Our intention is to describe empirical evidence for the variability in diagnostic sensitivity and specificity among (sub)populations (Section 2), and to review the objectives, methods and limitations of different approaches for test validation (Section 3). We emphasize epidemiologic approaches for estimation of operational test-performance parameters (Section 3.2), and develop a set of guidelines that can be used for validation of veterinary diagnostic tests (Section 4). In Appendix A, we present a glossary of terms and notations related to diagnostic testing. We do not consider multiple (or bulk sample) tests, aggregate (e.g. herd-level) testing and comparison of tests because these issues are addressed by Gardner et al. (2000), Christensen and Gardner (2000) and Greiner et al. (2000), respectively.

2. Influential factors for sensitivity and specificity

Sensitivity and specificity denote the conditional probabilities of a positive test result given the animal is truly diseased and the conditional probability of a negative test result given the animal is truly non-diseased, respectively (see glossary in Appendix A). It is a common observation that sensitivity and specificity estimates vary among published validation studies. This variation is mainly attributable to differences among the reference populations and sampling strategies that have been used for the validation procedure. However, true differences of test accuracy among studies are not directly observable because studies are not free of random and systematic errors. Other important factors include technical variation of test characteristics (among laboratories; by time), laboratory proficiency, choice of gold standard and cut-off value for interpretation, and handling of intermediate or uninterpretable results. Quantitative tools have been developed to summarize the diagnostic accuracy across individual primary studies (Irwig et al., 1994b) and to investigate the impact of covariate factors (e.g. Greiner et al., 1997b). In the medical literature, problems of selection and spectrum bias have been described in the context of the enrollment process of patients for test validation studies (Knottnerus and Leffers, 1992; Moons et al., 1997). The variation in sensitivity and specificity among subpopulations or even among farms results in a correlation of test errors that should be considered when applying tests at the aggregate (herd) level (Donald et al., 1994; see also Gardner et al., 2000).

Empirical evidence suggests that sensitivity and specificity vary within and among animal populations. Strictly speaking, the diagnostic sensitivity and specificity are population parameters that describe the test performance for a given reference population (having a given distribution of covariate factors) under defined conditions (laboratory, gold standard, cut-off, etc.). In this section, we discuss true differences among (sub)populations; biases are considered below (Section 3.2.1). An effect of factors such as stage of disease and animal age is biologically plausible for many test systems. In practice, the factors, their impacts and their interrelationships are mostly unknown and sensitivity and specificity estimates are average values calculated in non-homogenous
populations. We contend that stratum-specific estimates have greater diagnostic utility than crude (pooled) estimates; accordingly, we recommend that these values be calculated wherever possible (i.e. where covariate information exists). Factors that potentially influence the analyte (e.g. specific antibody in ELISA) concentration have been referred to as biological factors and their impact cannot be remedied technically (Greiner and Böhning, 1994; Greiner et al., 1997a). On a population level, biological factors affect the distribution of the test value and, consequently, also the diagnostic sensitivity and specificity. Some examples of known factors that affect sensitivity and specificity are described below. A systematic review of factors that affect antigen detection tests, bacterial culture, virus isolation and antibody tests is beyond the scope of the paper.

2.1. Biological factors affecting the sensitivity of a diagnostic test

A reasonable assumption for serologic tests is that the sensitivity varies with the stage of infection or with the immune status of the host. Therefore, the case mix in the diseased reference sample will influence the sensitivity estimate. It is known, for example, that tuberculosis skin testing has a low sensitivity in preallergic, inactive and anergic stages of the infection. The sensitivity of the tuberculin test, therefore, will vary with the proportion of non-responders in the population. Markedly different sensitivities of an adsorbed antibody ELISA for diagnosis of Johne’s disease (*Mycobacterium paratuberculosis*) have been described for cattle in stage I (preclinical/not shedding), II (preclinical/shedding) and III (clinical/shedding) of the infection (Ridge et al., 1991; Sockett et al., 1992). The sensitivity of a *Mycobacterium bovis* ELISA in badgers was reported to be positively associated with female gender and the presence of visible tuberculous lesions (Clifton-Hadley et al., 1995). The results of a serologic study on *Neospora*-associated abortions in cattle provide preliminary evidence that the sensitivity of an antibody ELISA test could be higher in the endemic than in the epidemic form of *Neospora*-associated abortions (Schares et al., 1999). The duration of infection (time since infection) affects the sensitivity of ELISA serodiagnosis of porcine trichinellosis by ELISA (Greiner et al., 1997b). For heartworm (*Dirofilaria immitis*) serodiagnosis in dogs, parasite numbers, the location in the host tissue, breed and the prevalence affect the test sensitivity (Courtney and Cornell, 1990). We note at this stage that no formal relationship exists between prevalence and sensitivity. However, Brenner and Gefeller (1997) assumed that the underlying distribution of ordinal and continuous test results relative to the diagnostic cut-off value and thus sensitivity and specificity are prevalence-related. We also assume that the distribution of biological factors related to infection (stage and severity of disease) is different in low- and high-prevalence populations. Influential factors for sensitivity may vary over time. Time-dependent changes in the sensitivity may be predictable in epidemic situations, where knowledge exists on the change in the proportions of disease stages over time.

