Diagnostic likelihood ratio – the next generation of diagnostic test accuracy measurement

C.G.B. Caraguel [1, 2]* & A. Colling [2, 3]

(1) School of Animal and Veterinary Sciences, The University of Adelaide, Roseworthy Campus, Roseworthy, South Australia, 5371, Australia
(2) OIE Collaborating Centre for Diagnostic Test Validation Science in the Asia-Pacific Region, CSIRO, 5 Portarlington Road, East Geelong, Victoria, 3219, Australia
(3) Australian Centre for Disease Preparedness, Commonwealth Scientific and Industrial Research Organisation (CSIRO), 5 Portarlington Road, East Geelong, Victoria, 3219, Australia
*Corresponding author: charles.caraguel@adelaide.edu.au

Summary
To select, interpret and assess the fitness-for-purpose of diagnostic tests, we need to compare the likelihoods of test results being true versus false across both infected and non-infected individuals. Diagnostic sensitivity (DSe) and specificity (DSp) report the accuracy of classification in infected and non-infected individuals separately and do not compare these likelihoods directly. Positive and negative predictive values combine these likelihoods, but they also heavily depend on the prevalence in the tested populations and, therefore, cannot be generalised. The authors propose the adoption of the diagnostic likelihood ratio (LR), which balances the likelihoods of true versus false results and is population independent. As a relative measure, LR ignores the absolute accuracy of tests, and two tests with different accuracy profiles may have the same LR. This can be easily mitigated by using listed complementary measures of accuracy, including DSe and DSp, or ancillary selection criteria. Over all, LR is a more relevant and universal measure of diagnostic test accuracy, which makes it the logical next-generation measure to adopt. The authors illustrate the applications and benefits of LR using three assays certified by the World Organisation for Animal Health as serological tests for bovine tuberculosis.

Keywords
Accuracy – Diagnostic sensitivity – Diagnostic specificity – Diagnostic test – Likelihood ratio.

Introduction

Understanding and appreciation of the accuracy of a test is necessary when designing and evaluating a detection algorithm to control and prevent disease no matter the level – individual animal or population. The *analytical* accuracy refers to the ability of a test to accurately measure the true status of a *specimen* submitted to the test (laboratory phase), while the *diagnostic* accuracy refers to the ability of a test to accurately measure the true status of the *sampled individual*. Diagnostic accuracy accounts for potential error(s) during specimen collection, transport and storage (field phase), as well as during the laboratory process and in the output reading. The diagnostic accuracy of tests is conventionally reported using diagnostic sensitivity (DSe) and diagnostic specificity (DSp) (1). It is difficult to determine when DSe and DSp were first used to report test accuracy, but there are reports of their use at least since the mid-1940s (see [2] as an example). Diagnostic sensitivity is the probability of a positive test result among infected individuals (i.e. test accuracy in infected) while DSp is the probability of a negative test result among non-infected individuals (i.e. test accuracy in non-infected) (3). These measures require the output of the tests to be binary – either positive or negative – otherwise the test output must first be dichotomised. For ordinal scale outputs this can be done by merging measurement categories (e.g. virus neutralisation titres), and for continuous scale outputs it can be done by setting a cut-off or breakpoint threshold (e.g. optical

https://doi.org/10.20506/rst.40.1.3226

density values for serum enzyme-linked immunosorbent assay or ELISA) (4).

The need for an alternative measure of diagnostic test accuracy

Diagnostic accuracy is typically reported using two separate measures, DSe and DSp, because a test is unlikely to be equally accurate in infected and non-infected individuals. A single measure of diagnostic accuracy would be the average of both DSe and DSp, weighted by the proportion of infected individuals (i.e. prevalence) in the study population and would, therefore, be population dependent. Because they report the accuracy of a test separately in infected and non-infected individuals, DSe and DSp are, in theory, prevalence and, therefore, population independent.

The DSe and DSp have a range of applications (5). They are used to design and assess detection algorithms at the individual level (accuracy estimation of multiple tests in combination) or at the population level (sample size calculations for surveys, 'herd' sensitivity and specificity estimation). They can also be combined with infection prevalence to facilitate interpretation of test results using Bayes' theorem (i.e. positive and negative predictive values at the individual or population level) or to adjust for misclassification in prevalence studies (i.e. adjust the true prevalence from the apparent prevalence). Finally, DSe and DSp can be used directly to assess the fitness of a particular test for one or several purposes (e.g. World Organisation for Animal Health [OIE] validation process) or to select, among several, the most suitable test for a given purpose. For instance, at the individual animal level, one would select the test with the highest DSp to minimise false positives and be more confident in a positive test result (i.e. ruling in infection) or, conversely, select the test with the highest DSe to minimise false negatives and be more confident in a negative test result (i.e. ruling out infection).

Although tempting to consider separately, DSe and DSp should always be used together when selecting a test or assessing its fitness for purpose. For example, in the absurd instance where a test always provided positive results regardless of the true status of the individual, the test's DSe would be perfect (i.e. 100%); however, its DSp would be zero (i.e. 0%). A test that never returned a negative result would mean infection could never be ruled out and it would, therefore, be of little use. Intuitively, DSp needs to be considered as well. This is because a negative test result can either be false, which depends on DSe (infected individuals falsely classified as negative), or be true, which depends on DSp (non-infected individuals truly classified as negative). Therefore, both DSe and DSp should be considered when selecting a test to rule out infection. This is particularly relevant in the likely scenario where one test has a higher DSe but lower DSp than another. Reciprocal reasoning applies when selecting test(s) for ruling in infection. Therefore, we need a measure of accuracy that can quantify the diagnostic value of a test result by weighting both DSe and DSp.

