STATISTICA, anno LXIX, n. 1, 2009

EVALUATING SENSITIVITY AND SPECIFICITY OF THREE DIAGNOSTIC TESTS WHEN THE "GOLD STANDARD" IS UNAVAILABLE, WITH APPLICATION TO THE CATTLE Q FEVER IN SMALL RUMINANTS CASE STUDY

D. Basso, K. Capello, L. Corain, L. Salmaso

1. INTRODUCTION AND PRESENTATION OF THE PROBLEM

In sanitation, zoology, and biology fields, wide use is made of diagnostic tests (or screening tests), which are used to determine whether or not a patient suffers from a specific illness. Typically, a diagnostic test is based on a quantitative measurement Y, and whenever an observation's measurement exceeds a pre-specified benchmark y_0, the related observation is classified as "sick", otherwise it is classified as "healthy".

As in the statistical hypotheses testing theory, there are two kinds of errors which may arise: declaring a subject sick when it is healthy, and classifying a subject as healthy when it is sick. The performances of a diagnostic test are usually evaluated in terms of percentage of good diagnosis: the Specificity (Sp) of a test is the percentage of healthy subject correctly classified, and the Sensitivity (Se) is the percentage of sick subject correctly classified. Obviously, the benchmark y_0 and the knowledge of the true state of the illness in the population play a major role in determining the performances of a diagnostic test.

In order to calibrate (i.e. specify y_0) and evaluate a diagnostic test, some information is required about the true state of the subject. This information can be obtained from a pre-existing test, known to be the "best" one, or by invasive analysis (e.g., autopsy). The "best" source of information is usually indicated as the "gold standard" test, meaning that its diagnosis is the most reliable one available about the prevalence of the illness in the population (i.e. the true proportion of sick subjects).

If a gold standard is available, then a diagnostic test can be calibrated by computing sensitivity and specificity as a function of several different benchmarks, and then by choosing y_0 as the measurement which maximizes sensitivity and specificity of the test. This method of calibration is usually known as the ROC (receiver operator characteristic) curve. Once the test has been calibrated, its performances are summarized by the two indexes Sp and Se.

When several tests are available to identify the same illness, and a gold standard is also available, the comparison among them can be done by looking at their sensitivities and specificities which can be estimated from the data. Obviously, the best test is the one with higher Sp and Se.

There are other situations where the gold standard is not available because there is no previous information on the incidence of an illness or because the only way to determine the state of a subject is invasive or may determine the death of the subject. In this situation, if several diagnostic tests is available and if there are no information about their sensitivities and specificities, it is rather hard to determine which one is the best-performing diagnostic test. In the literature there are some proposals to answer the specific problem. Johnson et al. (2001) extend the method first proposed by Hui and Walter (1980), by giving a solution in the case where two tests are applied to two distinct populations, and no gold standard is available. Their solution is based on the assumption of conditional independence among the tests, and the six parameters of interest (the specificities and sensitivity of the two tests, and the prevalence of the two populations) are estimated by looking at the joint distribution of the test outcomes. The estimate of the parameters is obtained through the maximum likelihood method. Other authors proposed a Bayesian approach by applying some a priori distribution to the unknown parameters (Joseph et al., 1995; Neath and Samaniego 1997).

The goal of this work is to evaluate, with application to a real case study of Q fever, the performances of three diagnostic tests applied to the same population, when no gold standard is available. Moreover, as in the case study, we considered the situations where there is no information about the prevalence of the illness in the population under study. In the next section, the real case study is introduced. In Section 3 we specify the theoretical assumptions and propose a possible way to estimate the parameters of interest. In Section 4 we evaluate the proposed methodolgy with a simulation study, and finally we apply it to the observed data.

2. A REAL CASE STUDY

The case study concerns the Q fever which is a disease caused by infection with *Coxiella burnetii* a bacterium that affects both humans and animals (Marrie, 1990). This organism is uncommon but may be found in cattle, sheep, goats and other domestic mammals, including cats and dogs. The infection results from inhalation of contaminated particles in the air, and from contact with the vaginal mucus, milk, feces, urine or semen of infected animals. The incubation period is 9-40 days. It is considered possibly the most infectious disease in the world, as a human being can be infected by a single bacterium (Beare et al., 2006). Ruminants are considered to be the main source of infection of humans, with the main route of infection being through inhalation of the organism of fine-particle aerosols. Abortion is the main clinical sign in ruminants.