2.2. Biological factors affecting the specificity of a diagnostic test

Most diagnostic tests are not perfectly analytically specific. In serology, non-target components in serum samples such as antibodies elicited by other cross-reacting
pathogens or rheumatic factors may cause false-positive results. Likewise, false-positive results may be induced by non-specific factors in milk-based assays (e.g. positive beta-lactam antibiotic residue assays in untreated cows recovering from clinical mastitis, Cullor et al., 1994). A frequent observation in tropical veterinary medicine is the decrease of specificity when — after preliminary validation using non-exposed populations — the test is used in the target population (e.g. Greiner et al., 1997c; Mboloi et al., 1999). This finding is not surprising because of gross differences in the challenge with cross-reacting pathogens, animal breeds and management conditions. Maternal antibodies or antibodies that persist after self-cure or treatment may cause false-positive results in terms of diagnostic decisions (although the detection of antibodies in these cases would be considered true-positive in the analytical sense). Similarly, the detection of antibodies after vaccination might be considered false or true-positive depending on whether the objective of testing was to diagnose infection or to determine the immune state, respectively. The specificity of a test may vary among populations (or subpopulations) if the above-mentioned biological factors are differently distributed. Changes of the specificity may occur over time. For example, the removal of test-positive animals from the population during disease-eradication programs would lead to both a decrease in prevalence (with the known effect of a lower positive and a higher negative predictive value) and an increase in specificity. The two effects are attributable to the removal of animals with true-positive and false-positive test results, respectively. On the other hand, a decrease in specificity would be expected in a population that has experienced an epidemic (i.e. infectious agent is not longer present in animals yet antibodies persist).

3. Objectives, methods and limitations of different approaches for test validation

A systematic approach to test validation requires that the objectives of testing (including unit of analysis) and the objectives of the validation study are clearly stated, appropriate methods selected and potential constraints and biases addressed. Subtle but practically important differences exist with regard to the objectives of test validation (where the interest is focused either on the test per se or on its diagnostic performance for a given target population). The latter case can be extended by considering consequences of testing in a given situation. “Standardized” and “operational” test parameters and validation of “endpoints of testing” are terms related to the different objectives of test validation and are discussed in the following sections. “Operational” in this context means that the test parameters will be further used to update diagnostic decisions. The emphasis of our presentation is on the estimation of operational test parameters. Therefore, we discuss aspects of internal and external validity in relation to the estimation of operational test parameters. For the remainder of the paper, we assume that the individual animal is the unit of concern — but the principles described below can be generalized to aggregate testing. We do not expand on reliability assessment (e.g. inter-assay and inter-observer agreement) because this is more related to the technical aspects of the validation process.
3.1. Validation aiming at standardized test parameters

Test certification or general statements about the diagnostic quality of a test are not related to a specific testing situation. In this case, test parameters will only be used to describe a test under well-defined conditions and no attempt will be made to use accuracy measures as population parameters. Therefore, the objective of validation is to establish standardized diagnostic parameters that are unlinked with the uncertainty of selecting and sampling natural reference populations. This could be realized by using national or international serum panels for estimation of sensitivity and specificity. For the set-up of such serum panels, an agreement of the involved institutions must be reached on inclusion criteria, protocols for submission and distribution, legal matters, funding, reasonable coverage of geographical and biological factors, data-base management and biosafety. The sample size of the serum panels should be great enough to include a range of disease stages, titers, and other relevant conditions. The standardized sensitivity (specificity) could be defined as the proportion of sera from the sensitivity (specificity) panel that give positive (negative) reactions with the test and should be reported without confidence intervals because we are not concerned with the estimation of a population parameter. Serum panels from animals with defined infections may be also used to assess cross-reactivity of heterologous diagnostic serum assays. Testing should be conducted on coded samples (blinding) to avoid review bias. Cut-off-independent indices (e.g. based on ROC analysis; see companion paper by Greiner et al., 2000) should be established if quantitative tests are used.

3.2. Validation aiming at operational test parameters

This section considers test parameters that are applicable to animal populations, and therefore, approaches test validation from an epidemiologic perspective. The objective is to establish estimates of the sensitivity and specificity of a diagnostic test that can be used for conditioning the interpretation of test results on misclassifications. Requirements are that parameters should be estimated with a predefined degree of precision (i.e. random errors through the sampling process should not compromise the information content) and that estimates are internally valid (i.e. biases that compromise the validity of the parameter estimates for the given study conditions should be avoided). Also, if covariate information on biological factors is collected, more efficient (i.e. stratified) adjustments for misclassification are possible and this may facilitate the extrapolation to other populations. Practical, logistic, ethical and budgetary considerations will guide the investigator to select a cost-effective study design to collect relevant data.

3.2.1. Internal and external validity

An internally valid validation study produces sensitivity and specificity estimates that are unbiased for the population under consideration. Internal validity can be compromised by various biases — most of which occur in relation to the selected study design. Therefore, in the following sections we differentiate among three major design options: cross-sectional sampling with complete and partial verification and use of convenience samples of known disease status. In all cases, lack of blinding
could lead to overoptimistic sensitivity and specificity estimates if there is a subjective classification process for the test and/or reference status (review bias). Animals with unexpected results should not be re-tested unless this is required for technical reasons (e.g. in the framework of quality-control rules), because this could result in a related type of review bias.

Internally valid results are not necessary externally valid. General skepticism is appropriate against uncritically extrapolating test parameters from the validation study to a target population. However, in the absence of more appropriate data, extrapolations are often necessary and warranted if supported by the expertise of experienced diagnosticians. Tests should be re-validated after substantial changes in the target population (OIE, 1996). In any case, prevalence-dependent measures such as efficiency and predictive values would not apply in populations with other prevalences or even not after alteration of the prevalence in the same population (e.g. after successful disease control). Also, for quantitative tests the sensitivity and specificity depend on the selected cut-off value, and therefore, should be interpreted as a parameter pair by receiver-operating characteristic (ROC) analysis (see Greiner et al., 2000). Biases in individual-animal estimates of sensitivity and specificity can be greatly magnified when tests are used for herd-level interpretation (see Christensen and Gardner, 2000).