One such measure is the *predictive value,* which reports the probability of a test result, either positive or negative, being true (1). However, this measure highly depends on the infection prevalence, as well as DSe and DSp, and varies accordingly with the population of interest. Therefore, predictive values are context-dependent and should never be reported and applied universally. Instead, we need a measure that weights both DSe and DSp and is population independent.

The **diagnostic likelihood ratio** (LR) is a measure that fits both of these requirements and that has been used by clinicians, primarily in human medicine, since the 1980s (6, 7). Within the clinical context, more than one test may be used, in addition to history and physical examination, to reach a final diagnosis. Test LRs are then used to assess the added diagnostic value of a given test result and refine the clinical diagnosis (8). The authors propose here to broaden the use of LR to the selection of diagnostic tests and the assessment of their fitness for purpose. After defining LR, the authors will outline its applications with examples that will highlight its strengths and limitations.

Diagnostic likelihood ratio – definition and calculation

Definition

The LR indicates the *diagnostic power* of a given test result. The diagnostic power combines the direction and the strength of the diagnostic evidence provided by a particular result (9).

Calculation

The likelihood ratio is calculated as the ratio of the likelihood of a given test result occurring in infected individuals to the likelihood of the same test result occurring in non-infected individuals (3):

$$LR_{\text{test result} = X} = \frac{\text{likelihood of test result X among } \mathbf{infected}}{\text{likelihood of test result X among } \mathbf{non\text{-}infected}} \quad \text{(Equation 1)}$$

There is a distinct LR for each output value of the test (11). A test with a binary output would have an LR for the positive test result (LR^+) and one for the negative test result (LR^-). LR^+ or LR^- can be directly calculated from DSe and DSp as follows (10):

$$LR^+ = \frac{DSe}{1 - DSp} \quad \text{(Equation 2)}$$

$$LR^- = \frac{1 - DSe}{DSp} \quad \text{(Equation 3)}$$

A test with a categorical scale output will have a different LR for each of the categories ($LR_{category}$). Tests with a continuous output have, in theory, an LR for each value within their operating range (LR_{value}). However, because of often limited sample size, this is impractical and tests with continuous output are instead categorised using one or more cut-off values into two (binary) or more (ordinal) range categories with one $LR_{category}$ each. When only a single cut-off is used, we obtain the default two result categories (positive and negative), each with their corresponding LR^+ or LR^-.

The authors use here, as an example, the OIE-certified Enferplex bovine tuberculosis (TB) antibody test, an indirect multiplex ELISA designed to detect antibody activity against up to 25 *Mycobacterium bovis* antigens in cattle sera and evaluated by Whelan *et al.* (12). The raw output of this assay is a measure of the intensity of the chemiluminescent signal produced by the labelled conjugate for each individual antigen and measured in relative light units (RLUs) using a digital imaging system. For each antigen, this measurement is dichotomised into positive or negative using an antigen-specific cut-off value set by the manufacturer. To be classified as positive, a serum sample needs to be positive to at least two antigens. In their study, Whelan *et al.* (12) investigated five different sets of RLU cut-offs (referred to as 'Levels'), resulting in distinct sets of DSe and DSp estimates. From the study results, the authors adapted the data for the Level-1 cut-off into a 2 × 2 contingency table and have provided an example of LR^+ and LR^- calculations in Table I. Similarly, the authors calculated separate sets of LR^+/LR^- for each of the other four presented cut-offs (Table II).

In the absence of a perfect reference test, the assessment of the accuracy of tests under evaluation is conducted using latent class analysis (LCA), using either a deterministic or a Bayesian approach (4). An example of the Bayesian version of LCA applied to TB serology evaluation is reported in Clegg *et al.* (13). In the Bayesian implementation of the LCA, it is also possible to estimate LR^+ and LR^- from the DSe and DSp posterior samples by adding into the LCA model the LR calculation steps corresponding to Eqs 2 and 3.

To also illustrate LR calculations for tests with categorical outputs, the authors reorganised the data from Whelan *et al.* (12), using the five RLU cut-off sets as the boundaries of six RLU category ranges (Table III). A separate $LR_{category}$ is calculated for each range by dividing the fraction of infected individuals by the fraction of those non-infected for results within the range of interest. Category-specific estimates of DSe and DSp may also be calculated ($DSe_{category}$ and $DSp_{category}$). As demonstrated in Table III, the $DSe_{category}$ corresponds to the fraction of infected individuals with results within the category of interest, and $DSp_{category}$ to the complement of the fraction of non-infected individuals with results within the category of interest.

Table I

Diagnostic accuracy estimates with 95% confidence intervals in parentheses (bold) calculated from a 2 × 2 contingency table using data adapted from Whelan *et al.* and corresponding to the 'Level 1' cut-off of the Enferplex bovine tuberculosis antibody test (12)

A total of 96 known infected and 93 known non-infected cattle with bovine tuberculosis (TB) were used in the study

		True tuberculosis status	
		Known infected (*n* = 96)	**Known non-infected (*n* = 93)**
Test results	Positive	83 TP	19 FP
	Negative	13 FN	74 TN
		Diagnostic sensitivity (DSe) TP/(TP + FN) **86.5% (78.0% – 92.6%)[a]**	Diagnostic specificity (DSp) TN/(TN + FP) **79.6% (69.9% – 87.2%)[a]**
		Diagnostic likelihood ratio for a positive test result (LR^+) [TP/(TP + FN)]/[FP/(TN + FP)] **4.23 (2.81 – 6.37)**	Diagnostic likelihood ratio for a negative test result (LR^-) [FN/(TP + FN)]/[TN/(TN + FP)] **0.170 (0.102 – 0.285)**

