ELSEVIER

Preventive Veterinary Medicine 45 (2000) 107–122

PREVENTIVE VETERINARY MEDICINE

www.elsevier.nl/locate/prevetmed

Conditional dependence between tests affects the diagnosis and surveillance of animal diseases

Ian A. Gardner[a,*], Henrik Stryhn[b], Peter Lind[b], Michael T. Collins[c]

[a]*Department of Medicine and Epidemiology, School of Veterinary Medicine, University of California, Davis, CA 95616, USA*

[b]*Danish Veterinary Laboratory, Bülowsvej 27, DK-1790 Copenhagen V, Denmark*

[c]*Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin, Madison, WI 53706, USA*

Abstract

Dependence between the sensitivities or specificities of pairs of tests affects the sensitivity and specificity of tests when used in combination. Compared with values expected if tests are conditionally independent, a positive dependence in test sensitivity reduces the sensitivity of parallel test interpretation and a positive dependence in test specificity reduces the specificity of serial interpretation. We calculate conditional covariances as a measure of dependence between binary tests and show their relationship to kappa (a chance-corrected measure of test agreement). We use published data for toxoplasmosis and brucellosis in swine, and Johne's disease in cattle to illustrate calculation methods and to indicate the likely magnitude of the dependence between serologic tests used for diagnosis and surveillance of animal diseases.

Keywords: Combined tests; Sensitivity; Specificity; Conditional dependence; Test covariance; Kappa; Parallel testing; Serial testing

1. Introduction

Combinations of tests are used in many diagnostic, health-certification, disease-surveillance and eradication programs for livestock disease. Multiple tests might be done on all samples or the tests might be done sequentially on a subset of samples depending on results of the initial test. Decision rules are then applied to the test results to classify

* Corresponding author. Tel.: +1-530-752-6992; fax: +1-530-752-0414.
E-mail address: iagardner@ucdavis.edu (I.A. Gardner)

0167-5877/00/$ – see front matter
PII: S0167-5877(00)00119-7

individuals as positive or negative. Typically, tests are assumed to be conditionally independent and the theoretical sensitivities and specificities of tests in combination are calculated directly from individual test values. However, for tests that measure similar biologic processes such as serum antibody responses to infectious agents, it is logical to expect that test results will be dependent, conditional on an animal's true status. For example, in an infected animal the serologic response measured by two different tests will tend to follow a similar time-dependent pattern. False-negative test results on both tests might be more likely early in the course of infection (e.g. first 14 days) or late in the infectious process (especially if organisms have intracellular sites, e.g. *Brucella* sp. *or Mycobacterium* sp.) or if agents become latent, e.g. pseudorabies virus. Similarly in non-infected animals, false-positive serologic responses attributable to vaccination or cross-reacting antibodies would also tend to be positively correlated on different serologic tests.

The need to consider the possibility of conditional dependence when multiple diagnostic tests are used has been discussed by several authors (Politser, 1982; Jones and McClatchey, 1988; Marshall, 1989; Chiecchio et al., 1994; Brenner, 1996). However, there is little evidence in the medical or veterinary that researchers who evaluate multiple tests for the same disease have considered test dependence. When we reviewed the veterinary literature, we found only one study of serologic tests for swine brucellosis (Ferris et al., 1995), in which the authors specifically referred to dependence of tests and no studies that estimated the magnitude of the dependence.

Because the extent and implications of conditional dependence of tests for animal diseases have not been adequately evaluated, we developed a mathematical model of dependence and determined its effects when two tests are used for disease diagnosis. We evaluated the extent of dependence using published data for tests for swine toxoplasmosis (Dubey et al., 1995a,b), swine brucellosis (Ferris et al., 1995) and Johne's disease (paratuberculosis) in cattle (Collins et al., 1991; Sockett et al., 1992a,b). We demonstrate that test dependence substantially changes the theoretical values of the sensitivity and specificity of combined tests from those obtained when conditional independence is assumed. In this paper, we only consider binary test results because although many diagnostic test outcomes are ordinal or continuous, such results usually are dichotomized to simplify interpretation and to allow decisions about interventions (treatment, culling, etc.).

2. Conditional dependence of tests

We begin with an example using two binary tests to explain the concepts of conditional dependence and independence. The term ''conditional'' refers to the infection status of animals as described in Section 2.1. We note that the terms ''dependence'' and ''correlation'' are used interchangeably by some authors (Politser, 1982; Chiecchio et al., 1994) but the former term is preferable when binary tests are used and will be used by us.

2.1. Definition and example

Two tests are conditionally independent when the sensitivity (or specificity) of the second test (T_2) does not depend on whether results of the first test (T_1) are positive or

Table 1
Expected cell counts and marginal totals for two conditionally independent binary tests when used on samples from 100 infected animals[a]

		Test 2 +	Test 2 -	
Test 1	+	72	18	90
	-	8	2	10
		80	20	100

[a] Sensitivities of tests 1 and 2 are 0.9 and 0.8, respectively.

negative among infected (or non-infected) individuals. For example, if a test with sensitivity=0.9 is used to test a population of 100 infected animals, we would expect 10 animals to yield false-negative test results. If a second test with a sensitivity=0.8 is used to test the 10 animals that tested negative initially and the two tests were conditionally independent, then 8 of 10 animals would be expected to test-positive on the second test (Table 1). Hence, the sensitivity of the second test is 0.8, regardless of results of the first test, i.e. $Pr(T_2+ \mid T_1+, \text{infected})=Pr(T_2+ \mid T_1-, \text{infected})=Pr(T_2+ \mid \text{infected})=0.8$. Similarly, if T_2 were done first, conditional independence means that $Pr(T_1+ \mid T_2+, \text{infected})=Pr(T_1+ \mid T_2-, \text{infected})=Pr(T_1+ \mid \text{infected})=0.9$. We note that if either of the tests is perfectly sensitive (specific), then the test sensitivities (specificities) are conditionally independent by definition.

