ELSEVIER

Preventive Veterinary Medicine 45 (2000) 61–81

PREVENTIVE VETERINARY MEDICINE

www.elsevier.nl/locate/prevetmed

Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown

Claes Enøe[a,*], Marios P. Georgiadis[b], Wesley O. Johnson[c]

[a]*Department of Animal Science and Animal Health, Division of Ethology and Health, Royal Veterinary and Agricultural University, DK-1870, Frederiksberg C, Denmark*

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

[c]*Division of Statistics, University of California, Davis, CA 95616, USA*

Abstract

The performance of a new diagnostic test is frequently evaluated by comparison to a perfect reference test (i.e. a gold standard). In many instances, however, a reference test is less than perfect. In this paper, we review methods for estimation of the accuracy of a diagnostic test when an imperfect reference test with known classification errors is available. Furthermore, we focus our presentation on available methods of estimation of test characteristics when the sensitivity and specificity of both tests are unknown. We present some of the available statistical methods for estimation of the accuracy of diagnostic tests when a reference test does not exist (including maximum likelihood estimation and Bayesian inference). We illustrate the application of the described methods using data from an evaluation of a nested polymerase chain reaction and microscopic examination of kidney imprints for detection of *Nucleospora salmonis* in rainbow trout.

Keywords: Diagnostic tests; Sensitivity; Specificity; Maximum likelihood; Bayesian approach; Latent data

* Corresponding author. Tel.: +45-35-28-30-10; fax: +45-35-28-30-22.
E-mail address: cle@kvl.dk (C. Enøe)

0167-5877/00/$ – see front matter
PII: S0167-5877(00)00117-3

1. Introduction

The sensitivity and specificity of a test are usually determined by comparison with a reference test (often referred to as a ''gold standard''), which is supposed to determine the true disease state of the animals unambiguously (Office International des Epizooties, 1996; Greiner and Gardner, 2000). When a gold standard is available, sensitivity and specificity can be estimated directly (Kraemer, 1992). The true disease state, however, is rarely known in practice, because perfect test results may be difficult or impossible to obtain (Tyler and Cullor, 1989).

If classification errors in the reference test are ignored, serious bias may be introduced in the assessment of the accuracies of the new test (Staquet et al., 1981; Valenstein, 1990). However, when the error probabilities of the reference test are known, it is possible to obtain unbiased estimates of the accuracies of the test in question (Gart and Buck, 1966; Staquet et al., 1981). The estimation is based on the assumption that the classification errors in the reference and the new test are independent, conditional on the true disease state. However, estimation is possible even when conditional independence is not assumed (Thibodeau, 1981).

Hui and Walter (1980) considered the case where two tests (both with unknown sensitivity and specificity) were simultaneously applied to individuals from two populations with different prevalences of disease. They showed that sensitivity and specificity of both tests (assuming conditional independence) — as well as true prevalence in both populations — could be estimated by maximum likelihood (ML). A thorough discussion of the applicability of the method in other settings (such as the case with one-population and three or more tests) is given by Walter and Irwig (1988). Bayesian methodology has also been used for the model proposed by Hui and Walter and ones like it (Joseph et al., 1995; Johnson et al., 2000). Computations are accomplished by Gibbs sampling (Gelfand and Smith, 1990). Hui and Zhou (1998) presented an overview of available methods for diagnostic test evaluation with an emphasis on methodology for estimation of sensitivity and specificity (without need for the assumption of conditional independence).

Although not widely adopted, the Hui and Walter model (and models similar to it) has been applied in statistical research (McClish and Quade, 1985; Vacek, 1985; Ashton and Moeschberger, 1988; Walter and Irwig, 1988; Qu et al., 1996; Sinclair and Gastwirth, 1996; Weng, 1996; Torrance-Rynard and Walter, 1997; Johnson and Pearson, 1999; Johnson et al., 2000) and in human medical science (van Ulsen et al., 1986; Shaw et al., 1987; Walter et al., 1991; de Bock et al., 1994; Faraone and Tsuang, 1994; Faraone et al., 1996; Line et al., 1997; McDermott et al., 1997; Mahoney et al., 1998; Rybicki et al., 1998). The methods have been introduced only recently in the evaluation of diagnostic tests used for detection of animal disease (Spangler et al., 1992; Agger et al., 1997; Chriel and Willeberg, 1997; Enøe et al., 1997; Sørensen et al., 1997; Willeberg et al., 1997; Georgiadis et al., 1998; Singer et al., 1998).

In this paper, we describe methods of estimating the sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. We present methods beginning with the case where an imperfect reference test is available, and we ultimately give special emphasis to the model and methods described by Hui and Walter (1980). Methods are illustrated using data from Georgiadis et al. (1998).

2. Estimating sensitivity, specificity and true prevalence when the true disease state is unknown

2.1. Reference test with known sensitivity and specificity

Consider the case where the true disease state of animals cannot be determined perfectly, but where the sensitivity (Se_R) and the specificity (Sp_R) of a reference test are presumed known. When each individual animal in a random sample of size n is tested by a new diagnostic test and a reference test, four outcomes are possible: both tests positive (T_1+, T_2+; denoted a); one test positive and one negative (T_1+, T_2-; denoted b) and (T_1-, T_2+; denoted c); both tests negative (T_1-, T_2-; denoted d). The resulting cross-classification for the n animals in population i is shown in Table 1. The data presented in the 2×2 tables are assumed to have a multinomial distribution throughout this discussion.

Gart and Buck (1966) considered this case and provided estimates of the sensitivity (Se_N) and the specificity (Sp_N) of the new test as well as the true prevalence (P) in the sample. Staquet et al. (1981) used an equivalent approach, but with simpler expressions. Using the notation in Table 1, Se_N, Sp_N and P, according to Staquet et al. (1981), are estimated as

$$\widehat{Se}_N = \frac{gSp_R - b}{n(Sp_R - 1) + e}, \qquad \widehat{Sp}_N = \frac{hSe_R - c}{nSe_R - e}, \qquad \hat{P} = \frac{n(Sp_R - 1) + e}{n(Se_R + Sp_R - 1)}.$$

Formulas for standard errors (S.E.s) of these estimates are provided by Gart and Buck (1966). In this situation, parameters can be estimated because both the data and the parameter space are three-dimensional. Thus, the estimates above are simply solutions to three equations with three unknowns.

Staquet et al. (1981) also considered the case where both tests have perfect specificity but unknown sensitivities, and the case where the reference test has perfect specificity but unknown sensitivity. In both instances, Se_N can be estimated.

All of the methods discussed assume that the two tests are conditionally independent (i.e. knowledge of the outcome of the reference test gives no information about the outcome of the new test, conditional on the true disease state). This assumption is not satisfied in many situations, especially when the two tests have the same basis (Gardner et al., 2000). Implications of this are discussed in Section 4.

Table 1
Test results stated as positive (T+) or negative (T−), cross-classified in a 2×2 tables, according to the status of each individual animal tested by test 1 and 2 in population i (i=1, 2)

Test 1	Test 2		
	T+	T−	
T+	a_i	b_i	g_i
T−	c_i	d_i	h_i
	e_i	f_i	n_i

2.2. Both tests with unknown sensitivity and specificity

Although the assumption of known sensitivity and specificity for the reference test is commonly made, it is rarely correct. Usually there is some uncertainty about the true sensitivity and specificity values, even when considerable data have been collected. When previously collected data are used to estimate sensitivity and specificity of a new test, the inherent variability in these estimates should be taken into account (Gastwirth, 1987).

When one needs to evaluate the performance of two tests (both with unknown accuracies), the observed data are three-dimensional, while there are five unknown parameters: two sensitivities, two specificities and the prevalence. Therefore, additional data/information are required to make inferences.

The model introduced by Hui and Walter (1980) allowed the estimation of sensitivity and specificity of two tests, based on their cross-classified results, when applied to individuals from two populations with different disease prevalences. In addition to the assumption of conditional independence between the two tests, Hui and Walter also assumed that the accuracy of each test was the same in both populations.