(a) Exact binomial 95% confidence interval

FN: false negative
FP: false positive
TN: true negative
TP: true positive

Confidence intervals (CIs) for LR are calculated as for other ratios of two independent proportions (e.g. risk ratio). The log method outlined by Altman (10) is the most commonly used approach and follows a three-step process:

Step 1 – Calculate the standard error (*SE*) of the natural log of the LR for the test result *X* of interest (LR_X):

$$SE_{\ln LRX} = \sqrt{\frac{1}{count\ of\ infected\ with\ result\ X} - \frac{1}{count\ of\ infected} + \frac{1}{count\ of\ non\text{-}infected\ with\ result\ X} - \frac{1}{count\ of\ non\text{-}infected}} \quad \text{(Equation 4)}$$

Step 2 – Calculate the boundaries of the 95% CI of the natural log of LR_X:

$$95\%\ CI\ {}^{lower\ limit}_{\ln LR_X} = \ln(LR_X) - 1.96 \times SE_{\ln LR_X} \quad \text{(Equation 5)}$$

$$95\%\ CI\ {}^{upper\ limit}_{\ln LR_X} = \ln(LR_X) + 1.96 \times SE_{\ln LR_X} \quad \text{(Equation 6)}$$

Step 3 – Back-transform the boundaries of the 95% CI of the natural log of LR_X:

$$95\%\ CI\ {}^{lower\ limit}_{LR_X} = e^{95\%\ CI\ {}^{lower\ limit}_{\ln LR_X}} \quad \text{(Equation 7)}$$

$$95\%\ CI\ {}^{upper\ limit}_{LR_X} = e^{95\%\ CI\ {}^{upper\ limit}_{\ln LR_X}} \quad \text{(Equation 8)}$$

Using the example of the LR+ estimated at 4.23 in Table I, the $SE_{\ln LR+i}$ is calculated as 0.2086

$$\text{(i.e. } \sqrt{\frac{1}{TP} - \frac{1}{TP + FN} + \frac{1}{FP} - \frac{1}{FP + TN}} =$$

$$\sqrt{\frac{1}{83} - \frac{1}{83 + 13} + \frac{1}{19} - \frac{1}{19 + 74}} \text{)}$$

and the boundaries of the 95% CI on the natural log scale as 1.03 and 1.85, respectively. After back-transformation, the boundaries of the 95% CI of LR+ become 2.81 and 6.37 (Table I).

Interpretation

The LR ranges from zero to infinity. If the LR of a given test result is greater than one, this test result supports the presence of the infection, i.e. this test result is more likely to occur in infected animals than in non-infected animals. Conversely, if the LR is lower than one, the test result supports the absence of the infection, i.e. this test result is less likely to occur in infected animals than in non-infected animals. An LR equal to '1' means that the test result has no diagnostic power (i.e. it is as likely to occur in infected animals as it is in non-infected animals). The further the LR is away from one, towards either zero or infinity, the stronger the evidence provided by the test result. In the clinical context, test outputs with an LR > 10 or < 0.1 are considered good diagnostic evidence of the infection being either present or absent, respectively (14).

In the case of the Enferplex assay for TB serology (Table I), the estimated LR+ for Level-1 cut-off was 4.2, which means that a seropositive result using this cut-off supports the presence of TB (LR > 1) and this diagnostic evidence should be considered as moderate (i.e. approximately four times

Table II

Estimates, with 95% confidence intervals (CI) in parentheses, of the cut-off specific diagnostic sensitivity ($DSe_{cut\text{-}off}$), cut-off specific diagnostic specificity ($DSp_{cut\text{-}off}$), likelihood ratio of a positive test result ($LR_{cut\text{-}off}^+$) and likelihood ratio of a negative test result ($LR_{cut\text{-}off}^-$) for five sets of cut-off thresholds used by Whelan *et al.* when evaluating the Enferplex bovine tuberculosis antibody test (12)

Cut-off	TP	FP	FN	TN	$DSe_{cut\text{-}off}$ (95% CI) [a]	$DSp_{cut\text{-}off}$ (95% CI) [a]	$LR_{cut\text{-}off}^+$ (95% CI)	$LR_{cut\text{-}off}^-$ (95% CI)
Level 1	83	19	13	74	86.5% (78.0%, 92.6%)	79.6% (69.9%, 87.2%)	4.2 (2.8, 6.4)	0.170 (0.102,0.285)
Level 2	79	8	17	85	82.3% (73.2%, 89.3%)	91.4% (83.8%, 96.2%)	9.6 (4.9, 18.7)	0.194 (0.125, 0.300)
Level 3	78	5	18	88	81.3% (72.0%, 88.5%)	94.6% (87.9%, 98.2%)	15.1 (6.4, 35.6)	0.198 (0.130, 0.301)
Level 4	76	2	20	91	79.2% (69.7%, 86.8%)	97.8% (92.4%, 99.7%)	36.8 (9.3, 146.0)	0.213 (0.144, 0.315)
Level 5	74	0	22	93	77.1% (67.4%, 85.0%)	100.0% (96.1%, 100.0%)	Infinity [b] (9.05, infinity)	0.229 (0.159, 0.331)

(a) Exact binomial 95% CI
(b) Denominator was zero, which can be interpreted as an infinitely large ratio