From a veterinary point of view, as reported by Guatteo et al. (2007), the characteristics of *Coxiella* shedding are still widely unknown, especially in dairy cattle. However, this information is crucial to assess the natural course of *Coxiella burnetii*

infection within a herd and then to elaborate strategies to limit the risks of transmission between animals and to humans.

Three diagnostic tests were applied to a sample of 1307 sera from sheep and goat collected during a monitoring activity of the disease in the Veneto region (northeast Italy). Each sample was tested with one complement fixation test (called FdC) and two indirect Elisa tests, one from IDEXX laboratories (called IDEXX) and the other from Pourquier Institute (called POURQUIER). For IDEXX and Pourquier some continuous measurements were available, whereas only the outcomes (positive/negative) of FdC were available. A synthesis of collected data is listed in Table 1.

TABLE 1

Marginal distribution of IDEXX, Pourquier and FdC

	IDEXX	**Pourquier**	**FdC**
–	1132	1197	1287
+	175	110	20
Total	**1307**	**1307**	**1307**

The estimated prevalence is 13.38% for IDEXX, 8.41% for Pourquier, and 1.53% for FdC. These results do not seem to agree with each other. As a further investigation, we considered Cohen's Kappa index of agreement between all possible pairs of tests. Since the observed frequencies of the joint distributions were sparse, we computed the exact test using a permutation approach (Mehta and Patel, 1998). The number of considered Monte Carlo permutations is 10,000.

As far as the IDEXX-Pourquier comparison is concerned, the bivariate joint distribution is displayed in Table 2. The resulting Kappa index is equal to 0.409, resulting in a high, significant (or highig significant **) p-value (exact $p < 0.0001$) against the null hypothesis that the observed agreement between tests is only due by chance.

The comparisons involving the FdC test gave significant results too, indicating that, at least, the concordance among the three tests is not due to chance. The joint distributions of [IDEXX, FdC] and [Pourquier, FdC] are listed in Tables 3 and 4.

TABLE 2

Joint distributions of IDEXX and Pourquier

		Pourquier		**Total**
		–	**+**	
IDEXX	**–**	1089	43	**1132**
	+	108	67	**175**
	Total	**1197**	**110**	**1307**

TABLE 3

Joint distributions of IDEXX and FdC

		FdC		**Total**
		–	**+**	
IDEXX	**–**	1119	13	**1132**
	+	168	7	**175**
	Total	1287	20	**1307**

TABLE 4

Joint distribution of Pourquier and FdC

		FdC		Total
		–	+	
Pourquier	**–**	1182	15	**1197**
	+	105	5	**110**
	Total	**1287**	**20**	**1307**

The Cohen index involving FdC and IDEXX is equal to 0.046, (exact $p = 0.0099$), and the Cohen index involving FdC and Pourquier is equal to 0.0524, (exact $p = 0.0231$). The asymptotic p-values of the comparisons involving FdC gave non significant results because of bad approximations due to sparse frequencies (they were equal to 0.2486 and 0.2675 respectively).

Finally, the joint distribution of the three tests is displayed in Table 5. This data will be the information required to apply our methodological proposal, which will be detail in the next section.

TABLE 5

Joint distributions of IDEXX, Pourquier, and FdC

Fdc = –		**Pourquier**			**Fdc = +**	**Pourquier**		
		–	**+**	**Total**		**–**	**+**	**Total**
IDEXX	**–**	1078	41	**1119**	**–**	11	2	**13**
	+	104	64	**168**	**+**	4	3	**7**
	Total	**1182**	**105**	**1287**	**Total**	**15**	**5**	**20**

3. PROPOSED METHODOLOGY

In this section, we are going to illustrate the parametric model we have chosen to describe the underlying results of the three diagnostic tests. We recall, from Section 2, that here no gold standard is supposed to be available. However, we will assume that an unknown gold standard exists, in order to model the "true" state of an observation. To this end, let G denote the gold standard and let π be the true, unknown incidence of the illness in the population. The possible outcomes of G on an observation can be "+" (Diseased) and "–" (Healthy). Then, by definition of gold standard, $\pi = \Pr[G = +]$, and $1 - \pi = \Pr[G = +]$.