Hui and Walter showed that when these assumptions hold, estimates of sensitivity and specificity of both tests, and of the prevalences in both populations can be obtained by ML. Data for this method can be obtained by sampling from two (or more) different populations (Georgiadis et al., 1998). Alternatively, two or more samples could be obtained from the same population by subdivision according to covariate information (such as health status or herd-size) (Enøe et al., 1997; Willeberg et al., 1997) as long as the resulting subpopulations have different prevalences.

For each population test results are cross-classified in a 2×2 table according to the status of each individual tested (as shown in Table 1). Each 2×2 table provides three degrees of freedom (d.f.) for estimation. When two populations are available, there are 6 d.f. for estimation. The number of parameters of interest is two for each of the two tests (sensitivities and specificities) and one for each of the two populations (prevalences). For this particular situation, explicit formulas for the maximum likelihood estimates (MLEs) are available (Hui and Walter, 1980). In this case, both the data and the parameter space are six-dimensional and the estimates are solutions to six equations with six unknowns.

Six parameters must be estimated from the observed data: sensitivity of test 1 (Se_1), sensitivity of test 2 (Se_2), specificity of test 1 (Sp_1), specificity of test 2 (Sp_2), prevalence in population 1 (P_1) and prevalence in population 2 (P_2). Formulas for the MLEs were given by Hui and Walter (1980):

$$\widehat{\mathrm{Se}}_1 = \frac{(g_1e_2 - e_1g_2)/n_1n_2 + a_2/n_2 - a_1/n_1 + F}{2(e_2/n_2 - e_1/n_1)},$$

$$\widehat{\mathrm{Se}}_2 = \frac{(g_2e_1 - e_2g_1)/n_1n_2 + a_2/n_2 - a_1/n_1 + F}{2(g_2/n_2 - g_1/n_1)},$$

$$\widehat{\mathrm{Sp}}_1 = \frac{(f_1h_2 - h_1f_2)/n_1n_2 + d_1/n_1 - d_2/n_2 + F}{2(e_2/n_2 - e_1/n_1)},$$

$$\widehat{\mathrm{Sp}}_2 = \frac{(f_2h_1 - h_2f_1)/n_1n_2 + d_1/n_1 - d_2/n_2 + F}{2(g_2/n_2 - g_1/n_1)},$$

$$\hat{P}_1 = 0.5 - \left\{ \frac{[(g_1/n_1)(e_1/n_1 - e_2/n_2) + (e_1/n_1)(g_1/n_1 - g_2/n_2) + a_2/n_2 - a_1/n_1]}{2F} \right\},$$

$$\hat{P}_2 = 0.5 - \left\{ \frac{[(g_2/n_2)(e_1/n_1 - e_2/n_2) + (e_2/n_2)(g_1/n_1 - g_2/n_2) + a_2/n_2 - a_1/n_1]}{2F} \right\},$$

where

$$F = \pm \left[\left(\frac{g_1 e_2 - g_2 e_1}{n_1 n_2} + \frac{a_1}{n_1} - \frac{a_2}{n_2} \right)^2 - 4 \left(\frac{g_1}{n_1} - \frac{g_2}{n_2} \right) \frac{a_1 e_2 - a_2 e_1}{n_1 n_2} \right]^{0.5}.$$

Two sets of solutions are provided by these equations (depending on the sign of F), only one of which gives reasonable estimates, assuming that $\mathrm{Se} + \mathrm{Sp} > 1$. For a more detailed discussion, see Hui and Walter (1980). Note, however, that in some situations there are no explicit solutions to the above equations (in which case it is necessary to apply an iterative procedure to obtain the MLEs). This is also the case when the model is extended to include data from more than two populations or tests. Moreover, even when explicit formulas for the MLEs are available, the formulas for the S.E.s are complicated. Therefore, we advocate the use of standard software (described below) for obtaining MLEs and associated S.E.s. We only discuss the two-population case. The extension to the multiple-population case is straightforward (Johnson et al., 2000).

3. Methods of estimation and computational techniques for the Hui and Walter model

ML estimates are a set of parameter estimates that were most ''likely'' to have generated the observed data and are obtained by maximizing the likelihood function (Tanner, 1996). Variances are obtained by calculating the Fisher Observed Information matrix and inverting it (Gelman et al., 1995, p. 100). The square roots of the diagonals of this matrix are the corresponding S.E.s. ML estimates have many optimal properties when sample sizes are large. They are asymptotically unbiased and efficient (i.e. large sample variances are relatively small, and asymptotically normal).

ML estimates and S.E.s generally are obtained through the Newton–Raphson (NR) technique (Tanner, 1996). The NR algorithm may fail to converge if the data are sparse or if the observed frequencies of some test result combinations are zero. In such instances, we recommend the use of data-augmentation approaches, which are better suited to this problem. One such approach is the Expectation–Maximization (EM) algorithm (Dempster et al., 1977), which takes advantage of a natural ''latent'' or ''missing-data'' structure for screening problems without a gold standard (namely, the missing information about true disease state). The computational advantage of data-augmentation methods derives from the simplicity of the augmented data likelihood, while the likelihood function based on the observed data is complicated. Details for the NR and EM algorithms are discussed in Sections 3.1 and 3.2, respectively.

Inferences based on the ML method rely on the assumption of a large sample size (n). Confidence intervals (CIs) obtained according to the formula $\hat{\theta} \pm z_{\alpha/2}$ S.E. (for a generic

parameter θ) will be valid only if n is sufficiently large. As a general rule of thumb, first calculate $\hat{\theta} \pm k$ S.E., where k is an integer constant. If the intervals obtained with k=3 exclude 0 and 1, experience suggests that 95% CI will be reasonably accurate (Johnson and Gastwirth, 1991; Johnson and Gastwirth, 2000; Johnson and Pearson, 1999; Johnson et al., 2000). With k=5, even 99% CIs should be highly accurate.

Alternatively, the Bayesian approach (Lee, 1989; Press, 1989; Gelman et al., 1995; Tanner, 1996) can be used to model a priori scientific knowledge about unknown parameters and to combine this information with the information contained in the likelihood based on observed data. Furthermore, it is possible to make inferences that are free from large sample theory assumptions. Virtually any inference can be made without having to resort to potentially difficult mathematical derivations; probability statements can be made about the parameters. The method is straightforward to implement by taking advantage of the missing-data structure and is accomplished by Gibbs sampling (Joseph et al., 1995; Mendoza-Blanco et al., 1996; Tanner, 1996; Johnson et al., 2000). Details of the Gibbs sampler are given in Section 3.3 and in Appendix A.

Bayesian statistical inferences require the modeling of all uncertainty with probability. Scientific information about the unknown accuracies and prevalences is thus incorporated into the model through the specification of a joint prior probability distribution on all parameters of interest. The information incorporated into this distribution should be elicited as scientific input and would ideally be based on previous studies that are similar to the current one. Posterior inferences are obtained by combining the actual likelihood with an assumed prior distribution. The Gibbs sampler can be used to obtain numerical approximations to exact posterior inferences. The previously considered situation, where sensitivity and specificity of the reference test were assumed known, could be regarded as a special case of Bayesian analysis, where there was no uncertainty about these two parameters. A more realistic Bayesian analysis simply attaches probabilistic measures of uncertainty to all the parameters of interest. Details about the specification of the prior distributions are given in Section 3.3.[1]

3.1. Maximum likelihood: Newton–Raphson

For each of the four cells in the 2×2 table (Table 1), the likelihood contributions are determined as the probability of observing data in each cell conditional on the parameters, raised to the power of the observed frequency for that cell. The likelihood contributions for the four cells corresponding to population i may be written as

$$(\mathrm{T_1}+,\mathrm{T_2}+): \quad [P_i\mathrm{Se_1Se_2} + (1-P_i)(1-\mathrm{Sp_1})(1-\mathrm{Sp_2})]^{a_i},$$

$$(\mathrm{T_1}+,\mathrm{T_2}-): \quad [P_i\mathrm{Se_1}(1-\mathrm{Se_2}) + (1-P_i)(1-\mathrm{Sp_1})\mathrm{Sp_2}]^{b_i},$$

$$(\mathrm{T_1}-,\mathrm{T_2}+): \quad [P_i(1-\mathrm{Se_1})\mathrm{Se_2} + (1-P_i)\mathrm{Sp_1}(1-\mathrm{Sp_2})]^{c_i},$$

$$(\mathrm{T_1}-,\mathrm{T_2}-): \quad [P_i(1-\mathrm{Se_1})(1-\mathrm{Se_2}) + (1-P_i)\mathrm{Sp_1Sp_2}]^{d_i}.$$

[1] Some readers may prefer to skip Sections 3.1–3.3, which are somewhat technical, during the first reading of the paper.