TP: true-positive count
FP: false-positive count
FN: false-negative count
TN: true-negative count

Table III
Estimates, with 95% confidence intervals (CI) in parentheses, of the diagnostic likelihood ratio ($LR_{category}$) of six range categories of relative light units using the data reported by Whelan *et al.* (12) These authors evaluating the Enferplex bovine tuberculosis (TB) antibody test. The $LR_{category}$ was calculated for each category range by dividing the fraction of TB known infected (equivalent to the category-specific diagnostic sensitivity, $DSe_{category}$) by the fraction of TB known non-infected (equivalent to the complement of the category-specific diagnostic specificity, $1\text{-}DSp_{category}$) with results within the range of interest

Category range	Tuberculosis known		$LR_{category}$ (95%CI)
	Infected (%)	Non-infected (%)	
< Level 1	13 (13.5%)	74 (79.6%)	0.170 (0.102, 0.285)
Level 1 to < Level 2	4 (4.2%)	11 (11.8%)	0.352 (0.116, 1.07)
Level 2 to < Level 3	1 (1.0%)	3 (3.2%)	0.323 (0.034, 3.05)
Level 3 to < Level 4	2 (2.1%)	3 (3.2%)	0.646 (0.11, 3.78)
Level 4 to < Level 5	2 (2.1%)	2 (2.2%)	0.969 (0.139, 6.74)
≥ Level 5	74 (77.1%)	0 (0.0%)	Infinity [(a)] (9.05, infinity)
Total	**96 (100.0%)**	**93 (100.0%)**	

(a) Denominator was zero, which can be interpreted as an infinitely large ratio

more likely to occur in seroconverted cattle than in non-seroconverted cattle). On the other hand, the estimated LR^- was 0.17, which suggests that a seronegative result based on this cut-off supports the absence of TB (LR < 1) and this diagnostic evidence is also moderate (i.e. approximately $1/0.17 \approx 6$ times more likely to occur in non-infected cattle than in infected cattle).

Particular attention should be given to the 95% confidence intervals. In Table III, all the intervals for the $LR_{category}$ ranging from 'Level 1 to < Level 2' to 'Level 4 to < Level 5' are relatively wide and include the value '1'. Wide intervals are primarily attributable to sparse data in these intermediate categories, with only between 4 and 15 observations each (Table III). The inclusion of '1' in the intervals indicates that the $LR_{category}$ for these categories is not significantly different from '1'. Therefore, the diagnostic value of these categories is worthless, as the likelihood of any of these categories occurring in infected animals is not different from the likelihood of them occurring in non-infected animals. In this example, only the categories '< Level 1' and '≥ Level 5' provide valuable diagnostic evidence ($LR_{category} \neq 1$), albeit in opposite directions.

Diagnostic likelihood ratio – applications

Assessment of a test's fitness for purpose

Tests submitted for inclusion in the OIE Register of test kits are validated by assessing if their accuracy fits their intended purpose(s). We saw earlier that using either DSe or DSp in isolation is a single-sided perspective and can lead to erroneous conclusions. Instead, the authors recommend considering using LR. To illustrate this, the authors compared the reported DSe and DSp estimates from the three OIE-certified serological assays for TB in cattle (15) and calculated the LR^+ and LR^- for each assay from the DSe and DSp estimates using Eq. 2 and Eq. 3, respectively (Table IV). Note that the estimates of DSe and DSp for the Enferflex assay in Table IV differ from Whelan *et al.* (12), reported in Tables I and II, because the data used for the OIE certification of this test originated from another study with a different set of data and cut-offs. It is also worth noting that the three assays were evaluated independently using separate animal populations at different times and locations, and that the observed variations in accuracy estimates could reflect differences in population characteristics. Ideally, the three assays should be evaluated and compared using the same animals to exclude a potential population effect.

First, the authors used the IDEXX *M. bovis* antibody test kit as an example. The stated intended purpose of this test is to act 'as a supplemental test, in conjunction with other methods, for diagnosing and managing *M. bovis* infection' (16). In other words, the assay is intended to confirm prior positive finding(s) and rule in TB infection. In this context, we would conventionally look for a high DSp to be highly confident in a positive test result and would pay less attention to the DSe. The DSp estimate of the IDEXX assay was estimated at 98% (Table IV), which suggests that the likelihood of false positives is low (i.e. 2%); however, the DSe was estimated at 64.6%, which means that the likelihood of true positives is, at best, moderate. So, the 'LR' true-to-false positive occurrence and, therefore, our confidence in a positive test result, may not be as high as first expected when we only considered DSp. The LR^+ quantifies this ratio as 32.3 (Table IV). From this estimate, we can confirm that a positive result from the IDEXX serology supports the presence of TB infection (LR > 1) and that this evidence is strong (a true positive is approximately 32 times more likely to occur than a false negative). Thus, the LR^+ measures the **confirmatory power** of a test and should be expected to be very large for any test aspiring to rule in infection. The LR^- , on the other hand, measures the **excluding power** of a test and should be expected to be very small (close to zero) for any test aspiring to rule out infection. The LR^- of the IDEXX serology assay was estimated at 0.361, which suggests that a seronegative result indeed

supports the absence of TB (LR < 1), but this evidence is, at best, moderate (a true negative is approximately 1/0.361 ≈ 3 times more likely to occur than a false negative).