From a statistical point of view, there are many similarities between diagnostic and statistical tests: they are functions of data into a response set $\{\Theta_0, \Theta_1\}$ and two kind of errors may arise. We will thus model the diagnostic test by the usual statistical notation. Let H_0 be the null hypothesis to be assessed on the ith observation, $i = 1,..., n$, where n is the sample size. Then we let H_0 be the event "the ith observation is healthy" versus the alternative hypothesis H_1: "the ith observation is sick". Let T_j be the jth diagnostic test. The possible outcomes of T_j are *positive* (i.e., the ith observation is sick, symbol "+") or *negative* (i.e. the ith observation is healthy, symbol "–"). Obviously, there are two kinds of errors that may arise and we will denote them using with the usual statistical notation:

$\alpha_j = \Pr\{T_j = + \mid G = -\}$, and $\beta_j = \Pr\{T_j = - \mid G = +\}$.

That is, α_j and β_j are probabilities of a 1st and a 2nd type error to occuring on the jth diagnostic test, respectively. In sanitation field, the jth diagnostic test is usually evaluated in terms of its *sensitivity* (Se_j) and *specificity* (Sp_j). The relationships among statistical and sanitation definitions are:

$Sp_j = 1 - \alpha_j$, and $Se_j = 1 - \beta_j$.

Clearly, a good diagnostic test shows high Specificity and Sensitivity. We have implicitly assumed that the gold standard G satisfies $Sp_G = Se_G = 1$ (i.e. the probability of misclassification of one observation is zero).

The joint distribution of the couple $\{T_j, G\}$ can be obtained by applying the Bayes relation $\Pr\{T_j, G\} = \Pr\{T_j \mid G\} \Pr\{G\}$, and it is illustrated in Table 6.

TABLE 6

Representation of the joint distribution of $\{T_j, G\}$

		Gold Standard (G)		
		+	–	$\Pr\{T_j\}$
T_j	+	$\pi(1-\beta_j)$	$(1-\pi)\alpha_j$	$(1-\pi)\alpha_j + \pi(1-\beta_j)$
	–	$\pi\beta_j$	$(1-\pi)(1-\alpha_j)$	$(1-\pi)(1-\alpha_j) + \pi\beta_j$
	$\Pr\{G\}$	π	$1-\pi$	1

Note that the marginal distribution of G is unknown, whereas the only information available from T_j is given by its marginal distribution $\Pr\{T_j\}$. Therefore, the marginal distribution of T_j is a function of three unknown parameters: π (common parameter), α_j, and β_j, $j = 1, 2, 3$.

The three diagnostic tests are applied to the same observations, and therefore their outcomes are dependent. However, we believe that the observed dependence is induced by the state of the ith observation, since the three tests are applied independently. Thus, in order to model the joint distribution of diagnostic tests, we assume conditional independence among tests; for instance, the joint distribution of $\{T_j, T_k\}$ is given by

$$\Pr\{T_j = t_j, T_k = t_k \mid G = g\} = \Pr\{T_j = t_j \mid G = g\} \Pr\{T_k = t_k \mid G = g\},$$

with t_j, t_k, g = "+", "–". Similarly, we assume the joint distribution of the three test is modelled by

$$\Pr\{T_1, T_2, T_3 | G\} = \Pr\{T_1 | G\} \Pr\{T_2 | G\} \Pr\{T_3 | G\},$$

where the symbol $\Pr\{T_j | G\}$ means $\Pr\{T_j = t_j | G = g\}$.

In order to model the joint distribution of the three tests, we need to integrate the joint ditribution of $\{T_1, T_2, T_3, G\}$ with respect to G. Therefore, by applying the Bayes theorem, we have:

$$\Pr\{T_1, T_2, T_3\} = \Pr\{T_1, T_2, T_3 | G\}\Pr\{G = +\} + \Pr\{T_1, T_2, T_3 | G\}\Pr\{G = -\}.$$