Conditional dependence of test sensitivities occurs when the second (first) test has different sensitivities for infected animals that test-positive and for those that test-negative on the first (second) test. If a positive dependence exists — which is the most biologically plausible scenario — the sensitivity of T_2 among T_1-negative infected animals would be <0.8, and hence, fewer than 8 of 10 infected animals would be expected to test-positive on the second test. If the second test failed to detect additional infected animals (Table 2), then the dependence between the test sensitivities would be complete and the combined sensitivity, assuming a positive on either test was positive, would be 0.9. In Table 2, we note that $Pr(T_2+ \mid T_1+, \text{infected})=80/90\approx0.89$ compared with $Pr(T_2+ \mid T_1-, \text{infected})=0$. Also, note that when there is complete dependence in test sensitivities but the sensitivities are not equal, there may be discordant test results.

We can express conditional dependence between the sensitivities of two tests as $Pr(T_1+ \text{ and } T_2+ \mid \text{infected}) \neq Pr(T_1+ \mid \text{infected})Pr(T_2+ \mid \text{infected})$. Alternatively, this relationship can be written as $Pr(T_2+ \mid T_1+, \text{infected}) \neq Pr(T_2+ \mid T_1-, \text{infected})$ and $Pr(T_1+ \mid T_2+, \text{infected}) \neq Pr(T_1+ \mid T_2-, \text{infected})$. Similar expressions apply to dependence of test specificities but a dependence of test sensitivities does not necessarily imply a dependence of test specificities and vice versa.

Table 2
Expected cell counts and marginal totals for two completely dependent binary tests when used on samples from 100 infected animals[a]

		Test 2 +	Test 2 -	
Test 1	+	80	10	90
	-	0	10	10
		80	20	100

[a] Sensitivities of tests 1 and 2 are 0.9 and 0.8, respectively, and the sensitivity covariance=0.08.

2.2. Estimation of dependence between test sensitivities and specificities

Assume that two binary tests are applied simultaneously to each individual in two separate populations of infected and non-infected animals (defined by a ''gold standard'' test). For each population, pairwise frequencies of test results can be cross-classified in a 2×2-table according to the true status of each animal tested (Table 3). Alternatively, cell frequencies in the cross-classified tables can be represented as probabilities (p_{ijk}) where i is the T_1 result (1=positive, 0=negative), j is the T_2 result (1=positive, 0=negative) and k is the infection status (1=infected, 0=non-infected). These probabilities (Table 4) are obtained by dividing the cell counts and marginal totals in the infected and non-infected populations by n_1 and n_2, respectively (Table 3). In the infected population table, the marginal probabilities are the sensitivities (Se) and false-negative proportions for the two

Table 3
Observed cell counts and marginal totals of pairwise test results for two binary tests when used on samples from infected and non-infected animals

		Infected Test 2 +	Infected Test 2 -			Non-infected Test 2 +	Non-infected Test 2 -	
Test 1	+	a	b	a+b	+	e	f	e+f
	-	c	d	c+d	-	g	h	g+h
		a+c	b+d	n_1		e+g	f+h	n_2

Table 4
Observed cell probabilities (p_{ijk}) of pairwise test results for two binary tests when used on samples from infected and non-infected animals[a]

		Infected Test 2 +	Infected Test 2 -			Non-infected Test 2 +	Non-infected Test 2 -	
Test 1	+	p_{111}	p_{101}	Se_1	+	p_{110}	p_{100}	$1\text{-}Sp_1$
	-	p_{011}	p_{001}	$1\text{-}Se_1$	-	p_{010}	p_{000}	Sp_1
		Se_2	$1\text{-}Se_2$	1		$1\text{-}Sp_2$	Sp_2	1

[a] Subscripts i and j denote test results (1=positive, 0=negative) for tests 1 and 2, respectively, and k denotes infection status (1=infected, 0=non-infected). Marginal probabilities are from the sensitivity (Se) and specificity (Sp) of each test.

tests. Similarly, in the non-infected-population table, the marginal probabilities are the respective test specificities (Sp) and false-positive proportions. If tests are conditionally independent, then the expected proportion of infected animals that will test positive on both tests is Se_1Se_2 (p_{111} in the left 2×2-table in Table 4). Similarly, the expected proportion of non-infected animals that will test negative on both tests is Sp_1Sp_2 (p_{000} in the right 2×2-table in Table 4). Other probabilities can be similarly calculated.

When tests are conditionally dependent, expected and observed proportions of the four pairwise combinations of test results (T_1+T_2+, T_1+T_2-, T_1-T_2+, and T_1-T_2-) differ. Denote $\gamma_{Se}=p_{111}-Se_1Se_2$ and $\gamma_{Sp}=p_{000}-Sp_1Sp_2$. These terms, γ_{Se} and γ_{Se}, are the conditional covariances between the test outcomes in the two populations of infected and non-infected animals, respectively. The magnitude of the conditional covariances is directly affected by the magnitude of the Se and Sp values. The limits for γ_{Se} and γ_{Sp} are

$$\max(-(1-Se_1)(1-Se_2), -Se_1Se_2) \leq \gamma_{Se} \leq \min(Se_1(1-Se_2), Se_2(1-Se_1))$$

$$\max(-(1-Sp_1)(1-Sp_2), -Sp_1Sp_2) \leq \gamma_{Sp} \leq \min(Sp_1(1-Sp_2), Sp_2(1-Sp_1))$$

These bounds apply both to true probabilities and to estimates based on an observed table of data (in which case the bounds apply to the estimates and not to the true parameter values). A $\gamma_{Se}>0$ means a positive dependence in sensitivities and a $\gamma_{Se}<0$ means a negative dependence in sensitivities. Similar interpretations apply to γ_{Sp} values. Low absolute values of γ_{Se} and γ_{Sp} are desirable if tests are to be used in combination.