3.2.2. Cross-sectional sampling with complete verification

In a cross-sectional design, the target population is sampled using simple random, systematic, cluster or stratified sampling. All sample estimates directly or after necessary adjustments reflect the respective population parameters unless biases are introduced through the sampling process and/or analysis. The sample reflects the spectrum of disease and the distribution of other relevant conditions and biological factors in the population. Therefore, the approach is sometimes referred to as “naturalistic sampling” (Kraemer, 1992).

The sample should include animals from different clusters (herds, flocks, etc.). Covariate factors may be cluster-related and extrapolation to the population may fail when test parameters are estimated using observations from a single cluster. For example, in a study of the specificity of serologic tests for *Brucella melitensis* in goats, samples were collected from 10 herds with 20 adult females per herd. This approach was adopted to cover a range of husbandry practices that might have potentially affected specificity (Mikolon et al., 1998). The variation of sensitivity and specificity among clusters represents a form of parameter heterogeneity. Graphical (histograms, density trace) and numerical (mixture-distribution) methods were described to assess the heterogeneity of a binomial parameter (Böhning and Greiner, 1998) — in our case, sensitivity and specificity. Alternatively, the distribution of the continuous test variable (stratified for the true disease status of animals) can be investigated. To our knowledge, adjustments of variance estimates for intracluster correlation (e.g. McDermott and Schukken, 1994) have not been done for validation studies. For validation studies, a stratification of the sample could be useful to accommodate logistic (e.g. districts, municipalities) and analytic (e.g. biological factors) requirements. Even in a non-stratified sampling, a post hoc stratification for biological factors (covariates) is recommended for analysis. Stratum-specific estimates of sensitivity and specificity should be the preferred operational test
parameters. Extrapolation to other populations may be attempted using estimates weighted by the distribution of the covariates in the target population.

The diagnostic test and the reference diagnosis (gold standard) are obtained for all elements of the sample. The sequence of testing with the new and the reference test is not important as long as inclusion of individuals is unconditional on test or reference status (complete verification). In unblinded studies, positive test results might receive more extensive verification than negative test results — which would result in an overestimation of the sensitivity (work-up bias). The probability of inclusion of an animal in the sample should not deviate from the theoretical sampling probability as given by the sampling design. If any non-random process leads to exclusion of animals from the sample (i.e. test values or complete-animal records are missing) selection bias might result. For example, the sensitivity of an ELISA for Johne’s disease would be seriously overestimated if cattle with stage-I infection were underrepresented in the sample. This may occur because of the difficulty to identify stage-I cattle using the reference test (spectrum-of-disease bias).

The availability of a suitable reference method is an important requirement for cross-sectional validation studies. The reference status can be established using a single test or multiple tests. In the latter case, the results of the new test must not be part of the definition of the reference status, because this would result in an overestimation of the test’s accuracy (incorporation bias). Usually, one endeavours to use a reference method with sensitivity and specificity close to 100%, and consequently, has to accept that the procedure may be invasive, risky, costly or fails to provide results in a timely fashion. If the reference method has a low accuracy or if a non-appropriate method is used, the relative sensitivity and specificity of the new test will be biased (information bias). Methods for obtaining sensitivity and specificity estimates when there is no gold standard or when the reference test is imperfect are described in a companion paper (Enøe et al., 2000).

The results of a validation study are usually summarized in a 2×2 table. Consider for example a *Trypanosoma* antibody ELISA, which has been validated using a random subsample of n=183 of cattle sera collected during a cross-sectional study on trypanocide resistance in Mukono county, Uganda (details of the study are described by Greiner et al., 1997c). A *Trypanosoma* spp.-specific DNA-detection system served as the reference test (Clausen et al., 1998) (Table 1).

According to the formulae provided in Appendix A, we obtain the estimates (and 0.95 confidence intervals) for prevalence, $P=0.44$ (0.36, 0.51), apparent prevalence, $AP=0.48$ (0.40, 0.55), sensitivity, $Se=0.64$ (0.52, 0.74), specificity, $Sp=0.65$ (0.55, 0.74), efficiency, $Ef=0.64$ (0.57, 0.71), Youden’s index, $J=0.29$ (0.15, 0.43), likelihood ratio of a positive test, $LR+=1.82$ (1.34, 2.49), likelihood ratio of a negative test, $LR-=0.56$ (0.4, 0.77), odds ratio, $OR=3.27$ (1.78, 6.02), positive predictive value, $PPV=0.59$ (0.48, 0.69) and negative predictive value, $NPV=0.7$ (0.6, 0.79).

Logistic-regression analysis can be used for estimation of test-performance parameters as described by Coughlin et al. (1992) and Knottnerus (1992) (see Appendix A). Both methods yield point estimates of sensitivity and specificity that are independent of prevalence and numerically identical to the respective simple proportions of a 2×2 table. The advantage of logistic regression lies in the option to include covariate information
The analogue of stratified 2×2 tables in case of categorical covariates and random-effects terms as described by Lindberg et al. (1999). Analysis of Table 1 data by logistic regression yielded slightly different 95% confidence intervals for Se (0.56, 0.7), Sp (0.58, 0.72), LR⁺ (1.34, 2.48), and LR⁻ (0.41, 0.75).

Due to the cross-sectional design, the sample prevalence is an unbiased estimate of the true prevalence. Therefore, besides sensitivity and specificity, the predictive values, and the estimated true (given by the results of the reference method) and apparent (given by the results of the test) prevalences are meaningful and should be reported together with confidence intervals and sample sizes. We also recommend that the observed results are displayed in a 2×2 table. This is slightly redundant but readers might be interested to recalculate reported indices. This set of information suffices to derive all other measure of accuracy. All results should be grouped by (sub)population if information exists (see paratuberculosis example in Section 2.1).