Let $F_j(t_j)$ denote the probability function of T_j, and let $F_{123}(t_1, t_2, t_3)$ denote the joint probability function of the three tests; then the joint distribution of $\{T_1, T_2, T_3\}$ is given by the following equations:

$$F_{123}(+, +, +) = (1-\pi)\alpha_1\alpha_2\alpha_3 + \pi(1-\beta_1)(1-\beta_2)(1-\beta_3) \quad (1)$$

$$F_{123}(+, +, -) = (1-\pi)\alpha_1\alpha_2(1-\alpha_3) + \pi(1-\beta_1)(1-\beta_2)\beta_3 \quad (2)$$

$$F_{123}(+, -, +) = (1-\pi)\alpha_1(1-\alpha_2)\alpha_3 + \pi(1-\beta_1)\beta_2(1-\beta_3) \quad (3)$$

$$F_{123}(+, -, -) = (1-\pi)\alpha_1(1-\alpha_2)(1-\alpha_3) + \pi(1-\beta_1)\beta_2\beta_3 \quad (4)$$

$$F_{123}(-, +, +) = (1-\pi)(1-\alpha_1)\alpha_2\alpha_3 + \pi\beta_1(1-\beta_2)(1-\beta_3) \quad (5)$$

$$F_{123}(-, +, -) = (1-\pi)(1-\alpha_1)\alpha_2(1-\alpha_3) + \pi\beta_1(1-\beta_2)\beta_3 \quad (6)$$

$$F_{123}(-, -, +) = (1-\pi)(1-\alpha_1)(1-\alpha_2)\alpha_3 + \pi\beta_1\beta_2(1-\beta_3) \quad (7)$$

$$F_{123}(-, -, -) = (1-\pi)(1-\alpha_1)(1-\alpha_2)(1-\alpha_3) + \pi\beta_1\beta_2\beta_3 \quad (8)$$

where the elements on the side of equation (1)→(8) can be estimated from the observed data.

Unfortunately, the above system of equations is unsuitable for obtaining proper estimates of the unknown parameters; indeed, they are linearly dependent since, for instance, we can obtain the probability of the event $\{T_1 = +, T_2 = +\}$ by adding (1) and (2). A further proof of this fact can be given by observing that the contingency table representing the joint distribution of $\{T_1, T_2, T_3\}$ has a single degree of freedom.

Thus, there are an infinite number of solutions to the system of equations (1)→(8). Moreover, the equations above are nonlinear in the parameters, although they can be linearized.

Therefore, no proper solution can be found for the given problem, and we have decided to apply nonlinear equation solving techniques to give a descriptive, rather than inferential, solution to the problem by setting the population parameter π, free and by expressing the remaining parameters as functions of $\pi \in [0, 1]$. This choice is motivated by the following reasons: (i) the sensitivity and specificity of the diagnostic tests should not depend on the true state of the illness incidence π; and (ii) no a priori information on π was given.

Thus, we let $\pi \sim U[0,1]$ and evaluate the performances of the three tests on the estimates of specificities and sensitivities from equations (1)→(8). The best performing test should be the one with higher power $(1-\beta)$ and smaller type I error rate (α). As a further (reasonable) assumption, we shall add the condition $1-\beta_j \geq \alpha_j \; \forall \; j$; that is, the power of one test should not be smaller than its type I error. This assumption is equivalent to the unbiasedness property of a statistical test, which states that the probability of rejecting the null hypothesis when it is true $(G = -)$ should not exceed the probability of rejecting the null hypothesis when the alternative is true $(G = +)$.

4. SIMULATIONS AND DISCUSSION OF THE RESULTS

We have run the computation and simulations by applying the `R` function "`optim`" (R Development Core Team, 2008). This function is usually applied to minimization problems, and returns a vector $\mathbf{x}$ that is the solution of the equation $g(\mathbf{x}) = c$, where $g : \mathbb{R}^p \to \mathbb{R}$ is a nonlinear function and $c = \min(g(\mathbf{x}))$ is a real value determining a (local) minimum of the function. The `optim` function requires a vector of initial values $\mathbf{x}_0$, then applies an optimization algorithm until the convergence criterion is satisfied (typically when $\|g(\mathbf{x}) - c\|_2 < \varepsilon$, where $\|\cdot\|_2$ is the L_2 norm and $\varepsilon > 0$ is the desired degree of precision.