Test selection for a given purpose

The likehood ratio is also advantageous over DSe and DSp when selecting a test for a given purpose. If one had to choose between IDEXX and BOVIGAM TB blood-based tests to confirm (rule in) a suspect case, the test with the highest DSp, in this case IDEXX, would normally be favoured (Table IV). However, when comparing the LR^+, which considers both DSp and DSe, BOVIGAM would become comparable, with a similar LR^+ (32.5 vs 32.3). This illustrates the fact that the evidence provided by a positive result from a test with inferior DSp can be even more valuable than that provided by a positive result from a test with higher DSp, if the DSe of the former is far higher than that of the latter. Therefore, selecting a test for confirmation solely using DSp could be an error. However, in this instance, the LR^+ of the IDEXX and BOVIGAM assays are not practically or statistically different (i.e. overlapping CIs), so which one is preferred? When the LR^+ among tests is similar, the authors recommend selecting the test with the smallest LR^-. This provides the best balance between LR^+ and LR^-, which is quantifiable by calculating the diagnostic odds ratio (DOR) as follows (19):

$$DOR = \frac{LR^+}{LR^-} \quad \text{(Equation 9)}$$

According to the DOR reported in Table IV, it appears that the BOVIGAM assay would be a better candidate than the IDEXX for TB blood testing in cattle (17, 18). This pairwise comparison ignores the Enferplex assay, whose LR^+ surpasses both other tests regardless of the cut-off and would be the favoured single test for ruling in TB (Table IV).

For ruling out (screening) TB, BOVIGAM and Enferplex-sensitive cut-off would be the tests of choice because they have the lowest LR^- (0.158 and 0.155, respectively, Table IV). If one had to choose between those two, Enferplex-sensitive cut-off might be preferred because it has the highest DOR (Table IV). However, the diagnostic evidence provided by a seronegative from any of those tests is moderate (1/0.16 ≈ 6 times more likely to occur in a non-infected animal than in an infected animal), and more sensitive or additional test(s) should be considered for screening purposes.

At the population level, animal disease surveillance is the equivalent of a diagnostic test used at the individual level and LR can be used to assess and compare the accuracy of surveillance activities when measuring infections in animal populations (see [20] as an example).

Test result interpretation

The most common application of DSe and DSp is to combine these values with an estimate of the prevalence and calculate the positive (PPV) or negative predictive values (NPV) (1). Similarly, LR can be used to calculate, on the odds scale, the probability of infection after accounting for the evidence provided by the test result (post-test probability) as well as the probability of infection before running the test (pre-test probability) (10). The pre-test (or prior) probability corresponds to the prevalence when calculating either the PPV or NPV, while the post-test (or posterior) probability corresponds to the PPV or 1-NPV, according to whether the test result was positive or negative, respectively. Its calculation follows a three-step process:

Step 1 – Convert the pre-test probability of infection (*ProbPre-Test*) into a pre-test odds (*OddsPre-Test*):

$$Odds_{Pre-Test} = \frac{Probability_{Pre-Test}}{1 - Probability_{Pre-Test}} \quad \text{(Equation 10)}$$

Step 2 – Calculate the post-test odds (*OddsPost-Test*) by multiplying the LR of the observed test result and the $Odds_{Pre-Test}$:

$$Odds_{Post-Test} = LR_{test\ result} \times Odds_{Pre-Test} \quad \text{(Equation 11)}$$

Step 3 – Convert the *OddsPost-Test* into post-test probability ($Prob_{Post-Test}$):

$$Prob_{Post-Test} = \frac{LR_{test\ result} \times Prob_{Pre-Test}}{1 - Prob_{Pre-Test} + LR_{test\ result} \times Prob_{Pre-Test}} \quad \text{(Equation 12)}$$

All these steps can be merged into one long single step (7):

$$Odds_{Pre-Test} = \frac{Probability_{Pre-Test}}{1 - Probability_{Pre-Test}} \quad \text{(Equation 13)}$$

For example, we can calculate the post-test probability of TB infection for cattle which tested positive to the Enferplex bovine TB antibody test. Using an arbitrary 5% pre-test probability of infection (estimated prevalence in the tested population) and combining it with the positive evidence provided by the test dichotomised using the 'higher sensitivity' cut-off (LR^+ = 52.9, Table IV), the post-test probability of TB infection of the test-positive cattle would be 73.6%. If the cattle tested positive using the 'higher specificity' cut-off instead (LR^+ = 273.3, Table IV), its post-test probability would increase to 93.5%. These calculations can be cumbersome for routine use, but adapted tools such as customised calculators and nomograms have been developed to facilitate this process for the user (see [9] as an example). McGee (21) offers a simplified approach based on a mnemonic. He noted that, as long as the pre-test probability is between 10% and 90%, the LR of the test

Table IV

Comparison of estimates, with 95% confidence intervals (CIs) in parentheses, of the diagnostic sensitivity (DSe), specificity (DSp), likelihood ratio of a positive test result (LR^+), likelihood ratio of a negative test result (LR^-) and diagnostic odds ratio (DOR) of three assays for bovine tuberculosis serology in cattle that are OIE certified and included in the OIE test register (15)

LR^+ and LR^- were calculated from DSe and DSp, and their 95% CIs were computed using a stochastic process from the boundaries of the 95% CIs of DSe and DSp estimates. Note that the Enferplex assay has two sets of estimates for each cut-off option – one with higher DSe and one with higher DSp