3.2.3. Cross-sectional sampling with partial verification

When disease events are not prevalent in a population, an efficient sampling design might be to screen the population with the test and follow-up all or some of the test-positive and test-negative animals using the gold-standard procedure (partial verification). The approach could also be motivated by ethical objections against subjecting test-negative individuals to gold-standard testing if the latter is invasive or risky. Under budgetary constraints, one would consider the relative importance of estimating the positive or negative predictive value more precisely. Design-efficiency issues were discussed in detail by Irwig et al. (1994a,b). The data from such a study would be arranged as shown in Table 2.

Using the notation indicated in Table 2 we define the proportion of individuals that receive verification given a positive test result as \(c_1 = (a+b)/n_1\) and the corresponding proportion given a negative test result as \(c_2 = (c+d)/n_2\). If the sampling proportions are unequal \((c_1 \neq c_2)\), the naive estimates of sensitivity \((Se_{nv} = a/(a+c))\) and specificity \((Sp_{nv} = d/(b+d))\) are biased (verification bias), because numerator and denominator refer

<table>
<thead>
<tr>
<th>DNA detection</th>
<th>ELISA</th>
<th>T+</th>
<th>T-</th>
</tr>
</thead>
<tbody>
<tr>
<td>T+</td>
<td>51</td>
<td></td>
<td>29</td>
</tr>
<tr>
<td>T-</td>
<td>36</td>
<td>67</td>
<td></td>
</tr>
</tbody>
</table>

*Notation is \(T^+, T^-, D^+, D^-\) for ELISA positive, negative, gold-standard (DNA detection) positive, negative, respectively.
to different sampling proportions. Application-specific considerations might dictate that \( c_1 > c_2 \). Under the so-called “conditional-independence assumption”, it is required that the probabilities of selection for verification given a positive and given a negative test result are independent of disease state; i.e.

\[
\Pr(\text{verification}|T+, D+) = \Pr(\text{verification}|T+, D-)
\]

and

\[
\Pr(\text{verification}|T-, D+) = \Pr(\text{verification}|T-, D-)
\]

and can be estimated by \( c 1 \) and \( c_2 \) respectively. The assumption is valid if animals are selected for verification at random from the \( T+ \) and \( T- \) groups. In this case, the positive (PPV = \( a/(a+b) \)) and negative (NPV = \( d/(c+d) \)) predictive values are unbiased; bias-corrected estimates of sensitivity (\( \text{Se} \)) and specificity (\( \text{Sp} \)) are given as

\[
\text{Se}_{\text{corr}} = \frac{n_1 \text{PPV}}{n_1 \text{PPV} + n_2 (1 - \text{NPV})} = \frac{a/c_1}{a/c_1 + c/c_2}
\]

and

\[
\text{Sp}_{\text{corr}} = \frac{n_2 \text{NPV}}{n_2 \text{NPV} + n_1 (1 - \text{PPV})} = \frac{d/c_2}{d/c_2 + b/c_1}
\]

respectively (Begg and Greenes, 1983). Consider, for example, that in a validation study with partial verification, all 47 \( T+ \) animals (\( c_1=1 \)) but only 47 of 153 \( T- \) animals (\( c_2=0.307 \)) receive the gold-standard test and that the observed results were \( a=46, b=1, c=17, d=30, e=0 \) and \( f=106 \) (Table 2). Naive estimates were \( \text{Se}_{\text{nv}}=0.73, \text{Sp}_{\text{nv}}=0.97 \) and bias-corrected estimates under the conditional-independence assumption were \( \text{Se}_{\text{corr}}=0.45, \text{Sp}_{\text{corr}}=0.99 \). We estimate the variance of the corrected parameters as

\[
\text{var(Se}_{\text{corr}})=\text{Se}_{\text{corr}}(1-\text{Se}_{\text{corr}})/(a+c) \quad \text{and} \quad \text{var(Sp}_{\text{corr}})=\text{Sp}_{\text{corr}}(1-\text{Sp}_{\text{corr}})/(b+d)
\]

and obtain (using the normal approximation) (0.33, 0.58) and (0.95, 1) as 95% confidence intervals for \( \text{Se}_{\text{corr}} \) and \( \text{Sp}_{\text{corr}} \), respectively. Zhou (1994) established the variance of the predictive values in case of partial verification as

\[
\text{var(PPV)} = \text{PPV}(1-\text{PPV})/
\]

Table 2
Cross-tabulation of the results of a validation study with partial verification (numbers in parentheses are used for illustration of calculations in the text)<sup>a</sup>

<table>
<thead>
<tr>
<th>New test</th>
<th>Gold-standard test</th>
<th>unverified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D+</td>
<td>D-</td>
</tr>
<tr>
<td>T+</td>
<td>a (46)</td>
<td>b (1)</td>
</tr>
<tr>
<td>T-</td>
<td>c (17)</td>
<td>d (30)</td>
</tr>
<tr>
<td></td>
<td>n2</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Notation is \( T+, T-, D+, D- \) for test positive, negative, gold-standard positive, negative, respectively.
\((a+b)\) and \(\text{var}(\text{NPV}) = \text{NPV}(1-\text{NPV})/(c+d)\) and used maximum likelihood to estimate predictive values under departure from the conditional-independence assumption.