The likelihood function for the unknown parameters based on the overall data is obtained by multiplying the likelihood contributions across the i populations, regardless of the number of populations sampled (provided they are sampled independently). The above formulas were derived using the conditional independence assumption, which is evident by noticing that the joint probabilities of test results, conditional on disease status ((D+)=truly diseased; (D−)=truly non-diseased), are obtained by multiplying the respective individual test probabilities, e.g. $\Pr(T_1+, T_2+ \mid D+) = Se_1Se_2$, etc.

After the likelihood function has been obtained, MLEs for the sensitivities, specificities and prevalences (as well as the asymptotic variance–covariance matrix) can be obtained by maximization of the likelihood by iterative methods. One approach is the NR algorithm (Tanner, 1996) which is available in statistical software packages, e.g. BMDP-LE (BMDP statistical software, Los Angeles, CA), S-Plus (Mathsoft, Seattle, WA) and SAS (SAS Institute, Cary, NC).

3.2. Maximum likelihood: EM algorithm

Both the EM algorithm and the Gibbs sampler make use of the idea that it is possible to allocate individuals from each population into a truly diseased or a truly non-diseased but unobservable (latent) class. In this case, the observed number of individual test results in each of the four cells in the 2×2 table is considered to be the sum of those that are truly diseased and those that are truly non-diseased. Therefore, the 2×2 table for each population is considered to result from collapsing the two tables of latent data for this population.

The EM algorithm is an iterative numerical technique for finding the MLEs. At a given iteration, the E step simply involves imputing a current surrogate for the missing disease status of all individual samples. The surrogate for the missing count (say, z_{11i} of truly diseased individuals given the test result (T_1+, T_2+)) is obtained as the conditional expectation of z_{11i} given the observed count a_i in Table 1. The corresponding distribution is binomial (a_{11i}, $\Pr_{11i}$), with $\Pr_{11i}=P_iSe_1Se_2/\{P_iSe_1Se_2+(1-P_i)(1-Sp_1)(1-Sp_2)\}$, where $\Pr_{11i}$ is the probability of disease given the test result (T_1+, T_2+) and is obtained using Bayes' theorem. Other probabilities are calculated similarly. The imputed data are used to create the ''augmented-data likelihood'' (ADL) with eight cell contributions from each population. The ADL function for the Hui and Walter model simplifies to the product of six binomial-like contributions. The ADL is maximized in the M step. The process begins with starting values for the parameters, which would be one's best available estimate for those values. Then, the E and M steps are iterated until convergence. The algorithm is described explicitly for the Hui and Walter model in Singer et al. (1998).

Once the MLEs have been obtained, it is necessary to obtain the large sample variance–covariance matrix. The simplest approach is to use the MLEs obtained in the EM algorithm as starting values in a standard computer program that performs the NR technique; the asymptotic variance–covariance matrix is standard output from this algorithm. With such starting values, the NR algorithm should converge in one iteration. In the event that it does not, the program should be modified so that the algorithm is forced to do so (assuming that one is confident that the EM algorithm

has converged to the global maximum). Alternatively, one could use the formulas given in Hui and Walter (1980). A third alternative would be to write software to perform the EM-related technique of Meng and Rubin (1991), as described in Tanner (1996, pp. 78–79).

3.3. Bayesian approach: specification of priors and Gibbs sampling

In Bayesian analysis involving proportions, we often specify the prior distributions by modeling subjective probability about the unknown parameters through the use of beta distributions (Johnson and Gastwirth, 1991; Joseph et al., 1995; Mendoza-Blanco et al., 1996; Bedrick et al., 1997). Additionally, the use of beta priors greatly simplifies calculations and beta distributions are quite flexible (because manipulation of their two parameters can yield a large array of potential shapes).

To construct a beta prior distribution for a particular parameter, one should seek expert opinion about two or three characteristics of each of the parameters of interest. First, elicit the most probable value or best guess (θ_0), which may be an actual estimate based on previous data. Then, determine a value (θ_L) for which the experimenter is $(1-\gamma/2)$ certain that the parameter will be larger, and/or a second value (θ_U) for which the experimenter is $(1-\gamma/2)$ certain that the parameter will be smaller. With $\gamma=0.1$, the experimenter is 95% sure that the parameter of interest is smaller than θ_U and 95% certain that it is larger than θ_L. These values are then the 95th and 5th percentiles of the prior distribution, respectively. The ''best guess'' can be selected to be the mode if a unimodal prior distribution is sought.

In the case of beta (a, b) priors, the mode of the distribution is given by the formula $\theta_0=(a-1)/(a+b-2)$ when $a>1$ and zero otherwise. Solving this equation, $a=(1+\theta_0(b-2))/(1-\theta_0)$. So for a given guess ($\theta_0$) and a given value of b, a is determined. Once a pair (a, b) has been obtained, software like S-plus can be used to determine whether the appropriate percentiles of the specified beta (a, b) distribution are θ_L and θ_U. If this constraint is not satisfied, another b is selected and the appropriate a is calculated. The process is repeated until a beta (a, b) distribution that satisfies the constraints posed by the prior specification is identified. The distribution is then presented graphically to the subject matter experts for verification; if not satisfactory, the process is repeated with another type of distribution. If $a<1$ is appropriate, then since the formula for the mode cannot be used, one can equate $\theta_0=a/(a+b)$ (which is the mean of the beta distribution) and proceed as described above.

All Bayesian analyses should include a sensitivity analysis, which involves consideration of a non-informative prior as well as a few perturbations of the given prior. If the corresponding posterior inferences change drastically, this should be reported. Some may prefer to simply use a non-informative prior, e.g. beta (1, 1), if a subject matter expert is not available. For more complete account of Bayesian screening in related settings, see Johnson and Gastwirth (1991), Gastwirth et al. (1991), Geisser and Johnson (1992), Mendoza-Blanco et al. (1996), Johnson and Pearson (1999) and Johnson et al. (2000).

After eliciting prior distributions for each of the six unknown parameters (the sensitivities and specificities of T_1 and T_2 and the prevalence in each of the

two populations), joint independence of these elicitations is assumed and the joint posterior distribution is obtained. Bayesian inferences are based on the combined input from the likelihood and the joint prior. The product of the likelihood function and the joint prior density is the conditional probability density of the parameters given the observed data (up to the constant of integration because the resulting posterior distribution must integrate to 1). The constant of integration will not be discussed further, because it plays no role. Ultimately, one can obtain (by integration) the marginal posteriors for each of the six parameters of interest. If the posterior is highly concentrated about a particular value, the implication is that there is a great deal of a posteriori certainty about that parameter. The median of the posterior is used for point estimates. The standard deviation (S.D.) of a marginal posterior distribution is a measure of the quality of the point estimate. Finally, intervals that have $1-\alpha$ probability content are obtained, using each of the marginal posteriors. So, with $\alpha=0.05$, such an interval can be interpreted to mean that one is 95% sure that the corresponding parameter is in that interval.

As previously indicated, the ADL is simple and — when combined with the prior — results in independent beta distributions. The latent data and observations from the joint posterior are simulated in the Bayesian approach by an iterative Markov chain Monte-Carlo technique using the Gibbs sampler (Gelfand and Smith, 1990; Gelman et al., 1995; Andersen, 1997). The simulated sample is then used to approximate the actual posterior distribution. For example, the procedure will generate many values from the posterior (e.g. Se_1). A plot of these values can then be regarded as a numerical approximation to the corresponding probability density function. The median of the Monte-Carlo sample is a numerical approximation to the actual value of the posterior that has half of the probability to the left and half to the right and this is our point estimate. The Bayesian intervals for each parameter are obtained by considering the percentiles of the Monte-Carlo sample.