Inclusion of conditional covariances in the calculations results in the general formulation of expected values for cell probabilities (Table 5). For the data in Table 2 (where complete dependence in test sensitivities exists), $\hat{\gamma}_{Se}=0.8-0.72=0.08$. The estimate of 0.08 is also the upper limit for γ_{Se}, i.e. minimum of (0.8)(0.1) and (0.9)(0.2). Because covariance values do not provide a direct measure of the magnitude of dependence, it is useful to express the covariance as a proportion or percentage of the

Table 5
Observed (p_{ijk}) and expected cell probabilities for results of two binary tests in infected and non-infected animals[a]

Observed probability	Expected value
Infected animals	
p_{111}	$Se_1Se_2+\gamma_{Se}$
p_{101}	$Se_1(1-Se_2)-\gamma_{Se}$
p_{011}	$(1-Se_1)Se_2-\gamma_{Se}$
p_{001}	$(1-Se_1)(1-Se_2)+\gamma_{Se}$
Non-infected animals	
p_{110}	$(1-Sp_1)(1-Sp_2)+\gamma_{Sp}$
p_{100}	$(1-Sp_1)Sp_2-\gamma_{Sp}$
p_{010}	$Sp_1(1-Sp_2)-\gamma_{Sp}$
p_{000}	$Sp_1Sp_2+\gamma_{Sp}$

[a] Probabilities are functions of sensitivity (Se), specificity (Sp), sensitivity covariance (γ_{Se}), and specificity covariance (γ_{Sp}).

maximal possible value. For example, two other tests which both have Se=0.5 could also have a γ_{Se}=0.08 but this would only be 32% of the upper bound for γ_{Se} of 0.25.

For the data in Tables 1 and 2, independence and complete dependence, respectively, we calculated 95% confidence intervals (CIs) for sensitivity covariances to be −0.02–0.03 and 0.04–0.13, respectively, using the method described below. In the former case, because the CI includes 0, one would not reject the null hypothesis that the test sensitivities are independent, whereas in the latter case one would. Also in the latter case, it may seem contradictory that the CI exceeds the maximal range of γ_{Se} determined from the observed test sensitivities. However, the true sensitivities are unknown and may differ from the observed values so as to enable a wider range of sensitivity covariances. Therefore, whenever the covariance bounds are determined from observed data, one should be aware that they are also subject to uncertainty.

2.2.1. Formal estimation of conditional covariances

From the observed cell counts (Table 3), the conditional covariances are estimated by $\gamma_{Se}=a/n_1-(a+b)(a+c)/{n_1}^2$ and similarly for γ_{Sp}. These intuitive estimates are also the maximum-likelihood estimates. A commonly used measure of the strength of association between two binary outcomes is the odds ratio (OR). There are separate ORs for infected and non-infected populations (denoted OR_1 and OR_0, respectively). In Table 5, $\widehat{OR}_1=p_{111}p_{001}/p_{011}p_{101}$ and $\widehat{OR}_0=p_{110}p_{000}/p_{010}p_{100}$.

Note that independence of sensitivities ($\gamma_{Se}=0$) corresponds to $OR_1=1$, so that standard tests apply to this hypothesis. However, because γ_{Se} is not only a function of OR but also of the marginal probabilities Se_1 and Se_2, CIs for the OR do not readily translate to CIs for γ_{Se}. To compute CIs for γ_{Se}, we used a general method based on the so-called profile likelihood (Venzon and Moolgavkar, 1988). Recent comparisons of CIs for related quantities in 2×2-tables have shown this method to perform well (Newcombe, 1998a,b). The 95% CI consists of those values of γ for which the likelihood-ratio test of the hypothesis $\gamma_{Se}=\gamma$ would not be rejected at the 5% level for the observed table of data. In

particular, the likelihood-ratio test of independence is significant at the 5% level exactly when 0 is not contained in the 95% CI. Our implementation of the method (available from the authors on request) performs a direct maximization of the profile likelihood for a range of fixed values of γ_{Se}. Similarly, OR_0 can be used to describe γ_{Sp}.

2.3. Relationship between test covariances and kappa

Kappa (κ) is a widely used measure of test agreement (Maclure and Willett, 1987) and is calculated as the ratio of the difference between the observed agreement (p_{obs}) and chance agreement (p_c) to the maximal possible agreement beyond chance

$$\hat{\kappa} = \frac{p_{obs} - p_c}{1 - p_c} \tag{1}$$

The magnitude of κ in a mixed population of infected and non-infected animals is dependent on seven factors: sensitivity and specificity of each test, prevalence of infection, and sensitivity and specificity covariances between the tests. Because prevalence strongly affects κ, comparison and interpretation of κ values from different studies can be problematic (Thompson and Walter, 1988). However, if estimates of κ are calculated separately for infected and non-infected populations — denoted as κ_1 and κ_0, respectively — it can be readily shown that these estimates depend only on the respective test parameters and covariances. For infected animals (Table 5), $p_{obs}=p_{111}+p_{001}=(Se_1Se_2+\gamma_{Se})+[(1-Se_1)(1-Se_2)+\gamma_{Se}]$. The proportion of chance agreement, p_c, is computed from the table margins and equals $Se_1Se_2+(1-Se_1)(1-Se_2)$. Substitution of these values for p_{obs} and p_c in Eq. (1) gives

$$\hat{\kappa}_1 = \frac{2\gamma_{Se}}{Se_1(1 - Se_2) + Se_2(1 - Se_1)} \tag{2}$$

For the data in Table 2, where $Se_1=0.9$ and $Se_2=0.8$ and there is complete dependence in test sensitivities ($\hat{\gamma}_{Se}=0.08$), $\hat{\kappa}_1=2(0.08)/[(0.9)(0.2)+(0.8)(0.1)]=0.62$.