We can apply the `optim` function as follows: from equation (1), we let

$$\mathbf{x} = [\pi, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3]^T,$$

and $\mathbf{f} = [f_1,..., f_8]$, where, for instance,

$$f_1 = f_1(\mathbf{x}) = (1-\pi)\alpha_1\alpha_2\alpha_3 + \pi(1-\beta_1)(1-\beta_2)(1-\beta_3) - \hat{F}_{123}(+, +, +), \tag{9}$$

where $\hat{F}_{123}(+,+,+)$ is the estimation of $F_{123}(+, +, +)$ obtained from the observed data. Clearly, $f_1 = 0$ implies that (9) is satisfied. In the same way, the remaining elements of $\mathbf{f}$ are equal to zero when the corresponding equations (2)→(8) are satisfied, provided that the probability functions are replaced by their estimates.

Finally, by setting $g(\mathbf{x}) = \mathbf{f}^T\mathbf{f}$ and $c = 0$ as the `optim` entries, we have that the minimum of function $g(\mathbf{x})$ is zero and the solution $\mathbf{x}^*$: $g(\mathbf{x}^*) = 0$ jointly satisfies the equations $f_k(\mathbf{x}) = 0$, $j = 1,...,8$, because

$$g(\mathbf{x}) = \mathbf{f}^T\mathbf{f} = \sum_{k=1}^{8} f_k^2 = 0,$$

implies $f_k(\mathbf{x}) = 0$, k = 1,..., 8.

As regards the initial vector of values $\mathbf{x}_0$, since we have no a priori information on α_j and β_j (j = 1, 2, 3), we set them equal to 0.5, and let π vary in [0, 1] in steps of 1/100. Finally, there are several minimization algorithms in the `optim` settings, and we have chosen the "L-BFGS-B" method (Byrd *et. al.*, 1995) since it is the only one which allows for constrained solutions (all the parameters take values in the [0,1] interval).

We begin with a simulation to evaluate the estimating method. To this end, we generated 100 independent data sets modelling either the true state of a sample of n = 1000 observations or the related outcomes of three diagnostic tests. As we mentioned before, the dependence among the outcomes is given by the fact that the three tests are applied to the same subjects.

The simulation settings are described in Table 7: here we have set T_1 as the best performing test and T_3 as the worst. The incidence of the illness is set equal to 0.1.

TABLE 7

Simulations estimates setting for illness incidence and specificity and sensitivity of three tests

		T_1		T_2		T_3	
Parameter	π	α	$1-\beta$	α	$1-\beta$	α	$1-\beta$
True value	**0.1**	**0.05**	**0.9**	**0.1**	**0.85**	**0.2**	**0.8**

As regards the simulation settings, we firstly generate the "true" state of illness on the ith observation by obtaining a random realization from the random variable $G_i \sim \text{Bi}(1, \pi)$, $i = 1,..., n$; then, conditionally on the outcome of G_i, we independently generated the one-to-one outcomes X_{ij} of each test on the ith observation with the conditional probabilities obtained from Table 1, namely:

$$\{X_{ij} | G_i = +\} \sim \text{Bi}(1, 1 - \beta_j),\ i = 1,..., n;$$

$$\{X_{ij} | G_i = -\} \sim \text{Bi}(1, \alpha_j),\ j = 1, 2, 3.$$

Then, for each data generation, we considered a lattice of 100 points representing the a priori distribution of π, from 0.01 to 1 in steps of 0.01. For each value of π, we ran the `optim` function with the vector of initial parameters given by $\mathbf{x_0} = [\pi, .5, .5, .5, .5, .5, .5,]$, and stored away the `optim` results. The graphical representation related to one generation of data is given in Figure 1: this figure represents the type I error rates (dotted lines) and the observed powers (solid lines) of each test as functions of π, with the settings given in Table 7.