3.2.4. Convenience samples

In the context of validation studies, we suggest that the term “convenience samples” denotes all collections of specimens of known disease status that were not established to represent a population of animals. The word “convenience” does not imply that the process of collection was not systematic or that no valuable information about test properties could be obtained. Experimental studies, for example, have been suggested for test validation (OIE, 1996; Jacobson, 1998). Experimental infections allow substantial control over characteristics and management conditions of the host animal — as well as over identity, amount and administration route of infective material. The artificial infection represents the gold standard so that misclassifications are highly unlikely. Exact knowledge of the infection stage permits insight into changes in the sensitivity during the course of infection. Also, some diseases may only be studied experimentally. However, we urge caution in extrapolation of sensitivity and specificity estimates to field situations. Experimental conditions typically lead to an overestimation of sensitivity and specificity (for example, by artificially high challenge doses and by using specific-pathogen-free animals as negative controls, respectively). The latter have had limited opportunity for exposure to organisms that might produce cross-reacting antibodies. False-positive results that might seriously compromise a test in field situations might remain undiscovered. Study costs typically are high with long-term follow-up; hence, sample sizes generally will be small. Sensitivity changes over the course of infection and will be seriously overestimated when reported only for time points when response or analyte levels reach peak values. Follow-up samples from individuals are sometime pooled to get overall estimates of sensitivity or specificity. We strongly recommend against treating repeated samples as independent because this creates serious biases of both point and variance estimates of test parameters. The results of experimental infections should be presented in the form of reaction profiles of the test value for individual animals. For example, Gamble (1996) showed the \(\text{Trichinella}\) antibody ELISA profiles of pigs infected with different doses of \(\text{Trichinella spiralis}\) larvae. Such profiles provide important information on the kinetics of \(\text{Trichinella}\) antibody production — but are not useful to estimate the sensitivity of the ELISA. The investigators should provide their opinion about whether the reaction profile seen in experimental animals can be expected under field conditions.

Collections of specimens submitted to a diagnostic laboratory or teaching hospital represent another form of convenience sample. Even a systematic collection process cannot guarantee reproducible results, because the spectrum of submissions might vary over time in an unforeseeable manner. Results obtained from submission-based collections cannot easily be extrapolated to other situations. The influence of referral patterns on the characteristics of diagnostic tests has been studied by Knottnerus and Leffers (1992). The lack of representativeness does not apply to laboratories that are involved in large-scale testing in the context of prescribed test procedures. Also, serum banks — mainly used for retrospective studies of livestock diseases (e.g. Elbers, 1996) —
often reflect large parts of the population and could effectively be used for validation of tests. In this case, they would represent the population from which samples are drawn (as described in Sections 3.2.2 and 3.2.3).

3.2.5. Sample size

Use of the standard formula for estimating a proportion provides approximate sample sizes to estimate sensitivity and specificity with fixed error limits. Separate sample-size \((n)\) calculations are required for sensitivity and specificity,

\[
 n = \left( \frac{z_{1-\alpha/2}}{e} \right)^2 \theta(1 - \theta)
\]

where \(z_{1-\alpha/2}=1.96\), \(\alpha=0.05\), \(\theta\) is the a priori estimate of sensitivity or specificity and \(e\) is the desired error margin on the estimate. This sample-size calculation applies to both crude and stratum-specific estimates. Bigger sample sizes are usually necessary to assure that exact binomial confidence limits have the desired coverage (Fig. 1). Where no a priori estimate of \(\theta\) is available, a reasonable approach might be to assume values for \(\theta\) of say 0.8 — a value of 0.5 might be overly conservative and be sufficient to dissuade the implementation of a study. Even if the preliminary sensitivity estimates were 1, we would recommend using a more conservative estimate (e.g. 0.9) since many preliminary estimates are too optimistic. Choice of an appropriate error-margin value would depend on issues such as the likely economic or public-health importance of the test, likely extent of use, whether the test will be used in national eradication programs, etc. The more important the disease and the greater the consequences of misdiagnosis, the more important it becomes to have narrow error margins. The same sample-size calculations apply to all study designs that are conditional on disease status. Explicit sample-size requirements may also arise in the context of test comparison (see Greiner et al., 2000) and in prevalence estimation with adjustment for misclassifications (see Greiner and Gardner, 2000, this issue). Resampling techniques with replacement (bootstrap; see Efron and Tibshirani, 1993) can be applied to establish purely data-based confidence intervals for all test-performance parameters.

Fig. 1. Sample size for estimation of a proportion \((\theta)\) with error limits \(e=\pm 0.05\) under assumption of simple random sampling approximated using \(n=(1.96^2\theta(1-\theta))e^2\) (solid line) and under the requirement that the half widths of the exact binomial confidence interval are not greater than 0.05 (circles).
3.3. Validation of the respective endpoint of testing

Estimates of sensitivity and specificity obtained as described above are only intermediate parameters if the costs and benefits of diagnostic testing are of concern. Therefore, the objective of evaluating the desired endpoint in the decision-making process aims at the estimation of the net effect of testing. Direct (laboratory costs and sampling costs) as well as indirect costs and benefits associated with false-positive, false-negative, true-positive and true-negative test results are weighted by factors including prevalence and other utilities such as economy, policy, trust, welfare and public health. Methods of test validation that account for prevalence and misclassification have been developed in the context of ROC analysis (see Greiner et al., 2000). Testing strategies have been evaluated from a herd-level financial viewpoint by Mohammed et al. (1990)