Because of the symmetric nature of calculations for p_{obs} and p_c (see Tables 4 and 5), it is straightforward to show that the value for κ_0 can be obtained by substitution of γ_{Sp}, Sp_1, and Sp_2 for γ_{Se}, Se_1, and Se_2, respectively:

$$\hat{\kappa}_0 = \frac{2\gamma_{Sp}}{Sp_1(1 - Sp_2) + Sp_2(1 - Sp_1)} \tag{3}$$

If test sensitivities are conditionally independent (Table 1), then $\kappa_1=\gamma_{Se}=0$ and if test specificities are conditionally independent, then $\kappa_0=\gamma_{Sp}=0$. However, when there is complete dependence in test sensitivities, $\kappa_1=1$ only if the test sensitivities are equal (compare Eq. (2) with the sensitivity covariance bounds in Section 2.2).

3. Sensitivities, specificities, and predictive values of combined tests

Although many interpretation schemes are possible (especially when >2 tests are used), parallel and serial interpretation are most commonly used. Parallel interpretation schemes — also known as ''OR'' schemes — (positive on at least one test is positive, negative

otherwise) are used to maximize sensitivity. Parallel interpretation typically is used in animal trade because the primary goal is to detect all infected animals in a shipment. Animals that are positive on at least one test usually are excluded from shipment and animals testing negative are usually considered eligible for movement (MacDiarmid, 1993). Serial interpretation schemes — also known as "AND" schemes (positive on all tests is positive, negative otherwise) — are used when the goal is to maximize the specificity of diagnosis. Serial interpretation is most appropriate when the cost of a false-positive diagnosis is high (e.g. Johne's disease in a bull of high genetic merit).

3.1. Sensitivity and specificity of combined tests

If values for sensitivities, specificities and test covariances have been estimated, it is possible to calculate general formulas for the sensitivities and specificities of combined tests using expected values from Table 5. For a parallel-interpretation scheme, the sensitivity is thus $1-p_{001}=1-(1-\mathrm{Se}_1)(1-\mathrm{Se}_2)-\gamma_{\mathrm{Se}}$, and the specificity is $p_{000}=\mathrm{Sp}_1\mathrm{Sp}_2+\gamma_{\mathrm{Sp}}$. For a serial-interpretation scheme, the sensitivity is $p_{111}=\mathrm{Se}_1\mathrm{Se}_2+\gamma_{\mathrm{Se}}$, and the specificity is $1-p_{110}=1-(1-\mathrm{Sp}_1)(1-\mathrm{Sp}_2)-\gamma_{\mathrm{Sp}}$. Compared with values calculated assuming conditional independence, a positive sensitivity covariance reduces the expected gain in sensitivity from parallel testing and a positive specificity covariance reduces the expected gain in specificity from serial testing.

When test covariances are at their maximum values, and the performance of tests differs (such that $\mathrm{Se}_1>\mathrm{Se}_2$ and $\mathrm{Sp}_1>\mathrm{Sp}_2$), then the above relationships can be simplified. For a parallel-interpretation scheme, the sensitivity and specificity of the combined tests reduce to Se_1 and Sp_2, respectively. For example, for tests of Se=0.9 and Se=0.8 and positive test dependence, the expected sensitivity of a parallel-interpretation scheme would range from 0.98 (conditional independence in Table 1) to 0.9 (complete dependence in Table 2). For a serial-interpretation scheme, the sensitivity and specificity of the combined tests, if complete dependence exists, are Se_2 and Sp_1, respectively.

Knowledge of test dependence can aid test selection. Suppose that an investigator had the opportunity to choose two of three tests, of comparable cost and ease of performance, for use in a parallel testing scheme for early detection of infection. These tests were shown to have sensitivities of 0.9, 0.8 and 0.7. If test sensitivities were independent, the best choice would be the two most-sensitive tests. However, if their sensitivities were dependent, a combination of the two least sensitive — but conditionally independent — tests might yield better or at least comparable sensitivity. In this example ($\gamma_{\mathrm{Se}}=0$), the sensitivity of the parallel testing scheme for the 0.8 and 0.7 combination of independent tests would be $1-(1-0.8)(1-0.7)=0.94$ compared with a value of 0.9 for the 0.9 and 0.8 combination, assuming that the latter combination was completely positively dependent. Negative dependence would yield even higher sensitivity of parallel tests, but we consider this scenario to be unlikely and practically unimportant, at least for serologic tests.

3.2. Predictive values

Estimates in Table 5 can be used with Bayes' theorem to derive predictive values for the four combinations of test results (Table 6). In a serial-interpretation scheme, only a

Table 6
Predictive values of two tests (1 and 2) applied to a population with disease prevalence, P, including sensitivity (γ_{Se}) and specificity (γ_{Sp}) covariances

Test results		Pr(D+ | test results)
1	2	
+	+	$\frac{(Se_1Se_2+\gamma_{Se})P}{(Se_1Se_2+\gamma_{Se})P+((1-Sp_1)(1-Sp_2)+\gamma_{Sp})(1-P)}$
+	−	$\frac{(Se_1(1-Se_2)-\gamma_{Se})P}{(Se_1(1-Se_2)-\gamma_{Se})P+((1-Sp_1)Sp_2-\gamma_{Sp})(1-P)}$
−	+	$\frac{((1-Se_1)Se_2-\gamma_{Se})P}{((1-Se_1)Se_2-\gamma_{Se})P+(Sp_1(1-Sp_2)-\gamma_{Sp})(1-P)}$
−	−	$\frac{((1-Se_1)(1-Se_2)+\gamma_{Se})P}{((1-Se_1)(1-Se_2)+\gamma_{Se})P+(Sp_1Sp_2+\gamma_{Sp})(1-P)}$

positive result on both tests would be considered positive and this value can be obtained directly from the table. For a parallel-interpretation scheme, only negative results on both tests would be considered negative; hence, the predictive value of a positive result (considering any positive result as positive) would be $1-\Pr(T_1-T_2-)$.