The Gibbs sampling approach proceeds as follows with technical details presented in Appendix A. The approach is iterative with two steps as with the EM algorithm. Initially, starting values for the parameters are selected. These parameter values can be one's best guesses based on the prior distributions, or they can simply be values that are sampled from the prior distributions. Sampling beta distributions is routine in packages such as S-plus, Gauss (Aptech, Kent, Washington) and SAS. With specified starting values, one proceeds to sample values for the missing (latent) z_{ijk}'s. These distributions are identical to those described in Section 3.2.

Using this approach, a set of z_{ijk}'s is sampled from the respective binomial distributions. These current values for the missing data are then substituted to create the current ADL, which in turn is combined with the prior resulting in independent beta posteriors for the six parameters. These distributions are then sampled, the new values used to re-sample the conditional binomial distributions for the missing z_{ijk}'s, and the process is continued to convergence. Convergence is assessed by considering plots of running means of the parameters of interest and is determined when these plots stabilize after a certain number of samples. Early samples in the iterative process are often discarded; this phase frequently is termed the "burn-in" phase. Details are given in Johnson et al. (2000).

3.4. Bayesian method for the one-population model

Joseph et al. (1995) considered the Gart and Buck (1966) situation with two tests and one sample of data without a gold standard. Thus, there were five unknown parameters and only three-dimensional data (Chriel and Willeberg, 1997). With only one population, the likelihood lacks identifiability; many distinct values of the parameters can result in the same value for the likelihood. But, with a proper prior (based on empiric or scientific knowledge) on the five parameters, this problem no longer occurs. The disadvantage, however, is that as the sample size increases, the posteriors will not necessarily focus on the ''true values'' of the parameters (Andersen, 1997; Johnson et al., 2000). However, from a purely Bayesian perspective, one is simply modeling uncertainty about parameters and does not expect ultimately to know the precise values unless the quality of the prior information is extremely precise. Certainly, such inferences are more realistic than simply substituting values for the sensitivity and specificity of the reference test as if they were truly known. The Bayesian approach allows one to be uncertain about these values and to still make valid inferences. We refer the reader to Joseph et al. (1995) and to Johnson et al. (2000) for the details of the implementation of this procedure with this kind of data.

4. Assumptions

The methods presented in this paper are based on several assumptions that — if not taken into careful consideration — can seriously invalidate the results.

In the models proposed by Gart and Buck (1966), Staquet et al. (1981) and Hui and Walter (1980), the two tests are assumed to be conditionally independent. The assumption of conditional independence implies that given that an animal is diseased (or not), the probability of positive (or negative) outcomes for T_1 is the same regardless of a known outcome for T_2. Thibodeau (1981) considered the case where the tests are conditionally dependent by modeling the correlation between tests (thus extending the results of Gart and Buck (1966)). Vacek (1985) showed that if conditional dependence exists between two tests, then classification errors for both tests will be substantially underestimated when using the Hui and Walter model (see also Brenner, 1996; Torrance-Rynard and Walter, 1997).

For the Hui and Walter model, an additional major assumption is that the accuracy of both tests remains constant over different populations. When this assumption fails, bias is introduced. A general discussion is provided by Choi (1997) and Brenner and Gefeller (1997). The reason for this assumption is a practical one: allowing the accuracies to be different from one population to the next would add four additional parameters for each population considered, while there are only three additional d.f. associated with each new population in the two-test case.

To illustrate some potential difficulties encountered with invalid assumptions, we generated data that are nearly perfectly consistent with certain parameter values. First, we consider data that are generated based on correlated tests but with constant accuracy across populations. We defined the correlation between two tests based on a known diseased sample to be $\rho_{D+} = \mathrm{corr}\,(T_1, T_2|D+)$ and the correlation between two test outcomes based on a sample that is known to be not diseased

as $\rho_{D-} = \text{corr}(T_1, T_2|D-)$. We generated data by choosing values that equal known expected values for the eight cell probabilities with given values for the accuracies, prevalences and correlations. The sample sizes were chosen to be large enough so that rounding expected values to integers would have little effect on the estimates. For example, $a_1 = n_1\{P_1 \Pr(T_1+, T_2 + |D+) + (1 - P_1)\Pr(T_1+, T_2 + |D-)\}$, where $\Pr(T_1+, T_2 + |D+) = Se_1Se_2 + \rho_{D+}\{Se_1(1 - Se_1)Se_2(1 - Se_2)\}^{1/2}$, etc.

Data were generated with both sensitivities equal to 0.95, both specificities equal to 0.70, and both correlations equal to 0.1, 0.5 and 0.9, respectively. The prevalences were 0.05 and 0.02 for the two hypothetical populations. Based on data generated with both correlations equal to 0.1 (a_1=181, b_1=c_1=218, d_1=583, a_2=206, b_2=c_2=302, d_2=813), where a_i, b_i, c_i and d_i refer to the cells of the two 2×2 tables, the resulting MLEs were 0.958 for the sensitivities, 0.732 for the specificities and 0.093 and 0.065, respectively, for the prevalences. With correlation of 0.5, the data were: a_1=273, b_1=c_1=119, d_1=668, a_2=361, b_2=c_2=178, d_2=1005. The corresponding MLEs for the sensitivities were 0.98, for the specificities were 0.855 and for the prevalences were 0.224 and 0.201, respectively. With data based on a correlation of 0.9, the NR algorithm did not converge. Thus, even with a small correlation of 0.1 between the tests, estimates can be biased and with a correlation of 0.5, the MLEs are strongly biased.

As a second illustration, we generated data under the condition of conditional independence, but without the assumption of constant accuracy across populations. Consider the situation with equal sensitivities of 0.95 for the two tests when used in population 1 and with equal sensitivities of 0.85 when used in population 2. Assume the specificities are the same and equal to 0.85 for both populations. Finally, assume a prevalence of 0.05 for population 1 and 0.02 for population 2. Based on the data (a_1=133, b_1=c_1=247, d_1=1373, a_2=73, b_2=c_2=255, d_2=1417), the resulting MLEs for the sensitivities were both 1, for the specificities were 0.84, and for the prevalences were 0.044 and 0.014, respectively. A second situation considered had sensitivities based on tests in population 1 equal to 0.95, and based on tests in population 2 equal to 0.65, specificities all set equal to 0.85 and the same prevalences as before. Based on the data (a_1=133, b_1=c_1=247, d_1=1373, a_2=61, b_2=c_2=259, d_2=1421), the estimated sensitivities were 1, estimated specificities were 0.847, and the estimated prevalences were 0.044 and 0.007, respectively. Hence, the assumption of equal accuracies is also crucial.

As previously described, the assumption of distinct prevalences is necessary to the Hui and Walter model because otherwise, the data can be collapsed into a single 2×2 table with only 3 d.f. for estimation. In our experience, MLEs exhibited little bias even when the difference in prevalences was as little as 0.01. Simulation studies could indicate whether this is a general finding or not. However, the variances of these estimates are inversely proportional to the squared difference in prevalences, and thus, CIs in this instance should be very wide. As a result, this is a built-in protection against making inappropriate inferences when this assumption is questionable.

5. Illustrations

To illustrate the reviewed methods and models, we used data from Georgiadis et al. (1998) who evaluated a nested polymerase chain reaction (PCR) test (Barlough et al.,

Table 2
Observed test results at three samplings of kidney tissue from rainbow trout, cross-classified as positive (T+) or negative (T−) for *N. salmonis* by ME and PCR

ME	PCR					
	Sample 1		Sample 2		Sample 3	
	T+	T−	T+	T−	T+	T−
T+	0	0	0	0	3	0
T−	1	99	2	30	24	3

1995) and microscopic examination (ME) of kidney imprints for detection of the microsporidian parasite *Nucleospora salmonis* in rainbow trout.