Table 3
Checklist for validation of diagnostic tests

<table>
<thead>
<tr>
<th>General</th>
</tr>
</thead>
<tbody>
<tr>
<td>The test purpose and the analytical unit are described</td>
</tr>
<tr>
<td>The test protocol is sufficiently described</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference test (gold standard)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The choice of the reference method is justified (being more accurate than the new is a necessary condition); method is fully described or referenced</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Selection of reference populations</th>
</tr>
</thead>
<tbody>
<tr>
<td>The reference population is sufficiently described (time, location, animal characteristics such as breed, age, gender, etc.)</td>
</tr>
<tr>
<td>The reference population should reflect the target population and include an appropriate spectrum of disease and spectrum of other conditions</td>
</tr>
<tr>
<td>The sampling frame should be an unbiased representation of the reference population</td>
</tr>
<tr>
<td>Selection criteria must be stated and should reflect the testing situation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sampling of the reference population</th>
</tr>
</thead>
<tbody>
<tr>
<td>The sampling procedure is described in detail</td>
</tr>
<tr>
<td>Exclusion or inclusion criteria (if any) are described</td>
</tr>
<tr>
<td>Sample sizes must be stated and should reflect the degree of the required statistical certainty (usually not below 100)</td>
</tr>
<tr>
<td>Random and systematic sampling are the preferred options</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance of test and reference test</th>
</tr>
</thead>
<tbody>
<tr>
<td>The testing protocols are sufficiently described (including definition of negative and positive results)</td>
</tr>
<tr>
<td>Results of test and reference test are evaluated independently (blinded)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Presentation of results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter estimators are explained by formulae; estimates are presented together with sample sizes and confidence intervals (exact binomial confidence intervals are preferred); sensitivity and specificity are always required, additional parameters may be presented as necessary; the source 2×2 table should be displayed; ROC analysis should be presented for test outcomes measured on ordinal, interval or ratio scales</td>
</tr>
<tr>
<td>The number of uninterpretable and intermediate results (if any) and reasons for missing data are given</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discussion of results</th>
</tr>
</thead>
<tbody>
<tr>
<td>The test-performance parameters should be discussed in relation to the study design and the intended or current use of the test; if the gold standard is imperfect, this should be discussed in relation to the effect on the study results</td>
</tr>
</tbody>
</table>
and Collins and Morgan (1991); the topic is also related to decision analysis (see Smith and Slenning, 2000).

4. Checklist of epidemiologic considerations in test validation studies

Several guidelines exist in the medical literature that address test validation and assessing the quality of a validation study (e.g. Mulrow et al., 1989; Jaeschke et al., 1994). We present a checklist which is an excerpt of existing guidelines, supplemented with some of our own epidemiologic comments (Table 3). Explicit recommendations that suit all situations occurring in practice cannot be given and we present the checklist as a guide.

5. Conclusions

We have outlined issues that people involved in validation of animal tests should consider before implementing an evaluation study. We consider that many of these issues should form the basis for the methods’ sections of published scientific papers and we recommend their description in papers: proposed specific application, gold standards, sample selection and design, and bias-avoidance strategies. Results sections should include confidence intervals or point and variance estimates of test performance. The author(s) should provide enough detail so that readers can draw their own conclusions about the internal and external validity of the study. We suggest guidelines which are flexible and account for different conditions of test application. We note that development of standard recommendations for a single validation procedure that is suitable for all veterinary diagnostic tests is not possible because of the many different situations that are encountered.

Appendix A.

A.1. Glossary of terms and standard notation

The following terminology and notation closely follows accepted standards. However, synonyms other than those indicated below are in use and some writers use some of the terms differently. Therefore, we emphasize that it is good scientific practice to provide clear definitions of all terms when they appear the first time in a text. The “hat” on estimates is omitted throughout.

Diagnostic test. A single- or multiple-step device or process designed to detect a sign, substance, tissue change or response, and which either directly or after dichotomization yields a positive \((T^+)\) or negative \((T^-)\) result. The result is used for a diagnostic (dichotomous) decision between the positive \((D^+)\) and negative \((D^-)\) true state of disease, infection or immunity of an animal or a group of animals (referred to as “unit of analysis”).
Subsequent definitions refer to disease as the state of interest.

**Gold standard.** Classification of a unit of analysis as positive (\(D^+\)) or negative (\(D^-\)) for validation of a diagnostic test. The ideal gold standard would be a perfect (definitive) test that produces no misclassifications. Practically, a (reference) test of known and high accuracy, or a clinical or experimental classification is used.

**Cut-off value** (syn. threshold, critical value, decision limit). A value on the measurement scale of quantitative or ordinal diagnostic tests that classifies test results as positive (\(T^+\)) or negative (\(T^-\)).

**Decision fractions.** The observed frequencies of true-positive (TP; \(T^+|D^+\)), false-positive (FP; \(T^+|D^-\)), false-negative (FN; \(T^-|D^+\)) and true-negative (TN; \(T^-|D^-\)) test results, where TP and TN represent correctly classified and FP and FN represent misclassified units of analysis, respectively. Decision fractions are presented in a 2×2 table based on a total of \(N\) of analytical units.

<table>
<thead>
<tr>
<th>True disease state</th>
<th>(D^+)</th>
<th>(D^-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test result</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(T^+)</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>(T^-)</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

**Prevalence** (\(P\)). The probability that a randomly selected unit of analysis has the disease; \(Pr(D^+)\). In the context of diagnostic testing, \(P\) is sometimes referred to as pretest or prior probability. If the sampling for the above table is unbiased, the prevalence can be estimated as \((TP+FN)/N\). Because \(P\) is a simple proportion, its variance is given as \(P(1-P)/N\). An exact binomial confidence interval for \(P\) can be obtained from statistical software (e.g. Epi Info version 6.04B: Epitable, Describe, Proportion, Simple Random Sampling; Dean et al., 1994).

**Apparent prevalence** (AP). The probability that a randomly selected unit of analysis has a positive test result; \(Pr(T^+)\). If the sampling for the above table is unbiased, the apparent prevalence can be estimated as \((TP+FP)/N\). In this case, the variance of AP is \(AP(1-AP)/N\) and the exact confidence interval can be obtained as described for the prevalence. The apparent prevalence can be expressed using the prevalence (\(P\)) and sensitivity (Se) and specificity (Sp) as \(Se\ P^+(1-\ Sp)\ (1-P)\).