When a single test is used, the gain in certainty from a positive or negative test can be assessed by comparing the pre-test and post-test probabilities. This gain can be expressed as a difference or as percentage change relative to the pre-test probability (Connell and Koepsell, 1985). The same approach to estimate gain in certainty can be readily extended to the use of multiple tests.

4. Examples

We obtained data from test-evaluation studies for toxoplasmosis (Dubey et al., 1995a,b) and brucellosis (Ferris et al., 1995) in pigs, and Johne's disease in cattle (Collins et al., 1991; Sockett et al., 1992a,b) from the respective authors. Each data set included test results and true infection status as defined by the ''gold standard'' test. We calculated covariances and 95% CI, covariances expressed as a percentage of their maximum possible values, and κ values for pairs of tests.

4.1. Swine toxoplasmosis

Ingestion of undercooked infected pork is considered an important source of human toxoplasmosis. Serologic screening may have future use for surveillance of animal populations and certification of freedom from *Toxoplasma gondii*. Dubey et al. (1995b) compared five serologic tests (modified agglutination test, latex-agglutination test, indirect hemagglutination test, enzyme-linked immunosorbent assay (ELISA) and Sabin–Feldman dye test) for the diagnosis of toxoplasmosis in 1000 naturally exposed sows (isolation of *T. gondii* from cardiac muscle was the reference test). Isolations were done in

Table 7
Test covariances (95% CI), covariances expressed as a proportion of their maximum value and kappa for five serologic tests for *T. gondii* in pigs[a]

Tests		Sensitivity covariance (95% CI)	Covariance/ maximum value	Kappa	Specificity covariance (95% CI)	Covariance/ maximum value	Kappa
IHAT[b]	MAT[c]	0.05 (0.03, 0.07)	1.0	0.16	0.01 (0.00, 0.02)	0.53	0.14
LAT[d]	MAT	0.08 (0.06, 0.10)	1.0	0.30	0.02 (0.01, 0.03)	0.70	0.32
ELISA[e]	MAT	0.08 (0.05, 0.11)	0.66	0.47	0.04 (0.03, 0.06)	0.51	0.41
DT[f]	MAT	0.09 (0.07, 0.12)	1.0	0.40	0.04 (0.03, 0.05)	0.45	0.40
LAT	IHAT	0.14 (0.11, 0.16)	0.85	0.56	0.01 (0.00, 0.02)	0.63	0.44
ELISA	IHAT	0.06 (0.04, 0.09)	0.78	0.21	0.01 (0.00, 0.01)	0.50	0.09
DT	IHAT	0.06 (0.03, 0.10)	0.47	0.25	0.01 (0.00, 0.01)	0.29	0.09
ELISA	LAT	0.08 (0.05, 0.11)	0.66	0.32	0.02 (0.01, 0.03)	0.60	0.20
DT	LAT	0.08 (0.05, 0.12)	0.40	0.33	0.02 (0.01, 0.03)	0.52	0.25
DT	ELISA	0.09 (0.06, 0.12)	0.59	0.36	0.03 (0.02, 0.04)	0.38	0.28

[a] Data from Dubey et al. (1995a,b).
[b] Indirect hemagglutination test (Se=0.29, Sp=0.98).
[c] Modified agglutination test (Se=0.83, Sp=0.90).
[d] Latex-agglutination test (Se=0.46, Sp=0.97).
[e] Enzyme-linked immunosorbent assay (Se=0.73, Sp=0.86).
[f] Sabin–Feldman dye test (Se=0.54, Sp=0.91).

mice (all sows) and cats (183 sows). If *T. gondii* was isolated from either mice or cats, the sow was considered infected. Our analysis indicated pairwise sensitivity and specificity covariances from 40 to 100% and from 29 to 70% of their maximum values, respectively (Table 7). Confidence intervals for all sensitivity and specificity covariances excluded 0.

4.2. Swine brucellosis

Screening for brucellosis in swine herds has typically been done by serologic methods. Ferris et al. (1995) compared six serologic tests commonly used at diagnostic laboratories in USA: particle-concentration fluorescence immunoassay, the automated complement-fixation assay, the card test, the buffered acidified-plate antigen assay, the standard tube test and the rivanol test. Serologic-test results were compared with culture results of eight lymph nodes collected at slaughter from 221 swine from 39 known brucellosis-infected herds. Consistent with the authors' conclusion that there was a high degree of dependence among test results of infected pigs, we found high positive ($\geq$77% of maximum value) sensitivity and specificity covariances for all pairs of tests. Nine of 15 pairwise test combinations had complete dependence in both sensitivity and specificity (Table 8).

4.3. Johnes disease in cattle

Ante-mortem diagnosis of subclinical Johne's disease in cattle is based on culture of feces for *M. paratuberculosis* and serologic testing. Until about 1990, the complement-fixation (CF) test was the standard serologic test for Johne's disease but ELISAs and an agar-gel immunodiffusion (AGID) were also developed. A repository of samples from

Table 8
Test covariances (95% CI), covariances expressed as a proportion of their maximum values and kappa for six serologic tests for swine brucellosis[a]