Name of assay	Manufacturer	OIE approval no.	Date of certification	Type of assay (specimen)	Cut-off criteria	DSe (95% CI)	DSp (95% CI)	LR^+ (95% CI)	LR^- (95% CI)	DOR (95% CI)
IDEXX *M. bovis* antibody test kit	IDEXX Laboratories	20120107	May 2012	Indirect ELISA (plasma, serum)	na	64.6%[a] (59.7–69.5)	98.0%[a] (97.5–98.4)	32.3 (25.4–40.9)	0.361 (0.314–0.412)	89.4 (65.1–121.7)
BOVIGAM® - *Mycobacterium bovis* gamma interferon test kit for cattle	Prionics AG	20150110	May 2015	Sandwich ELISA (whole blood, plasma)	na	84.6%[b] (73.0–95.5)	97.4%[b] (87.5–99.6)	32.5 (6.6–125.7)	0.158 (0.088–0.285)	205.8 (28.8–864.7)
Enferplex bovine TB antibody test	Enfer Scientific ULC	20190113	May 2019	Indirect chemiluminescent multiplex ELISA * (serum, milk)	Higher DSe	84.7% (81.2–87.7)	98.4[c] (98.0%–99.8)	52.9 (42.1–66.4)	0.155 (0.126–0.191)	340.5 (241.8–474.1)
					Higher DSp	82.0%[c] (78.3%–85.0%)	99.7%[c] (99.5%–99.8%)	273.3 (164.5–449.8)	0.181 (0.149–0.218)	1514.0 (856.0–2637.5)

(a) Estimates from the table on page 2 of (16)
(b) Estimates for cattle using classical statistics version from the table on page 4 of (17)
(c) DSe estimates for 'all samples' from Table 4 and DSp estimates for 'non-boosted' samples in Table 3 of (18)

* Provisional recognition
ELISA: enzyme-linked inmmunosorbent assay
na: not applicable

Table V

Approximate impact of the diagnostic evidence provided by a given test result on the post-test probability of infection according to its diagnostic likelihood ratio (LR) value

These approximations apply when the pre-test probability of infection ranges between 10% and 90%. Adapted from McGee (21)

LR value	Impact on post-test probability of infection
10	+45%
5	+30%
2	+15%
1	No change
0.5	–15%
0.2	–30%
0.1	–45%

result changes the post-test probability by approximately a constant and that, consequently, only a few benchmark values need to be remembered (Table V). For LR > 1, the reference LRs are 2, 5 and 10, which each increase the post-test probability by increments of 15% approximately, i.e. + 15%, + 30% and + 45%, respectively. For LR < 1, the benchmark values are 0.5, 0.2 and 0.1, which each decrease the post-test probability by increments of 15%, i.e. – 15%, – 30% and – 45%. Even in the absence of a precise estimate of the pre-test probability, the LR value provides valuable insight into the strength and impact of the diagnostic evidence provided by a test.

Diagnostic likelihood ratio – conclusion

The DSe and DSp address the accuracy of diagnostic tests separately in infected and non-infected individuals and can sometimes be misleading when assessing the diagnostic value of a test's result. For instance, we have seen that, despite a marginally lower DSp (increased risk of false positive in the non-infected), the BOVIGAM TB serology assay may be an equivalent or better choice to be confident in a positive test result and rule in infection when compared to the IDEXX assay because of its much higher DSe (higher likelihood of true positives in the infected) (Table IV). Therefore, to select, interpret, and assess the fitness for purpose of diagnostic tests, we need to compare the likelihoods of test results being true vs false across both infected and non-infected individuals. Predictive values do

weight these likelihoods, but they also depend heavily on the prevalence of infection in the tested populations and, therefore, cannot be generalised. The authors propose the adoption of LR as an additional measure of test accuracy because it balances the likelihood of true vs false results and is of a more universal nature because it is population independent (6).

Nevertheless, LR does have some inherent limitations. Diagnostic sensitivity is known to vary according to the representation of the disease spectrum (e.g. clinical vs. subclinical) across the infected individuals and this may make the estimate population dependent (22). We expect LR to suffer from the same limitation, as it strongly depends on DSe. One mitigation option would be to breakdown the test outputs into incremental categories and to use category-specific estimates ($DSe_{category}$ and $DSp_{category}$ or $LR_{category}$). Each of these category-specific estimates would cover a smaller fraction of the infected population and would be less dependent on the study population's overall representation. The LR is also a relative measure and, as such, it ignores the absolute value of diagnostic accuracy of the test. For instance, a test with DSe = 40% and DSp = 96% would have the same LR^+ as a test with DSe = 80% and DSp = 92% (LR^+ = DSe/(1–DSp) = 10 for both) despite their absolute diagnostic accuracy being very distinct. In this scenario, complementary measures, such as the DOR, can be used instead to discriminate between these two tests (DOR of 16 vs 46, respectively). Equally, one could simply compare directly the two sets of DSe and DSp. Besides diagnostic accuracy, other ancillary criteria should be carefully considered when comparing a test's fitness for purpose. These considerations vary according to the intended purpose and the level of the decision-making (individual animal or population level), and may include fixed and variable costs, repeatability and reproducibility, technical complexity, high-throughput capacity and turn-around time, to list a few. Some category-specific LRs may suffer from less precise estimation due to lower sample sizes if too narrow or towards the tails of the distributions. In the Enferplex assay's example (Table III), the distributions of TB infected and non-infected cattle appeared distinct, which resulted in the intermediate categories having fewer observations and, therefore, less precise LR estimates. In other instances, these distributions are expected to overlap more and, therefore, the extreme categories are expected to have less precise estimates. When interpreting a test result using LR, the back-and-forth conversion of pre- and post-probabilities in the odds scale may be cumbersome and may increase the risk of miscalculations. However, to reduce this risk, a single formula calculation exists (Eq. 13), as well as numerous calculators and graphical tools (see [23] as an example).