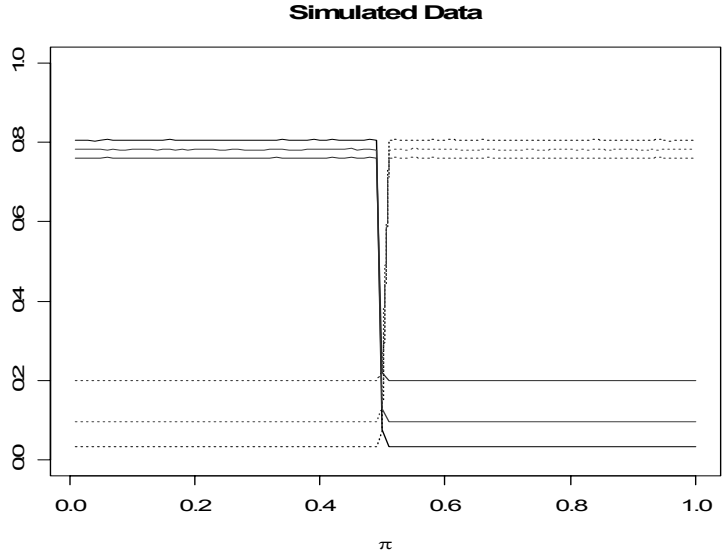

Figure 1 – Results of the optim function with L-BFGS-B minimization algorithm for one data generation. Dotted lines are type I° errors and solid lines represent the powers of each diagnostic test. T_1 = black lines; T_2 = red lines, and T_3 = blue lines. True illness incidence equal to 0.1.

The performances of T_1 are represented in black, those of T_2 in red and those of T_3 in blue. Note how the curves of powers and type I errors interchange after the point $\pi = 0.5$. This behaviour is due to the symmetry in formulas (1)→(8); i.e. when π is replaced by $1-\pi$ and powers are replaced by error rates. Assuming the condition $1 - \beta_j \geq \alpha_j$, our attention focuses on the first half of the [0, 1] interval. In this data generation, T_1 has a pretty constant power close to 80%, which is the highest, and a type I error close to 5%. The performances of the simulated tests are in accordance with the simulation settings, in the sense that here T_1 is the best performing test. Also note that the estimated power and type I error variability changes very little for $\pi \in [0, 0.5]$, indicating that the `optim` solution is a global minimum, rather than a local one.

Of course, if we set the true illness incidence above 0.5, the corresponding lines are interchanged, and the assumption $1 - \beta_j \geq \alpha_j$ suggests considering only the [0.5, 1] interval. Figure 2 shows the results of a generation with the same settings of Table 7, but here the true illness incidence was set equal at 0.75.

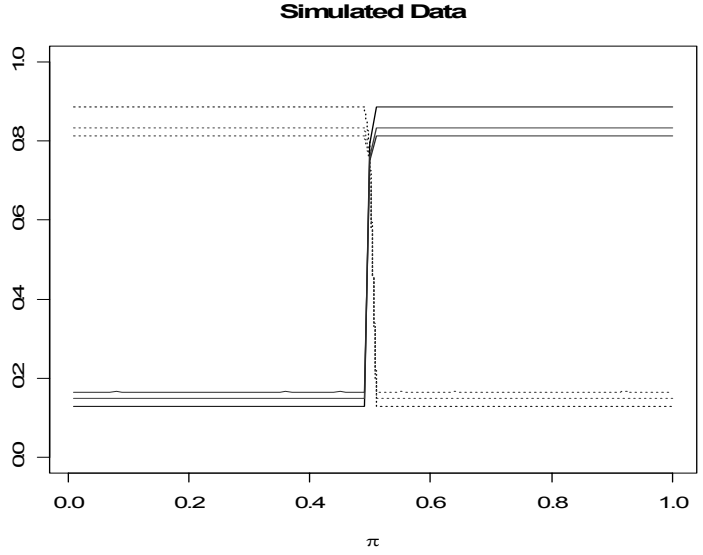

Figure 2 – Results of the optim function with L-BFGS-B minimization algorithm for one data generation. Dotted lines are type I errors and solid lines represent the powers of each diagnostic test. T_1 = black lines; T_2 = red lines, and T_3 = blue lines. True illness incidence equals to 0.75.

The simulation results relating to the Table 7 settings are summarized in Table 8: these results were obtained by storing the average estimated parameter given by the `optim` function corresponding to values of π smaller than 0.5.

Note how the estimates of the parameters are on average very close to the true values, and this is true also for the illness incidence.

The simulation shows that, at least on average, the proposed method gives unbiased estimates of the parameters. With this in mind, we have ran the estimating procedure on the observed data, and found the results represented in Figure 3.