Tests		Sensitivity covariance (95% CI)	Covariance/ maximum value	Kappa	Specificity covariance (95% CI)	Covariance/ maximum value	Kappa
PCFIA[b]	CARD[c]	0.13 (0.07, 0.19)	1.0	0.67	0.07 (0.04, 0.11)	1.0	0.83
ACF[d]	CARD	0.16 (0.10, 0.21)	0.88	0.68	0.05 (0.03, 0.09)	0.90	0.78
BAPA[e]	CARD	0.16 (0.10, 0.21)	1.0	0.79	0.07 (0.04, 0.11)	1.0	0.89
RIV[f]	CARD	0.16 (0.09, 0.21)	0.77	0.66	0.04 (0.02, 0.08)	1.0	0.75
STT[g]	CARD	0.12 (0.06, 0.18)	1.0	0.61	0.05 (0.03, 0.08)	1.0	0.25
ACF	PCFIA	0.11 (0.06,0.16)	1.0	0.48	0.06 (0.03, 0.09)	1.0	0.71
BAPA	PCFIA	0.15 (0.08, 0.21)	1.0	0.87	0.09 (0.06, 0.13)	1.0	0.94
RIV	PCFIA	0.12 (0.07, 0.18)	1.0	0.54	0.04 (0.02, 0.08)	1.0	0.59
STT	PCFIA	0.14 (0.07, 0.21)	1.0	0.93	0.06 (0.04, 0.09)	0.92	0.31
BAPA	ACF	0.14 (0.08, 0.19)	1.0	0.58	0.06 (0.03, 0.09)	1.0	0.77
RIV	ACF	0.20 (0.14, 0.23)	0.90	0.81	0.04 (0.02, 0.07)	0.87	0.77
STT	ACF	0.10 (0.05, 0.15)	1.0	0.43	0.04 (0.02, 0.06)	1.0	0.20
RIV	BAPA	0.13 (0.07, 0.18)	0.84	0.55	0.04 (0.02, 0.07)	1.0	0.64
STT	BAPA	0.13 (0.07, 0.20)	1.0	0.80	0.05 (0.03, 0.08)	0.91	0.27
STT	RIV	0.11 (0.06, 0.17)	1.0	0.48	0.03 (0.01, 0.05)	1.0	0.15

[a] Data from Ferris et al. (1995).
[b] Particle-concentration fluorescence immunoassay (Se=0.80, Sp=0.89).
[c] Card test (Se=0.67, Sp=0.92).
[d] Automated complement-fixation test (Se=0.57, Sp=0.94).
[e] Buffered acidified-plate antigen test (Se=0.76, Sp=0.90).
[f] Rivanol test (Se=0.58, Sp=0.95).
[g] Standard tube test (Se=0.83, Sp=0.62).

animals of known infection status allowed blinded evaluation of the sensitivity and specificity of tests by several investigators (Collins et al., 1991; Ridge et al., 1991; Sockett et al., 1992a,b). Our analysis of test dependence among pairs of five serologic tests (three ELISAs, a CF and an AGID) indicated moderate-to-high positive sensitivity covariances (53–89% of maximum values) and all CI for sensitivity covariances included positive values only. On the other hand, specificity covariances were small and the CI included 0 (Table 9).

4.4. Risk assessment of disease introduction

We follow with an application of test dependence to a simple risk assessment of disease introduction when all inputs are known with certainty. Suppose that a dairy herd has a history of Johne's disease and the true prevalence (P) of disease is 2%. A dairyman in another herd wants to buy $n=10$ replacement heifers from the source herd. Three options are considered for introduction of heifers: no testing, testing by ELISA only, or testing by ELISA and CF (estimates are done incorporating the sensitivity covariance estimate and also ignoring the covariance, i.e. assuming conditional independence). What are the comparative risks assuming that the recipient herd will become infected if at least one

Table 9
Test covariances (95% CI), covariances expressed as a proportion of their maximum value and kappa for five serologic tests for subclinical Johne's disease in cattle[a]

Tests		Sensitivity covariance (95% CI)	Covariance/ maximum value	Kappa	Specificity covariance (95% CI)	Covariance/ maximum value	Kappa
UWELISA[b]	ALLELISA[c]	0.09 (0.06, 0.13)	0.53	0.41	0.008 (−0.005, 0.027)	0.24	0.06
CSLELISA[d]	ALLELISA	0.15 (0.12, 0.18)	0.81	0.59	−0.000 (−0.002, 0.009)	0	−0.02
AGID[e]	ALLELISA	0.09 (0.06, 0.11)	0.79	0.32	0.000 (−0.001, 0.009)	0	0
CF[f]	ALLELISA	0.12 (0.09, 0.15)	0.75	0.47	−0.000 (−0.002, 0.009)	0	−0.02
CSLELISA	UWELISA	0.10 (0.07, 0.13)	0.71	0.37	0.002 (−0.004, 0.015)	0.32	0.02
AGID	UWELISA	0.08 (0.05, 0.10)	0.93	0.26	0.000 (−0.003, 0.007)	0	0
CF	UWELISA	0.10 (0.08, 0.13)	0.86	0.37	0.002 (−0.004, 0.015)	0.32	0.02
AGID	CSLELISA	0.13 (0.10, 0.16)	0.89	0.56	0.000 (−0.000, 0.010)	0	0
CF	CSLELISA	0.14 (0.10, 0.17)	0.63	0.56	−0.000 (−0.001, 0.009)	0	−0.01
CF	AGID	0.13 (0.10, 0.16)	0.79	0.57	0.000 (−0.000, 0.010)	0	0

[a] Data from Collins et al. (1991) and Sockett et al. (1992a,b).
[b] University of Wisconsin enzyme-linked immunosorbent assay (Se=0.69, Sp=0.73).
[c] Allied Laboratories enzyme-linked immunosorbent assay (Se=0.59, Sp=0.95).
[d] Commonwealth Serum Laboratories enzyme-linked immunosorbent assay (Se=0.44, Sp=0.99).
[e] Agar-gel immunodiffusion test (Se=0.27, Sp=1).
[f] Complement fixation test (Se=0.39, Sp=0.99).

infected heifer is introduced? For the two testing options, assume that test-positive heifers are rejected and only test-negative heifers are introduced. Another option is to reject the shipment if any test-positive results are obtained but we do not consider this possibility in our example.