**Accuracy** (syn. validity). The ability of a diagnostic test to produce correct test results. Measures of diagnostic accuracy include sensitivity and specificity.

**Test validation** (syn. test evaluation). A study conducted with the objective of obtaining unbiased estimates of the diagnostic accuracy of a diagnostic test for a given target population.

**Sensitivity** (Se, syn. diagnostic sensitivity). The conditional probability that a unit of analysis has a positive test result, given the disease is present; \(Pr(T^+|D^+)\). Se can be estimated as \(TP/(TP+FN)\). In this case the variance of Se is \(Se\ (1-\ Se)/(TP+FN)\) and the exact confidence interval can be obtained as described for the prevalence. In
a cross-sectional sampling (whereby \( P \) is a random variable and a multinomial rather than a binomial sampling model is invoked; Abel, 1993, pp.70–72), \( \text{Se} \) and \( \text{Sp} \) are not stochastically independent — but often this is ignored. In a cluster-sampling design, an adjustment of the variance of a binomial parameter (here, \( \text{Se} \)) is often necessary. A simple approach is to use the herd-specific values \( \text{TP} \) and \( (\text{TP} + \text{FN}) \) for estimation of the “design effect” (i.e. the ratio of the design-based variance and the variance expected under simple random sampling) (Epi Info version 6.04b: Epitable, Describe, Proportion, Design effect; Dean et al., 1994). The design effect is then used to obtain an adjusted confidence interval for the global \( \text{Se} \) (Epitable, Describe, Proportion, Cluster sampling).

**Specificity** (\( \text{Sp} \), syn. *diagnostic specificity*). The conditional probability that a unit of analysis has a negative test result, given the disease is not present; \( \text{Pr}(T^-|D^-) \). \( \text{Sp} \) can be estimated as \( \text{TN}/(\text{TN} + \text{FP}) \). In this case, the variance of \( \text{Sp} \) is \( \text{Sp}(1-\text{Sp})/(\text{TN} + \text{FP}) \). An exact confidence interval (simple random-sampling assumption) and an adjusted confidence interval for cluster sampling can be obtained as described above.

**Efficiency** (\( \text{Ef} \)). The probability of correct classification given the prevalence \( P \). Estimated as the overall proportion of agreement (\( \text{TP} + \text{TN} \))/\( N \). The variance of \( \text{Ef} \) is \( \text{Ef}(1-\text{Ef})/N \) and the exact confidence interval can be obtained as described for the prevalence. \( \text{Ef} \) is prevalence-dependent because it is equivalent to \( \text{Se} P + (\text{Sp}) (1-P) \).

**Youden's index** (\( J \)). A summary measure of sensitivity and specificity invariant to prevalence; equals \( \text{Se} + \text{Sp} - 1 \). If the \( \text{Se} \) and \( \text{Sp} \) estimates are stochastically independent (as it is in the case of prestratified sampling where \( P \) is a fixed value), the variance of \( J \) is the sum of the variances of \( \text{Se} \) and \( \text{Sp} \). A large-sample \( (1-\alpha) \) confidence interval can be constructed using the normal approximation.

**Likelihood ratio of a positive test result** (\( \text{LR}^+ \)). A prevalence-independent, combined measure of sensitivity and specificity that represents the link between the odds of the pretest and post-test probability of disease, given a positive test result. \( \text{Odds} = \text{probability}/(1-\text{probability}). \)

\[
\text{post-test odds (given } T^+ \text{)} = \text{LR}^+ \times \text{pretest odds} \\
\frac{\text{Pr}(D^+|T^+)}{\text{Pr}(D^-|T^+)} = \text{LR}^+ \times \frac{\text{Pr}(D^+)}{\text{Pr}(D^-)} \\
\frac{\text{PPV}}{1-\text{PPV}} = \text{LR}^+ \times \frac{P}{1-P}
\]

where \( \text{PPV} \) denotes the positive predictive value (see below). Using algebra, one can solve for \( \text{LR}^+ \), which is given as \( \text{Se}/(1-\text{Sp}) \). If, more generally, we express the likelihood ratio \( \text{LR}^+ = p_1/p_2 \), the variance of log (\( \text{LR} \)) can be estimated as \( \text{var}(\text{LR}) = (1-p_1)/a + (1-p_2)/b \), where \( a \) and \( b \) are the denominators of \( p_1 \) and \( p_2 \), respectively. A \((1-\alpha)\% \) confidence interval can be obtained as \( \text{LR} \times \exp\left(\pm z_{1-\alpha/2} \sqrt{\text{var}(\text{LR})}\right) \) (see Centor, 1992; a spreadsheet template for calculation of confidence intervals for \( \text{LR} \) can be obtained from R.M. Hamm, University of Oklahoma Health Services Center, www.fammed.ouhs-c.edu/robham/cdmcalc.htm). Likelihood ratios for continuous test results are described elsewhere (Greiner et al., 2000).
**Likelihood ratio of a negative test result** (LR−). A prevalence-independent, combined measure of sensitivity and specificity that represents the link between the odds of the pretest and post-test probability of disease, given a negative test result.

\[
\text{post-test odds (given } T^-) = \text{LR} - \times \text{pretest odds}
\]

\[
\frac{\Pr(D+ | T^-)}{\Pr(D- | T^-)} = \text{LR} - \times \frac{\Pr(D+)}{\Pr(D-)}
\]

\[
\frac{1 - \text{NPV}}{\text{NPV}} = \text{LR} - \times \frac{P}{1 - P}
\]

where NPV denotes the negative predictive value. Algebraic manipulation results in \(\text{LR} = (1 - \text{Se})/\text{Sp}\). Refer to LR+ for variance and confidence intervals.