Over all, the strengths of the LR largely outweigh its limitations, and its relevance and universality make it the logical next-generation measure of diagnostic test accuracy to adopt when selecting and interpreting a test and assessing its fitness for purpose. This is true for detection of any pathogen but is particularly important when testing for transboundary diseases, because the consequences of test misclassifications may impact animal health and welfare on the global scale.

Acknowledgements

We thank Prof. Ian Gardner and two anonymous reviewers, whose useful comments helped to improve the manuscript.

Le ratio de vraisemblance diagnostique – une nouvelle génération d'outils pour mesurer l'exactitude d'un test diagnostique

C.G.B. Caraguel & A. Colling

Résumé

Afin de sélectionner, interpréter et évaluer l'aptitude à l'emploi de tests diagnostiques, nous devons comparer les probabilités de vrais *vs* faux résultats tant chez les individus infectés que chez les individus non infectés. La sensibilité diagnostique (DSe) et la spécificité diagnostique (DSp) correspondent à la capacité de classifier correctement les individus infectés ou non infectés séparément, sans toutefois comparer directement ces probabilités. Les valeurs prédictives positive et négative combinent ces probabilités mais sont fortement tributaires de la prévalence au sein de la population testée et par conséquent ne peuvent être généralisées. Les auteurs proposent d'adopter le ratio de

vraisemblance diagnostique (likehood ratio : LR), qui compare les probabilités que les résultats diagnostiques soient vrais ou faux indépendamment de la population testeé. En tant que mesure relative, le LR ignore l'exactitude absolue d'un test, de sorte que deux tests présentant des profils d'exactitude différents peuvent avoir le même LR. Ces cas peuvent être résolus en faisant appel à d'autres mesures d'exactitude, y compris la DSe et la DSp, ou à des critères alternatifs. Globalement, le LR constitue une mesure plus pertinente et universelle, ce qui justifie son adoption comme la nouvelle génération de mesure pour l'exactitude diagnostique. Les auteurs illustrent les applications et les avantages du LR à travers l'exemple de trois tests validés par l'Organisation mondiale de la santé animale pour la sérologie de la tuberculose bovine.

Mots-clés
Exactitude – Sensibilité diagnostique – Spécificité diagnostique – Test diagnostique – Ratio de vraisemblance.

La razón de verosimilitudes del diagnóstico, parámetro de próxima generación para medir la exactitud de las pruebas de diagnóstico

C.G.B. Caraguel & A. Colling

Resumen
Para seleccionar e interpretar pruebas de diagnóstico y valorar en qué medida cumplen su propósito debemos comparar la probabilidad de que los resultados que deparen sean verdaderos o falsos, y ello en individuos tanto infectados como no infectados. La sensibilidad y la especificidad de diagnóstico dan cuenta de la exactitud con que se pueden catalogar los individuos como infectados y no infectados separadamente, sin comparación directa entre estas dos probabilidades. Los valores predictivos positivos y negativos combinan ambas probabilidades, pero son muy dependientes de la prevalencia en la población analizada, por lo que no es posible generalizarlos. Los autores proponen que se adopte la razón de verosimilitudes (*likelihood ratio*: LR) del diagnóstico, que da cuenta de la probabilidad de obtener un resultado verdadero frente a la de obtener un resultado falso con independencia de la población de que se trate. Siendo un parámetro relativo, la LR es ajena a la exactitud absoluta de la prueba. Puede ocurrir por lo tanto que dos pruebas con distintas características de exactitud presenten la misma LR, extremo que se puede paliar fácilmente empleando una lista de medidas complementarias de la exactitud, entre ellas la DSe y la DSp, o criterios de selección auxiliares. Globalmente, la LR ofrece una medida más pertinente y universal de la exactitud de una prueba de diagnóstico, lo que justifica que sea adoptada como parámetro de próxima generación. Los autores ilustran las aplicaciones y ventajas de la LR aplicándola a tres ensayos aprobados por la Organización Mundial de Sanidad Animal como pruebas serológicas de detección de la tuberculosis bovina.

Palabras clave
Especificidad de diagnóstico – Exactitud – Prueba de diagnóstico – Razón de verosimilitudes – Sensibilidad de diagnóstico.

References

1. World Organisation for Animal Health (OIE) (2018). – Chapter 1.1.6. Principles and methods of validation of diagnostic assays for infectious diseases. *In* Manual of Diagnostic Tests and Vaccines for Terrestrial Animals, 25th Ed. OIE, Paris, France. Available at: www.oie.int/fileadmin/Home/eng/Health_standards/tahm/1.01.06_VALIDATION.pdf (accessed on 1 April 2020).

2. Yerushalmy J. (1947). – Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Pub. Hlth Rep.*, **62** (40), 1432–1449.

3. Dohoo I., Martin W. & Stryhn H. (2009). – Chapter 5: Screening and diagnostic tests. *In* Veterinary epidemiologic research (I. Dohoo, W. Martin & H. Stryhn, eds), 2nd Ed. VER Inc, Charlottetown, Canada, 91–134.

4. World Organisation for Animal Health (OIE) (2018). – Chapter 2.2.5. Statistical approaches to validation. *In* Manual of Diagnostic Tests and Vaccines for Terrestrial Animals, 25th Ed. OIE, Paris, France. Available at: www.oie.int/fileadmin/Home/eng/Health_standards/tahm/2.02.05_STATISTICAL_VALIDATION.pdf (accessed on 1 April 2020).