From Table 9, ALLELISA has Se=0.59 and Sp=0.95, CF has Se=0.39 and Sp=0.99, γ_{Se}=0.12 and γ_{sp}=0 between the tests. Substitution of these values in the formulas in Section 3.1 yields a sensitivity of the ALLELISA and CF in parallel of 0.63 (γ_{Se}=0.12) or 0.75 (γ_{Se} assumed to be 0), and a specificity of 0.9405. Using a simple binomial model of risk introduction, the probability of disease introduction if there is no testing=1−Pr (all heifers are non-diseased)=$1-(1-P)^n=1-(0.98)^{10}=0.18$. The negative predictive value (NPV) for each testing scenario: ALLELISA, ALLELISA and CF with γ_{Se}=0.12, and ALLELISA and CF assuming γ_{Se}=0 are then calculated. The resulting values are 0.9913, 0.9920, and 0.9946, respectively. Substitution of NPV values for $(1-P)$ in the equation above results in estimates of disease introduction risk of approximately 0.084, 0.077 and 0.053, respectively, compared with the unmitigated risk of 0.18. Because of the high sensitivity covariance (75% of its maximum value), the addition of the CF test to the testing protocol only minimally reduces the risk. Moreover, had the ALLELISA and CF been assumed to be conditionally independent — which is typically done — the reduction in disease risk would have been overestimated. The effects of uncertainty in input parameters can be further assessed by stochastic risk modeling and the value of use of the CF in addition to the ALLELISA can be evaluated by decision analysis (Smith, 1993; see also Smith and Slenning, 2000).

5. Discussion

For all three diseases, there was a moderate-to-high positive dependence in test sensitivities. On the other hand, the magnitude of positive dependence in specificities varied with disease. For brucellosis, pairwise specificity and sensitivity covariances were very high, which indicates that these tests measure essentially the same entity. Hence, substantially less than six tests likely are needed for serologic screening programs for brucellosis. For toxoplasmosis, specificities of the five serologic tests were moderately dependent but for Johne's disease, the specificities were independent. We assumed that the ''gold standard'' test result was valid for each disease. For brucellosis, lymph-node culture was used as the standard and because culture-negative animals were used from infected herds, it is possible that some animals used in the specificity covariance calculations were truly infected. For toxoplasmosis, two types of direct detection tests — mouse inoculation and cat feeding — were available as ''gold standards''. Although the cat data may be considered more definitive, they were biased in favor of sows with low antibody titers in the MAT and 6-fold more data were available by including the mouse inoculations. Therefore, we decided to use the combination of cat and mouse data for the present evaluation. For Johne's disease, the definitive test was isolation of *M. paratuberculosis* from fecal samples or tissues (terminal ileum or ileocecal lymph node). We consider this to be an appropriate standard for comparison of serologic tests.

We selected these three diseases for study because authors were willing to provide the original data and because test results were evaluated in a blinded fashion (which did not introduce additional observer dependence). The magnitude of pairwise dependence for other tests and diseases will differ, but our estimates should provide serologic-test users with guidance as to typical numeric values for dependence. If data on covariances are not available, it is reasonable to expect that the magnitude of conditional dependence should depend on the degree of relatedness of the tests and type of antigen used. For example, two enzyme-linked immunosorbent antibody (ELISA) assays using the same antigen might be expected to have a stronger positive dependence than an ELISA and an AGID test, which in turn should have a stronger dependence than an antibody ELISA and bacterial culture of a different sample type. In our study, however, the sensitivity covariances between all serologic tests were high regardless of test type. For example, two ELISAs for Johne's disease (UWELISA and ALLELISA, Table 9), which use the same antigen and protocol but different cut-off values for interpretation, had similar covariance values to other pairs of serologic tests. Similarly for toxoplasmosis, the positive dependence in sensitivity was high regardless of test type or antigen source: IHAT and LAT use soluble antigens attached to particles, the ELISA uses both soluble antigens and suspended membrane antigens, and the dye test and MAT use unfragmented parasites that are live and formalin-fixed, respectively.

We calculated conditional covariances as a simple measure of pairwise test dependence because they can be used directly to calculate the sensitivity and specificity of tests in combination. Covariance calculations can be extended to three or more tests; however, other analytic techniques such as log-linear and logistic modeling are preferable to investigate complicated dependence structures (see Hanson et al., 2000).

We emphasize that all covariance calculations were done at cut-off values for sensitivity and specificity as reported in the published papers and it is likely that use of other cut-off values for ordinal or continuous test results would have yielded different findings. For continuous tests, it is also possible to evaluate quantitative results for tests in combination. Instead of single cut-off points, methods such as discriminant analysis can be used to construct classification contours in the space of the combined test results (Schneider, 1994).

We demonstrated that κ — when calculated based on true status — depended only on test sensitivities or specificities and the respective covariances. However, it was interesting to see that for brucellosis tests with complete pairwise dependence, κ values ranged from 0.15 to 0.93. Hence, arbitrary criteria recommended for interpretation of κ by authors such as Landis and Koch (1977) may be misleading when tested populations only consist of infected or non-infected animals. We believe that test covariances likely will have greater utility than κ as a measure of pairwise dependence because covariances can be used directly to calculate the expected sensitivity and specificity of tests in combination.

In addition to use of two tests on a single occasion, there are other laboratory and field situations in which test dependence might be important. For example, test–retest of the same sample should yield test results that are strongly positively dependent, especially if laboratory error is minimal. Second, repeated testing of animals over time (as occurs in disease eradication programs) should also result in positive dependence. If an animal was

chronically infected and test-negative, sequential samples would also tend to be test-negative. Third, use of multiple tests on samples for several animals in a herd creates another form of dependence because factors that affect sensitivity and specificity (and presumably their respective covariances) might cluster among herds (Donald et al., 1994). Investigation of test dependence in these other situations is warranted but was outside the scope of the present study. Methods that address the problem of paired-sample designs for evaluating two tests are described elsewhere (see Greiner et al., 2000).