Briefly, Georgiadis et al. (1998) used the NR and the EM algorithms to assess the accuracy of the PCR test and ME using the Hui and Walter two-population model. Thus, some of the results in the present study have been previously published. However, we discuss the estimates in further detail and extend the analyses for illustrative purposes.

A rainbow-trout hatchery that was historically found infected with *N. salmonis* was identified. Between March 1995 and August 1996, this population was sampled three times. Kidney samples from each sampled fish were examined by PCR and ME. Test results are shown in Table 2.

5.1. Known reference test characteristics

Initially, the best available estimates for the Se and Sp of each of the two tests were used to estimate the respective characteristics of the other using the method of Staquet et al. (1981). The best available estimates for each parameter were chosen to be equal to the ''the most probable value'' based on knowledge about the morphology and biology of the parasite, the epidemiology of the respective disease as well as information on the PCR and ME tests.

The best available estimates for the Se and Sp of the ME were believed to be 0.55 and 0.98, respectively. The best available estimates for the Se and Sp of the PCR were 0.90 and 0.85, respectively.

Using the method of Staquet et al. (1981) and data from sample 3 in Table 2, estimates and S.E.s were $\hat{P} = 0.151(0.103)$, $\widehat{\mathrm{Se}}_{\mathrm{PCR}} = 1.0$ and $\widehat{\mathrm{Sp}}_{\mathrm{PCR}} = 0.122(0.067)$, when ME was used as reference. Using the PCR test as reference, estimates were $\hat{P} = 1.0$ and $\widehat{\mathrm{Se}}_{\mathrm{ME}} = 0.113(0.062)$, whereas $\widehat{\mathrm{Sp}}_{\mathrm{ME}}$ could not be determined from the present data by this method. Standard errors were not determined when estimates were 1 (or 0) because standard asymptotic theory for derivation of S.E.s did not apply in those cases.

5.2. Hui–Walter model

We decided to collapse results from the two samplings into a single sample.

The assumption of conditional independence was considered to be reasonable because the two tests have different bases: microscopy relies on visual observation of the parasite, while PCR is a DNA-based technique. However, the assumption of constant Se for the two populations may be tenuous, because the Se of a diagnostic test (especially when detecting an infectious agent) may depend on the prevalence and the stage of the disease it detects. The two sampled populations probably had very different prevalences (Georgiadis et al., 1998). On the other hand, Sp of each test should be similar in the two populations, because all samples were obtained from the same location (which minimizes the likelihood of differences in cross-reacting microorganisms present).

We used the formulas provided by Hui and Walter to obtain MLEs of the six parameters of interest. The estimates were $\hat{P}_1 = 0.0$, $\hat{P}_2 = 0.898$, $\widehat{\mathrm{Se}}_{\mathrm{PCR}} = 1.0$, $\widehat{\mathrm{Sp}}_{\mathrm{PCR}} = 0.977$, $\widehat{\mathrm{Se}}_{\mathrm{ME}} = 0.111$ and $\widehat{\mathrm{Sp}}_{\mathrm{ME}} = 1.0$. We then used the NR algorithm to obtain the MLEs by use of BMDP-LE (Georgiadis et al., 1998). Several choices of starting values were used to check that a global rather than local maximum was achieved. However, because of the zeros in some of the cells, the NR algorithm failed to converge to the MLEs. However, convergence was achieved by adding a small number to all eight cells of the 2×2 tables. After adding 0.185 to all cells, the estimates were $\hat{P}_1 = 0.025$, $\hat{P}_2 = 1.0$, $\widehat{\mathrm{Se}}_{\mathrm{PCR}} = 0.999$, $\widehat{\mathrm{Sp}}_{\mathrm{PCR}} = 0.998$, $\widehat{\mathrm{Se}}_{\mathrm{ME}} = 0.111$ and $\widehat{\mathrm{Sp}}_{\mathrm{ME}} = 1.0$. Unfortunately, this approach introduced bias (Georgiadis et al., 1998), and thus we proceeded to apply the aforementioned data-augmentation approaches.

5.2.1. Maximum likelihood: EM algorithm implementation

The EM algorithm was implemented using Gauss and SAS software. We obtained estimates of the parameters using the data from the two populations described above (population 1: first and second sampling, population 2: third sampling) (Georgiadis et al., 1998).

To obtain a measure of goodness-of-fit, we computed MLEs using the EM algorithm based on the data from all three samplings. By using three-population data, we increased the number of d.f. by 3. One additional parameter (the prevalence of the corresponding third population) required estimation, leaving 2 d.f. for the goodness-of-fit test. The fit of the model was evaluated by Pearson's chi-square, making use of observed and predicted numbers calculated from the MLEs. The observed chi-square statistic was 3.01 with 2 d.f. (P=0.22) (indicating a reasonable fit). This should be interpreted with caution because the large sample theory basis for the chi-square distribution is questionable for these data.

The estimates of Se and Sp for the two tests were identical to those obtained with collapsed data (Table 3). The estimated prevalences using all three samples were 0 for the first two and 0.898 for the third sample. S.E.s for the estimates can be obtained using the NR algorithm with starting values equal to the MLEs obtained by the EM algorithm. Construction of CIs for parameters is based on the asymptotic normality of the MLEs which depends on the availability of moderate to large samples. This assumption was not deemed to be appropriate for our data. S.E.s were not obtained when the corresponding estimates were 0 or 1. None of the other estimates satisfied our rule-of-thumb for determining the adequacy of the sample size for obtaining CIs ($\hat{\theta} \pm k$ S.E. excludes 0 and 1, where k=3 for 95% CI and k=5 for 99% CI).

Table 3
Estimates of sensitivity and specificity of ME and PCR for detection of *N. salmonis* in rainbow trout and disease prevalence obtained by the Hui–Walter model

	Original data[a]			Perturbed data[b]							
	0, 0, 3, 129/3, 0, 24, 3[c]			1, 0, 3, 129/3, 0, 24, 3		0, 1, 3, 129/3, 0, 24, 3		0, 0, 3, 129/3, 1, 24, 3		1, 1, 3, 129/3, 1, 24, 3	
	EM[d]	Bayes[d]	Bayes with NIP[e]	EM	Bayes	EM	Bayes	EM	Bayes	EM	Bayes
$\hat{P}_1$	0.00	0.012 (0–0.05)[f]	<0.01 (0–0.02)	0.03	0.02	<0.01	0.01	<0.01	0.01	0.04	0.02
$\hat{P}_2$	0.90	0.86 (0.7–0.96)	1.00 (0.87–1.00)	0.90	0.86	0.90	0.86	1.00	0.87	0.90	0.87
$\widehat{\mathrm{Se}}_{\mathrm{ME}}$	0.11	0.17 (0.07–0.32)	0.12 (0.04–0.26)	0.13	0.18	0.11	0.16	0.13	0.19	0.16	0.20
$\widehat{\mathrm{Se}}_{\mathrm{PCR}}$	1.00	0.94 (0.81–0.99)	0.89 (0.75–0.98)	1.00	0.94	1.00	0.94	0.87	0.91	1.00	0.91
$\widehat{\mathrm{Sp}}_{\mathrm{ME}}$	1.00	0.99 (0.97–1)	0.99 (0.97–1.00)	1.00	0.99	0.99	0.99	1.00	0.99	1.00	0.99
$\widehat{\mathrm{Sp}}_{\mathrm{PCR}}$	0.98	0.97 (0.93–0.99)	0.97 (0.94–0.99)	1.00	0.97	0.98	0.98	0.98	0.97	1.00	0.97

[a] Data from Georgiadis et al. (1998).
[b] Data modified by changing one or more cell frequencies from 0 to 1 and used in case-influence analysis.
[c] Observed frequencies in the eight cells of the two 2×2 tables (a_1, b_1, c_1, d_1/a_2, b_2, c_2, d_2).
[d] Parameter estimates obtained using the EM algorithm and Bayesian estimation (Bayes).
[e] Sensitivity analysis using non-informative priors (NIPs).
[f] Bayesian 95% probability interval.