5. Caraguel C.G.B., Gardner I.A. & Hammell L.K. (2015). – Selection and interpretation of diagnostic tests in aquaculture biosecurity. *J. Appl. Aquacult.*, **27** (3), 279–298. doi:10.1080/10454438.2014.922154.

6. Dujardin B., Van den Ende J., Van Gompel A., Unger J.P. & Van der Stuyft P. (1994). – Likelihood ratios: a real improvement for clinical decision making? *Eur. J. Epidemiol.*, **10** (1), 29–36. doi:10.1007/BF01717448.

7. Timsit E., Léguillette R., White B.J., Larson R.L. & Buczinski S. (2018). – Likelihood ratios: an intuitive tool for incorporating diagnostic test results into decision-making. *JAVMA*, **252** (11), 1362–1366. doi:10.2460/javma.252.11.1362.

8. Grimes D.A. & Schulz K.F. (2005). – Refining clinical diagnosis with likelihood ratios. *Lancet*, **365** (9469), 1500–1505. doi:10.1016/S0140-6736(05)66422-7.

9. Caraguel C.G.B. & Vanderstichel R. (2013). – The two-step Fagan's nomogram: *ad hoc* interpretation of a diagnostic test result without calculation. *BMJ Evid. Based Med.*, **18** (4), 125–128. doi:10.1136/eb-2013-101243.

10. Altman D.G. (2000). – Diagnostic tests. *In* Statistics with confidence (D.G. Altman, D. Machin, T.N. Bryant & M.J.Gardner, eds), 2nd Edition. BMJ Books, London, United Kingdom, 105–119.

11. Gardner I.A. & Greiner M. (2006). – Receiver-operating characteristic curves and likelihood ratios: improvements over traditional methods for the evaluation and application of veterinary clinical pathology tests. *Vet. Clin. Pathol.*, **35** (1), 8–17. doi:10.1111/j.1939-165X.2006.tb00082.x.

12. Whelan C., Whelan A.O., Shuralev E., Kwok H.F., Hewinson G., Clarke J. & Vordermeier H.M. (2010). – Performance of the Enferplex TB assay with cattle in Great Britain and assessment of its suitability as a test to distinguish infected and vaccinated animals. *Clin. Vacc. Immunol.*, **17** (5), 813–817. doi:10.1128/CVI.00489-09.

13. Clegg T.A., Duignan A., Whelan C., Gormley E., Good M., Clarke J., Toft N. & More S.J. (2011). – Using latent class analysis to estimate the test characteristics of the γ-interferon test, the single intradermal comparative tuberculin test and a multiplex immunoassay under Irish conditions. *Vet. Microbiol.*, **151** (1), 68–76. doi:10.1016/j.vetmic.2011.02.027.

14. Hayden S.R. & Brown M.D. (1999). – Likelihood ratio: a powerful tool for incorporating the results of a diagnostic test into clinical decision making. *Ann. Emerg. Med.*, **33** (5), 575–580. doi:10.1016/S0196-0644(99)70346-X.

15. World Organisation for Animal Health (OIE) (2020). – Register of diagnostic kits certified by the OIE as validated as fit for purpose. OIE, Paris, France. Available at: www.oie.int/en/scientific-expertise/registration-of-diagnostic-kits/the-register-of-diagnostic-kits/ (accessed on 1 April 2020).

16. World Organisation for Animal Health (OIE) (2012). – IDEXX *M. bovis* antibody test kit. OIE procedure for registration of diagnostic kits – abstract sheet. OIE, Paris, France. Available at: www.oie.int/fileadmin/Home/eng/Our_scientific_expertise/docs/pdf/OIE_Register_MBAT_Abstract_v1_06.2012.pdf (accessed on 1 April 2020).

17. World Organisation for Animal Health (OIE) (2015). – BOVIGAM® – *Mycobacterium bovis* gamma interferon test kit for cattle. OIE procedure for registration of diagnostic kits – abstract sheet. OIE, Paris, France. Available at: www.oie.int/fileadmin/Home/eng/Our_scientific_expertise/docs/pdf/OIE_Register_Bovigam_Abstract_v1_05.2015.pdf (accessed on 1 April 2020).

18. World Organisation for Animal Health (OIE) (2019). – Enferplex bovine TB antibody test. OIE procedure for registration of diagnostic kits – abstract sheet. OIE, Paris, France. Available at: www.oie.int/fileadmin/Home/eng/Our_scientific_expertise/docs/pdf/OIE%20Abstract%2026.07.19.pdf (accessed on 1 April 2020).

19. Glas A.S., Lijmer J.G., Prins M.H., Bonsel G.J. & Bossuyt P.M.M. (2003). – The diagnostic odds ratio: a single indicator of test performance. *J. Clin. Epidemiol.*, **56** (11), 1129–1135. doi:10.1016/S0895-4356(03)00177-X.

20. Gustafson L., Klotins K., Tomlinson S., Karreman G., Cameron A., Wagner B., Remmenga M., Bruneau N. & Scott A. (2010). – Combining surveillance and expert evidence of viral hemorrhagic septicemia freedom: a decision science approach. *Prev. Vet. Med.*, **94** (1), 140–153. doi:10.1016/j.prevetmed.2009.11.021.

21. McGee S. (2002). – Simplifying likelihood ratios. *J. Gen. Intern. Med.*, **17** (8), 647–650. doi:10.1046/j.1525-1497.2002.10750.x.

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

23. DocNomo v2.0.1 (2018). – iOs application to enhance the interpretation of diagnostic test results computer [software]. Apple App Store. Available at: apps.apple.com/us/app/docnomo/id901279945 (accessed on 5 January 2021).