TABLE 8

Simulation Results for illness incidence and specificity and sensitivity of three tests

		T_1		T_2		T_3	
Parameter	π	α	$1-\beta$	α	$1-\beta$	α	$1-\beta$
True value	**0.1**	**0.05**	**0.9**	**0.1**	**0.85**	**0.2**	**0.8**
Min.	0.0699	0.0234	0.7346	0.0670	0.6728	0.1597	0.6659
1st Qu.	0.0903	0.0430	0.8601	0.0908	0.8204	0.1907	0.7650
Median	0.0998	0.0491	0.9140	0.0991	0.8591	0.2031	0.8071
Mean	**0.1001**	**0.0495**	**0.9092**	**0.0994**	**0.8562**	**0.2032**	**0.7993**
3rd Qu.	0.1077	0.0562	0.9785	0.1080	0.8942	0.2125	0.8353
Max.	0.1505	0.0753	1.0000	0.1388	0.9951	0.2376	0.9461

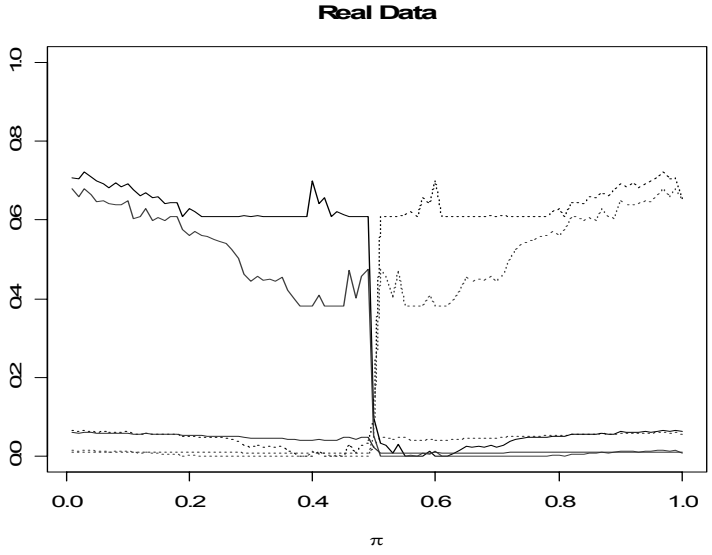

Figure 3 – Results of the optim function with L-BFGS-B minimization algorithm on the observed data. Dotted lines are type I errors and solid lines represent the powers of each diagnostic test. IDEXX = black lines; Pourquier = red lines, and FdC = blue lines.

Figure 3 indicates that the illness incidence is lower than 0.5, therefore we will take as the estimates of the parameters the average of the optim results corresponding to values of π smaller than 0.5. These results are shown in Table 9.

TABLE 9

Simulation estimates settings for illness incidence and specificity and sensitivity of three tests

		IDEXX		Pourquier		FdC	
Parameter	π	α	$1-\beta$	α	$1-\beta$	α	$1-\beta$
Av. estimate	0.1601	0.0376	0.6398	0.0045	0.5256	0.0088	0.0503

From these results, the illness incidence is equal to 16%, and the FdC test is clearly the worst, since its power estimate is only 5% (remember that, according to FdC, only 20 cattle out of 1307 were affected by the illness). As regards the comparison between IDEXX and Pourquier, no clear indication is given from the

observed data: IDEXX is more sensitive than Pourquier, but less specific. In order to give a final answer to the problem, it really depends on what consequences the Q - fever has on the cattle: if the illness affecting the cattle is not dangerous to humans, then probably Pourquier is to be preferred, since it shows a small chances of commiting a type I error. On the other hand, if the illness is dangerous for humans, then IDEXX is to be preferred, since it is more powerful in detecting the illness and can better prevent potential dangers in human alimentation.

5. CONCLUSION

At the very end of this study, we finally obtained some information from experts on the possible range of illness incidence in the population and on sensitivities and specificities of the three tests. The experts stated that a possible range for π is [0.14, 0.40], and $Sp_{IDEXX} = 0.95$, $Se_{IDEXX} = 0.92$; $Sp_{Pourquier} = 1$, $Se_{IDEXX} = 0.982$; $Sp_{dCF} = 1$, $Se_{FdC} = 0.48$ (the information on sensitivities and specificities was taken from official declarations made by pharmaceutical companies that produced the tests). The results we have obtained substantially confirm the feelings of experts on π and the specificities of the tests declared by the pharmaceutical companies. On the contrary, the sensitivities declared by the pharmaceutical companies differ considerably from our results.