6. Conclusions

Estimation of the dependence of sensitivities and specificities should be considered as part of the analysis of studies comparing binary tests. We recommend that authors of these studies report test covariances so that end-users who select and interpret results of combinations of tests can have some indication whether the addition of more tests will increase the certainty of diagnosis or just the cost of the testing program. As an alternative, we recommend that authors report frequencies of all combinations of test results by infection status so that the extent of test dependence can be evaluated by others. Although our analysis was based on knowledge of infection status, it is possible to use latent-class methods to evaluate conditional dependence in the absence of a ''gold standard'' — provided that there are sufficient populations or tests (see Enøe et al., 2000).

Acknowledgements

We thank R.A. Ferris and J.P. Dubey for providing the swine brucellosis and toxoplasmosis data and B. Norby for technical assistance. The study was supported in part by the NRI Competitive Grants Program/USDA award No. 98-35204-6535.

References

Brenner, H., 1996. How independent are multiple independent diagnostic classifications? Statist. Med. 15, 1377–1386.

Chiecchio, A., Malvano, R., Giglioli, F., Bo, A., 1994. Performance assessment of coupled tests: the effects of statistical non-independence. Eur. J. Clin. Chem. Clin. Biochem. 32, 169–175.

Collins, M.T., Sockett, D.C., Ridge, S., Cox, J.C., 1991. Evaluation of a commercial enzyme-linked immunosorbent assay for Johne's disease. J. Clin. Microbiol. 29, 272–276.

Connell, F.A., Koepsell, T.D., 1985. Measures of gain in certainty from a diagnostic test. Am. J. Epidemiol. 121, 744–753.

Donald, A.W., Gardner, I.A., Wiggins, A.D., 1994. Cut-off points for aggregate testing in the presence of disease clustering and correlation of test errors. Prev. Vet. Med. 19, 167–187.

Dubey, J.P., Thulliez, P., Powel, E.C., 1995a. *Toxoplasma gondii* in Iowa sows: comparison of antibody titers to isolation of *T. gondii* by bioassays in mice and cats. J. Parasitol. 81, 48–53.

Dubey, J.P., Thulliez, P., Weigel, R.M., Andrews, C.D., Lind, P., Powel, E.C., 1995b. Sensitivity and specificity of various serologic tests for detection of *Toxoplasma gondii* infection in naturally infected sows. Am. J. Vet. Res. 56, 1030–1036.

Enøe, C., Georgiadis, M.P., Johnson, W.O., 2000. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Prev. Vet. Med. 45, 61–81.

Ferris, R.A., Schoenbaum, M.A., Crawford, R.P., 1995. Comparison of serologic tests and bacteriologic culture for detection of brucellosis in swine from naturally infected herds. J. Am. Vet. Med. Assoc. 207, 1332–1333.

Greiner, M., Pfeiffer, D., Smith, R.D., 2000. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Prev. Vet. Med. 45, 23–41.

Hanson, T.E., Johnson, W.O., Gardner, I.A., 2000. Log-linear and logistic modeling of dependence among diagnostic tests. Prev. Vet. Med. 45, 123–137.

Jones, R.H., McClatchey, M.W., 1988. Beyond sensitivity, specificity and statistical independence. Statist. Med. 7, 1289–1295.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

MacDiarmid, S.C., 1993. Risk analysis and the importation of animals and animal products. Rev. Sci. Tech. Off. Int. Epiz. 12, 1093–1107.

Maclure, M., Willett, W.C., 1987. Misinterpretation and misuse of the kappa statistic. Am. J. Epidemiol. 126, 161–169.

Marshall, R.J., 1989. The predictive value of simple rules for combining two diagnostic tests. Biometrics 45, 1213–1222.

Newcombe, R., 1998a. Interval estimation for the difference between independent proportions: comparison of eleven methods. Statist. Med. 17, 873–890.

Newcombe, R., 1998b. Improved confidence intervals for the difference between binomial proportions based on paired data. Statist. Med. 17, 2635–2650.

Politser, P., 1982. Reliability, decision rules and the value of repeated tests. Med. Decision Making 2, 47–69.

Ridge, S.E., Morgan, I.R., Sockett, D.C., Collins, M.T., Condron, R.J., Skilbeck, N.W., Webber, J.J., 1991. Comparison of the Johne's absorbed EIA and the complement-fixation test for the diagnosis of Johne's disease in cattle. Aust. Vet. J. 68, 253–257.

Schneider, B., 1994. Combining laboratory tests for diagnostic decisions. Eur. J. Clin. Chem. Clin. Biochem. 32, 177–178.

Smith, R., 1993. Decision analysis in the evaluation of diagnostic tests. J. Am. Vet. Med. Assoc. 203, 1184–1192.

Smith, R.D., Slenning, B.D., 2000. Decision analysis: dealing with uncertainty in diagnostic testing. Prev. Vet. Med. 45, 139–162.

Sockett, D.C., Carr, D.J., Richards, W.D., Collins, M.T., 1992a. A repository of specimens for comparison of diagnostic testing procedures for bovine paratuberculosis. J. Vet. Diagn. Invest. 4, 188–191.

Sockett, D.C., Conrad, T.A., Thomas, C.B., Collins, M.T., 1992b. Evaluation of four serologic tests for bovine paratuberculosis. J. Clin. Microbiol. 30, 1134–1139.

Thompson, W.D., Walter, S.D., 1988. A reappraisal of the kappa coefficient. J. Clin. Epidemiol. 41, 949–958.

Venzon, D.J., Moolgavkar, S.H., 1988. A method for computing profile-likelihood-based confidence intervals. Appl. Statist. 37, 87–94.