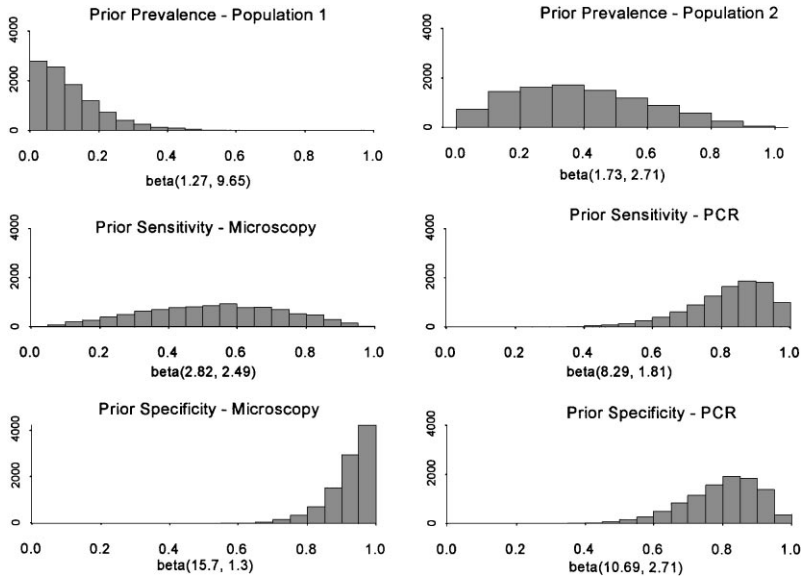

Fig. 1. Prior distributions for the six parameters of interest in the two-population and two-test problem for the Bayesian analysis of Georgiadis et al. (1998) data.

5.2.2. Bayesian approach: Gibbs sampling implementation

We require prior probability distributions for the six parameters of interest. These distributions are based on expert opinion and represent our scientific knowledge about the parameters.

The most probable value of Se_{ME} was determined to be 0.55, while we were 95% sure that it was less than 0.85. Following the outlined procedure, the beta distribution (2.82, 2.49) with mode 0.55 and 95th percentile 0.85 was identified. The most probable value of Sp_{ME} was 0.98, and it was thought to be at least 0.8 with 95% certainty. The beta distribution with a mode of 0.98 and 5th percentile of 0.8 is beta (15.7, 1.3).

The most probable value of Se_{PCR} was determined to be 0.9 and we were 95% sure that it was at least 0.6, which results in a beta (8.29, 1.81) distribution. The distribution for Sp_{PCR} was based on a best guess of 0.85 and a 5th percentile of 0.60. The corresponding distribution was beta (10.69, 2.71). Our best guess for P_1 in population 1 was 0.03, and we assumed that it should not be more than 0.30 with 95% certainty. The corresponding distribution with mode 0.03 and 95th percentile 0.30 is beta (1.27, 9.65). Finally, P_2 in population 2 was believed to be considerably higher than P_1 with the most probable value set to 0.30 and the 5th percentile set to 0.08. This resulted in a beta (1.73, 2.71) distribution. The six prior distributions are shown in Fig. 1. The Bayesian analysis was performed in S-Plus and results are given in Table 3. Posterior probability distributions are shown in Fig. 2.

We finally considered a Bayesian analysis of one-population data. We used the data from sample 3 in Table 2 and the same prior distributions that were used for the two-population analysis. The estimates and S.D.s were $\hat{P} = 0.78(0.13)$, $\widehat{Se}_{ME} = 0.19(0.08)$,

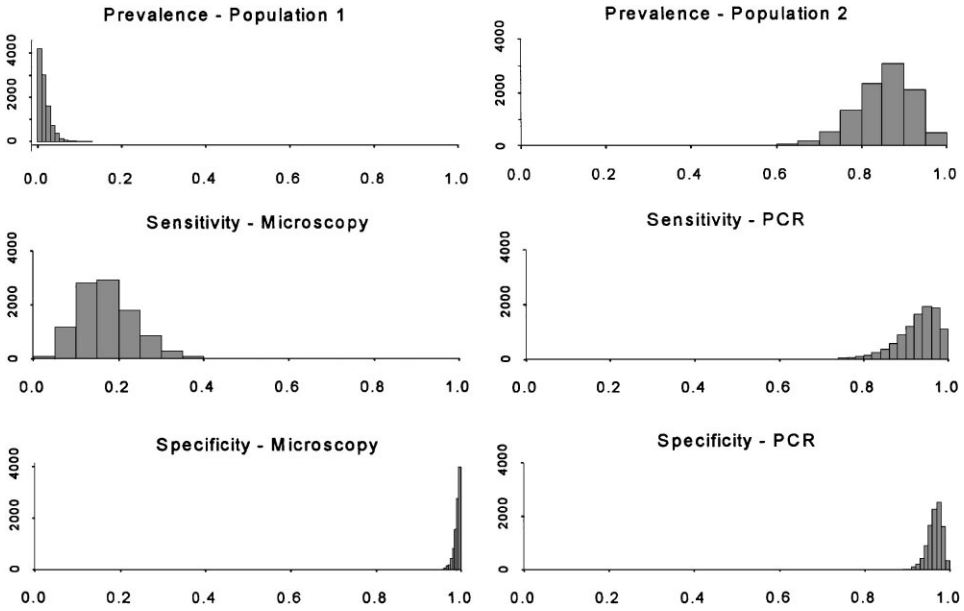

Fig. 2. Posterior distributions for the six parameters of interest in the two-population and two-test problem obtained from the Bayesian analysis of Georgiadis et al. (1998) data.

$\widehat{\mathrm{Sp}}_{\mathrm{ME}} = 0.93(0.05)$, $\widehat{\mathrm{Se}}_{\mathrm{PCR}} = 0.93(0.05)$, and $\widehat{\mathrm{Sp}}_{\mathrm{PCR}} = 0.72(0.14)$. The point estimates are consistent with those presented in Table 3.

5.3. Comparisons

The point estimates of the six parameters for both methods were similar. The Bayesian probability intervals for all the parameters, except for the two specificities, were very wide, which was attributed to the small sample size and the sparseness of the data. The amount of information in the data was insufficient to provide great precision, regardless of the method used.

We considered the effect of slight perturbations of the data on the final inferences. The primary reason was to determine the impact on our analysis if, e.g. one or more of the zeros in the data had actually been observed to be one. This is akin to doing a case-influence analysis (Cook, 1977; Johnson, 1985). If modifying one of the zeros in this way has a large impact on the results, we should not be overly confident in our analysis. Results are presented in Table 3. Both methods gave unstable estimates of prevalence and sensitivity with these sparse data; Bayesian estimates were slightly more stable due to the additional information introduced by the prior.

We next conducted a sensitivity analysis by using relatively non-informative priors. We used uniform (beta (1, 1)) prior probability distributions for all the accuracy parameters, a beta (0.1, 0.9) for P_1 and a beta (0.9, 0.1) for P_2. All estimates were qualitatively similar (see Table 3) except for the estimate of P_2 using the non-informative prior distributions.

Large samples are required to obtain valid CIs based on the ML approach. The sample size for our data was small, which suggests that large sample CIs may be inappropriate. Further evidence is suggested by consideration of the posterior distributions plotted in Fig. 2. If the sample size was large enough, the likelihood would be multivariate normal in shape, and this would cause marginal posteriors to be normal in shape, provided the prior information was not too strong, relative to that in the data. The skewness of the plots in Fig. 2 suggests the lack of normality in the shape of the likelihood, which indicates that large sample normal theory is not valid for these data.

The Bayesian method provided point estimates and intervals without the necessity of a large sample size. Furthermore, the Bayesian intervals were true probability intervals (a 95% Bayesian interval contains the true parameter value with 95% certainty). This interpretation is preferable to the corresponding interpretation for a 95% frequentist CI, which is considered to include the true parameter value 95% of the time in repeated implementations of the data collection-and-analysis procedure. Another advantage of the Bayesian method is that it allows the use of well-documented prior information, which summarizes expert opinion. Traditional frequentist analyses tend to omit this valuable information, while the Bayesian approach embraces it.