The expected number of positive outcomes if the true illness incidence was $\pi = 0.16$ and sensitivities and specificities were as expected by the pharmaceutical companies, would be 247, 205, and 100 for IDEXX, Pourquier, and FdC respectively. However, the estimates of π obtained from data and the declared specificities and sensitivities would be equal to 0.0964, 0.0857 and 0.0318 for IDEXX, Pourquier, and FdC respectively. A 95% confidence interval for π obtained from the `optim` results is [0.1276, 0.2199].

The application of this approach to other kind of real situations and in different fields needs further investigation in order to explore the possible flexibility of the proposed solution.

Department of Management and Engineering DARIO BASSO
University of Padova, Italy

Istituto Zooprofilattico Sperimentale delle Venezie KATIA CAPELLO

Department of Management and Engineering LIVIO CORAIN
University of Padova, Italy LUIGI SALMASO

REFERENCES

P.A. BEARE, J.E. SAMUEL, D. HOWE, K. VIRTANEVA, S.F. PORCELLA, R.A. HEINZEN (2006). *Genetic diversity of the Q fever agent, Coxiella burnetii, assessed by microarray-based whole-genome comparisons*. "Journal of Bacteriology", 188, 7, pp. 2309-2324.

R.H. BYRD, P. LU, J. NOCEDAL, C. ZHU (1995). *A limited memory algorithm for bound constrained optimization*, "SIAM J. Scientific Computing", 16, pp. 1190-1208.

R. GUATTEO, F. BEAUDEAU, A. JOLY, H. SEEGERS (2007). *Coxiella burnetii shedding by dairy cows*. "Veterinary Research", 38, 6, pp. 849-60.

S.L. HUI, S.D. WALTER (2001). *Estimating the error rates of diagnostic tests*, "Biometrics", 36, pp. 167-71.

W.O. JOHNSON, J.L. GASTWIRTH, L.M. PEARSON (2001). *Screening without a "Gold Standard": The Hui-Walter Paradigm Revisited*, "American Journal of Epidemiology", 153, 9, pp. 921-924.

L. JOSEPH, T.W. GYORKOS, L. COUPAL (1995). *Bayesian estimation of disease prevalence and parameters for diagnostic tests in the absence of a gold standard*. "American Journal of Epidemiology", 141, pp. 263-72.

T.J. MARRIE (EDITOR) (1990). *Q Fever*. CRC Press.

C. MEHTA, N. PATEL (1998). *Exact Inference for Categorical Data*. In: P. Armitage, T. Coltin (Eds.). "Encyclopedia of Biostatistics" (Vols. 1-6). NY: John Wiley.

A. NEATH, F.J. SAMANIEGO (1997). *On the efficacy of Bayesian inference for nonidentifiable models*. "American Statistician", 51, pp. 225-32.

R DEVELOPMENT CORE TEAM (2008). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Wien, Austria, http://www.R-project.org.

SUMMARY

Evaluating sensitivity and specificity of three diagnostic tests when the "gold standard" is unavailable, with application to the cattle Q fever in small ruminants case study

In the context diagnostic tests may be assessed through indicators of diagnosis reliability called specificity and sensitivity. In practice, these indicators can be estimated only if a "gold standard" test is available, meaning that its diagnosis is the most reliable one available as to the prevalence of an illness in a population.

Starting from a real case study related to cattle Q fever disease in small ruminants, the aim of this work is to determine which of the three examined diagnostic tests is the best, taking into account the fact that there is neither any a priori information on the sensitivity and specificity of the three tests, nor a reference "gold standard" diagnostic test. Moreover, the incidence of the disease in the reference population is unknown.

Our approach, which is mainly descriptive in nature, derived estimates of sensitivity and specificity of the diagnostic tests from incidence of the disease. The estimates are obtained by minimizing the least squares and a performed simulation study shows that on average the method provides unbiased estimates of unknown parameters. The application of the method to a real case study make it possible to establish a hierarchy among the three diagnostic tests in question.