**Odds ratio** (OR). The cross-product ratio \((TP \times TN)/(FP \times FN)\), which is equivalent to \(\text{LR}+ / \text{LR}−\) and to \(\text{antilog (logit (Se) + logit (Sp))}\). An estimate of the variance of \(\log (\text{OR}) \) is \(\text{var(OR)} = 1/\text{TP} + 1/\text{FP} + 1/\text{FN} + 1/\text{TN}\), which can be used to construct a \((1-\alpha)\)-confidence interval around \(\text{OR}\) as \(\text{OR} \times \exp \left( \pm \frac{z_{1-\alpha/2}}{\text{var(OR)}} \right)\) (Woolf’s method). Exact confidence intervals for OR can be obtained from statistical software (e.g. Epi Info version 6.04B: Epitable, Study, Case-Control, Unmatched; Dean et al., 1994).

**Positive predictive value** (PPV). The conditional (post-test) probability of disease, given a positive test result; \(\Pr(D+ | T+)\). If the \(2 \times 2\) table reflects an unbiased sample of the target population in terms of disease prevalence, PPV can be estimated as \(\text{TP}/(\text{TP} + \text{FP})\). Otherwise, PPV is established using

\[
\frac{P \times \text{Se}}{P \times \text{Se} + (1 - P)(1 - \text{Sp})}
\]

**Negative predictive value** (NPV). The conditional (post-test) probability of “no disease”, given a negative test result; \(\Pr(D- | T-)\). If the \(2 \times 2\) table reflects an unbiased sample of the target population in terms of disease prevalence, NPV can be estimated as \(\text{TN}/(\text{TN} + \text{FN})\). Otherwise, NPV is established using

\[
\frac{(1 - P) \times \text{Sp}}{(1 - P) \times \text{Sp} + P(1 - \text{Se})}
\]

**Precision.** The ability of a diagnostic test to produce consistent results within tolerable analytical-error limits when a specimen is re-tested. According to the re-testing condition, one differentiates between repeatability (intra-assay, run-to-run, day-to-day, operator-to-operator precision) and reproducibility (laboratory-to-laboratory precision).

**Logistic regression.** A mathematical model to describe the relationship between a dichotomous response variable and a single or a set of explanatory variables. In the context of validation studies, we can define the variable \(X\) so that \(X=0\) and \(X=1\) denote negative and positive test results, respectively, and the variable \(Y\) so that \(Y=0\) and \(Y=1\) denote negative and positive true disease status, respectively. The logistic-regression model

\[
\text{logit} \Pr(X=1 | Y) = a + bY + \varepsilon
\]
where \(a\) and \(b\) are estimated coefficients and \(\varepsilon\) is the variation not explained by the model, describes the probability of positive test results conditional on disease state. This is model 1. The coefficients \(a=\logit(1-\text{Sp})\) and \(b=\logit(\text{Se})-\logit(1-\text{Sp})=\ln(\text{OR})\) are both prevalence-independent and can be used for estimation of \(\text{Se}=\exp(a+b)/(\exp(a+b)+1)\) and \(\text{Sp}=1/(\exp(a)+1)\) (Coughlin et al., 1992). The model

\[
\logit \Pr(Y=1|X) = a + bX + \varepsilon
\]

(2)
describes the logit of the predictive value of \(X\). This is model 2. The coefficients \(a\) (which depends on the sample prevalence \(P_0\)) and \(b\) (which is not dependent on \(P_0\)) have the interpretation of \(a=\logit(1-\text{NPV})\) and \(b=\ln(\text{OR})\), respectively. Using the estimated coefficients, we can derive \(\text{PPV}=\exp(a+b)/(\exp(a+b)+1)\), \(\text{NPV}=1/(\exp(a)+1)\). \(\text{Se, Sp, LR+}\), and \(\text{LR}–\) can be derived from model 2 but this has no advantage. If we regard the independent variable \(X\) in model 2 as latent continuous, we can find a value \(x_0\) which does not change the prior (sample) prevalence \(P_0\). In this case, \(\logit \Pr(Y=1|x_0)=a + bx_0=\logit P'\). We solve for \(x_0=(\logit P' - a)/b = -\ln(\text{LR}–)/\ln(\text{OR})\) and subtract this constant from \(X\) in the model

\[
\logit \Pr(Y=1|X) = a + b(X - x_0) + \varepsilon.
\]

(3)

This is model 3. The coefficients have the interpretation \(a=\logit(P')\) and \(b=\ln(\text{OR})\). From model 3, we can derive \(\text{LR+}=\exp(b(1-x'))\), \(\text{LR}=\exp(b(-x'))\), \(\text{Se}=\exp(-bx')/(\exp(-b)-1)\) and \(\text{Sp}=\exp(-b(1-x'))/(\exp(-b)-1)\) (Knottnerus, 1992). Using the standard error (S.E.) of \(b\), we can construct approximate \((1-\alpha)\)-confidence intervals for \(\text{LR+}, \text{LR}, \text{Se}\) and \(\text{Sp}\) by replacing \(b\) with \(b\pm z_{1-\alpha/2} S.E.\) (\(b\)). Note that in Knottnerus’ equations (10) and (27), the term \((x'-1)\) is to be replaced by \((1-x')\). The logistic-regression models could be extended by inclusion of covariate factors as explanatory variables into the linear part.

**Biological factors (syn. influential covariate factors).** In the context of test validation, acquired or inherited, stable or variable traits of animals that affect sensitivity or specificity of a diagnostic test through alteration of physiological or pathophysiological pathways (e.g. analyte concentration or reactivity) (see Greiner et al., 1997a).

**References**