6. Conclusions

When evaluating a new diagnostic test, it is generally wise to assume that the sensitivity and specificity of the reference test are not precisely known, and to use available methods to estimate them as well. For this purpose, we advocate the use of the ML methods when two or more populations can be sampled, and when the assumptions for the Hui and Walter model can be justified. A simple Newton–Raphson approach should suffice when the cells of the 2×2 tables displaying the cross-classified test results have large frequencies. When the NR algorithm fails to converge because of small sample sizes, the EM algorithm should be used. Bayesian methods require the modeling of uncertainty about the actual values for test accuracies and prevalences with probability. The Bayesian approach provides somewhat more stable estimates than the EM algorithm in this case (and easily interpretable intervals), but requires the extra step of specifying prior distributions.

All the presented methods are relatively easy to implement using an advanced statistical software program like S-Plus or SAS. Both the EM and the Gibbs sampler require short, and relatively user-friendly program code. Code is available from the authors upon request. When applying these methods, attention should be paid to the validity of assumptions. The work of Vacek (1985) and our examples indicate that violations of these assumptions may lead to severely biased results. To our knowledge, there is currently a lack of known, reliable methods for assessing these assumptions, although some work has been done in this area (Qu et al., 1996; Hadgu and Qu, 1998; Qu and Hadgu, 1998), see also Hui and Zhou (1998)). While it is expected that adding tests and populations to the scenario will allow for testing the Hui and Walter model, additional work is necessary to provide the theoretical basis for such an expectation.

Acknowledgements

The study was supported in part by the NRI Competitive Grants Program/USDA Award No. 98-35204-6535. Additionally, we thank S. Andersen, J. Barlough, W. Cox, I.A. Gardner, R.P. Hedrick, L.M. Pearson, R. Singh and M. Thurmond for valuable contributions to this work.

Appendix A. Technical details for the Gibbs sampler

Define the missing or latent $2\times2\times2$ tables of counts for those individuals who are diseased (D+) to be $\{z_{ijk}\}$. Given the observed counts $\{a_k, b_k, c_k, d_k\}=\{y_{11k}, y_{12k}, y_{21k}, y_{22k}\}$, the table of missing counts consists of independently distributed binomial variates with $z_{ijk}|y_{ijk}$ distributed binomial $(y_{ijk}, \mathrm{Pr}_{ijk})$, where Pr_{ijk} is the conditional probability of being a D+ given the individual is from row i, column j and population k, i.e. has test combination (i, j). For example, using Bayes' theorem

$$\mathrm{Pr}_{111} = \mathrm{Pr}(\mathrm{D}+|\mathrm{T}_1+, \mathrm{T}_2+, \text{ Population 1}) = \frac{P_1\mathrm{Se}_1\mathrm{Se}_2}{\{P_1\mathrm{Se}_1\mathrm{Se}_2+(1-P_1)(1-\mathrm{Sp}_1)(1-\mathrm{Sp}_2)\}},$$

where P_1 is the prevalence of population 1, cf. (Gastwirth, 1987; Brookmeyer and Gail, 1994). Furthermore, the posterior distribution of the parameters, given the data and the missing data $\{z_{ijk}\}$ is the product of independent beta posteriors for each parameter. For example, the augmented data posterior for P_k is

$$\mathrm{beta}(a_{P_k} + z_{\cdot\cdot k}, b_{P_k} + n_k - z_{\cdot\cdot k}),$$

where the parameters a and b have subscripts which indicate that this is the prior for P_k, and the dot subscripts indicate the sum over that index. The corresponding distributions for Se_1 and Se_2 are

$$\mathrm{beta}(a_{\mathrm{Se}_1} + z_{1\cdot\cdot}, b_{\mathrm{Se}_1} + z_{2\cdot\cdot}), \qquad \mathrm{beta}(a_{\mathrm{Se}_2} + z_{\cdot1\cdot}, b_{\mathrm{Se}_2} + z_{\cdot2\cdot}),$$

respectively, and for Sp_1 and Sp_2, they are

$$\mathrm{beta}(a_{\mathrm{Sp}_1} + y_{2\cdot\cdot} - z_{2\cdot\cdot}, b_{\mathrm{Sp}_1} + y_{1\cdot\cdot} - z_{1\cdot\cdot}), \qquad \mathrm{beta}(a_{\mathrm{Sp}_2} + y_{\cdot2\cdot} - z_{\cdot2\cdot}, b_{\mathrm{Sp}_2} + y_{\cdot1\cdot} - z_{\cdot1\cdot}).$$

Thus, given starting values for the parameters, one can alternately sample from these two sets of distributions to obtain a Gibbs sample from the joint distribution of the posterior and proceed exactly as in Joseph et al. (1995).

References

Agger, J.F., Bartlett, P.C., Woudstra, I., Willeberg, P., Houe, H., Lawson, L., Enøe, C., 1997. Evaluation of clinical mastitis and somatic cell count as diagnostic tests for surveillance of udder health in dairy herds. Epidémiologie et Santé Animale 31/32, 12.07.1–12.07.3.

Andersen, S., 1997. Re: Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am. J. Epidemiol. 145, 290–291.

Ashton, J.J., Moeschberger, M.L., 1988. An SAS macro for estimating the error rates of two diagnostic tests, neither being a gold standard. In: Proceedings of the 13th SAS Users Group International Conference, March 27–30, Orlando, FL, pp. 995–996.

Barlough, J.E., McDowell, T.S., Milani, A., Bigornia, L., Pieniazek, N.J., Hedrick, R.P., 1995. Nested polymerase chain reaction for detection of *Enterocytozoon salmonis* genomic DNA in chinook salmon *Oncorhynchus tshawytscha*. Dis. Aquat. Org. 23, 17–23.

Bedrick, E.J., Christensen, R., Johnson, W.O., 1997. Bayesian binomial regression: predicting survival at a trauma center. Am. Statist. 51, 211–218.

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

Brenner, H., Gefeller, O., 1997. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Statist. Med. 16, 981–991.

Brookmeyer, R., Gail, M.H., 1994. AIDS Epidemiology: A Quantitative Approach. Oxford University Press, London, 354 pp.

Choi, B.C.K., 1997. Causal modeling to estimate sensitivity and specificity of a test when prevalence changes. Epidemiology 8, 80–86.

Chriel, M., Willeberg, P., 1997. Dependency between sensitivity, specificity and prevalence analysed by means of Gibbs sampling. Epidémiologie et Santé Animale 31/32, 12.03.1–12.03.3.

Cook, R.D., 1977. Detection of influential observations in linear regression. Technometrics 19, 15–18.

de Bock, G.H., Houwing-Duistermaat, J.J., Springer, M.P., Kievit, J., van Houwelingen, J.C., 1994. Sensitivity and specificity of diagnostic tests in acute maxillary sinusitis determined by maximum likelihood in the absence of an external standard. J. Clin. Epidemiol. 47, 1343–1352.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B 39, 1–38.

Enøe, C., Andersen, S., Thomsen, L.K., Mousing, J., Leontides, L., Sørensen, V., Willeberg, P., 1997. Estimation of the sensitivity and the specificity of two diagnostic tests for the detection of antibodies against *Actinobacillus pleuropneumoniae* serotype 2 in pigs by maximum-likelihood-estimation and Gibbs sampling. Epidémiologie et Santé Animale 31/32, 12.C.34.

Faraone, S.V., Tsuang, M.T., 1994. Measuring diagnostic accuracy in the absence of a gold standard. Am. J. Psychiatr. 151, 650–657.

Faraone, S.V., Blehar, M., Pepple, J., Moldin, S.O., Norton, J., Nurnberger, J.I., Malaspina, D., Kaufmann, C.A., Reich, T., Cloninger, C.R., DePaulo, J.R., Berg, K., Gershon, E.S., Kirch, D.G., Tsuang, M.T., 1996. Diagnostic accuracy and confusability analyses: an application to the diagnostic interview for genetic studies. Psychol. Med. 26, 401–410.

Gardner, I.A., Stryhn, H., Lind, P., Collins, M.T., 2000. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. Prev. Vet. Med. 45, 107–122.

Gart, J.J., Buck, A.A., 1966. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. Am. J. Epidemiol. 83, 593–602.

Gastwirth, J.L., 1987. The statistical precision of medical screening tests. Statist. Sci. 2, 213–238.

Gastwirth, J.L., Johnson, W.O., Reneau, D.M., 1991. Bayesian analysis of screening data: application to AIDS in blood donors. Can. J. Statist. 19, 135–150.

Gelfand, A., Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. J. Am. Statist. Assoc. 85, 398–409.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analaysis. Chapman & Hall, London, 528 pp.

Geisser, S., Johnson, W.O., 1992. Optimal administration of dual screening tests for detecting a characteristic with special reference to low prevalence diseases. Biometrics 48, 839–852.

Georgiadis, M.P., Gardner, I.A., Hedrick, R.P., 1998. Field evaluation of sensitivity and specificity of a polymerase chain reaction (PCR) for detection of *N. salmonis* in rainbow trout. J. Aquat. Anim. Health 10, 372–380.

Greiner, M., Gardner, I.A., 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev. Vet. Med. 45, 3–22.

Hadgu, A., Qu, Y., 1998. A biomedical application of latent class models with random effects. Appl. Statist. 47, 603–616.

Hui, S.L., Walter, S.D., 1980. Estimating the error rates of diagnostic tests. Biometrics 36, 167–171.
Hui, S.L., Zhou, X.H., 1998. Evaluation of diagnostic tests without gold standards. Statist. Meth. Med. Res. 7, 354–370.
Johnson, W.O., 1985. Influence measures for logistic regression: another point of view. Biometrika 72, 59–65.
Johnson, W.O., Gastwirth, J.L., 1991. Bayesian inference for medical screening tests: approximations useful for the analysis of acquired immune deficiency syndrome. J. Roy. Statist. Soc. Ser. B 53, 427–439.
Johnson, W.O., Gastwirth, J.L., 2000. Dual group screening. J. Statist. Plann. Inference 83, 449–473.
Johnson, W.O., Pearson, L.M., 1999. Dual screening. Biometrics 55, 276–282.
Johnson, W.O., Gastwirth, J.L., Pearson, L.M., 2000. Screening without a gold standard: the Hui–Walter paradigm revisited. Am. J. Epidemiol., in press.
Joseph, L., Gyorkos, T.W., Coupal, L., 1995. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am. J. Epidemol. 141, 263–272.
Kraemer, H.C., 1992. Evaluating Medical Tests: Objective and Quantitative Guidelines. Sage, Beverley Hills, CA, 294 pp.
Lee, P.M., 1989. Bayesian Statistics: An Introduction. Oxford University Press, New York, 294 pp.
Line, B.R., Peters, T.L., Keenan, J., 1997. Diagnostic test comparisons in patients with deep venous thrombosis. J. Nucl. Med. 38, 89–92.
Mahoney, W.J., Szatmari, P., Maclean, J.E., Bryson, S.E., Bartolucci, G., Walter, S.D., Marshall, B.J., Zwaigenbaum, L., 1998. Reliability and accuracy of differentiating pervasive developmental disorder subtypes. J. Am. Acad. Child Adolesc. Psych. 37, 278–285.
McClish, D., Quade, D., 1985. Improving estimates of prevalence by repeated testing. Biometrics 41, 81–89.
McDermott, J., Drews, C., Green, D., Berg, C., 1997. Evaluation of prenatal care information on birth certificates. Paediatr. Perinat. Epidemiol. 11, 105–121.
Mendoza-Blanco, J., Tu, X., Iyengar, S., 1996. Bayesian inference on prevalence using a missing-data approach with simulation-based techniques: applications to HIV screening. Statist. Med. 15, 2161–2176.
Meng, X., Rubin, D., 1991. Using EM to obtain asymptotic variance–covariance matrices. J. Am. Statist. Assoc. 86, 899–909.
Office International des Epizooties (OIE), 1996. Principles of validation of diagnostic assays for infectious diseases. In: Manual of Standards for Diagnostic Tests and Vaccines, 3rd Edition. OIE, Paris, pp. 8–15.
Press, S.J., 1989. Bayesian Statistics: Principles, Models, and Applications. Wiley, New York, 237 pp.
Qu, Y., Hadgu, A., 1998. A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. J. Am. Statist. Assoc. 93, 920–928.
Qu, Y., Tan, M., Kutner, M.H., 1996. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics 52, 797–810.
Rybicki, B.A., Peterson, E.L., Johnson, C.C., Kortsha, G.X., Cleary, W.M., Gorell, J.M., 1998. Intra- and inter-rater agreement in the assessment of occupational exposure to metals. Int. J. Epidemiol. 27, 269–273.
Shaw, P.C., van Romunde, L.K.J., Griffion, G., Janssens, A.R., Kreuning, J., Eilers, G.A.M., 1987. Peptic ulcer and gastric carcinoma: diagnosis with biphasic radiography compared with fiberoptic endoscopy. Radiology 163, 39–42.
Sinclair, M.D., Gastwirth, J.L., 1996. On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data. J. Am. Statist. Assoc. 91, 961–969.
Singer, R.S., Boyce, W.M., Gardner, I.A., Johnson, W.O., Fisher, A.S., 1998. Evaluation of bluetongue virus diagnostic tests in free-ranging bighorn sheep. Prev. Vet. Med. 35, 265–282.
Sørensen, V., Barfod, K., Feld, N.C., Nielsen, J.P., Enøe, C., Willeberg, P., 1997. Evaluation of a polyclonal blocking ELISA and a complement fixation test detecting antibodies to *Actinobacillus pleuropneumoniae* serotype 2 in pig serum. Epidémiologie et Santé Animale 31/32, 12.C.43.
Spangler, E., Bech-Nielsen, S., Heider, L.E., 1992. Diagnostic performance of two tests and fecal culture for subclinical paratuberculosis and associations with production. Prev. Vet. Med. 13, 185–195.
Staquet, M., Rozencweig, M., Lee, Y.J., Muggia, F.M., 1981. Methodology for assessment of new dichotomous diagnostic tests. J. Chronic. Dis. 34, 599–610.
Tanner, M.A., 1996. Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 3rd Edition. Springer, New York, 207 pp.
Thibodeau, L.A., 1981. Evaluating diagnostic tests. Biometrics 37, 801–804.

Torrance-Rynard, V.L., Walter, S.D., 1997. Effects of dependent errors in the assessment of diagnostic test performance. Statist. Med. 16, 2157–2175.

Tyler, J.W., Cullor, J.S., 1989. Titers, tests, and truisms: rational interpretation of diagnostic serologic testing. J. Am. Vet. Med. Assoc. 194, 1550–1558.

Vacek, P.M., 1985. The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics 41, 959–968.

Valenstein, P.N., 1990. Evaluating diagnostic tests with imperfect standards. Am. J. Clin. Pathol. 93, 252–258.

van Ulsen, J., Michel, M.F., van Strick, R., van Eijk, R.V.W., van Joost, T., Stolz, E., 1986. Experience with a modified solid-phase enzyme immunoassay for detection of gonorrhea in prostitutes. Sex. Transm. Dis. 13, 1–4.

Walter, S.D., Irwig, L.M., 1988. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J. Clin. Epidemiol. 41, 923–937.

Walter, S.D., Frommer, D.J., Cook, R.J., 1991. The estimation of sensitivity and specificity in colorectal cancer screening methods. Cancer Detec. Prev. 15, 465–469.

Weng, T.S., 1996. Evaluation of a new diagnostic test against a reference test less than perfect in accuracy. Commun. Statist. Simul. Comput. 35, 533–555.

Willeberg, P., Wedam, J.M., Gardner, I.A., Holmes, J.C., Mousing, J.A., Kyrval, J., Enøe, C., Andersen, S., Leontides, L., 1997. A comparative study of visual and traditional post-mortem inspection of slaughter pigs: estimation of sensitivity, specificity and differences in non-detection rates. Epidémiolgie et Santé Animale 31/32, 04.20.1–04.20.3.