Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Preventive Veterinary Medicine 68 (2005) 19–33
PREVENTIVE VETERINARY MEDICINE
www.elsevier.com/locate/prevetmed

Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard

Nils Toft[a,*], Erik Jørgensen[b], Søren Højsgaard[b]

[a]*Department of Animal Science and Animal Health, The Royal Veterinary and Agricultural University, Grønnegårdsvej 8, DK-1870 Frederiksberg C, Denmark*
[b]*Biometry Research Unit, Danish Institute of Agricultural Sciences, P.O. Box 50, DK-8830 Tjele, Denmark*

Abstract

Latent class analysis to assess the sensitivity and specificity of a diagnostic test can be carried out under different assumptions. An often applied set of assumptions is known as the Hui–Walter paradigm, which essentially states that: (i) the population is divided into two or more populations in which two or more tests are evaluated under assumption that (ii) sensitivity and specificity of the tests are the same in all populations; and (iii) the tests are conditionally independent given the disease status. This study explores the implications of these assumptions. Through simulation studies, it is shown how the size of the difference between disease prevalences within the populations influences the precision of the estimates. It is also illustrated by a simulation study how a difference in a test sensitivity between populations may result in estimates that are biased towards the sensitivity of the test in the population with highest disease prevalence, since that population estimate is supported by most of the data. It is shown that the assumption of conditional independence between tests in general cannot be ignored in latent class models. Failure to impose conditional independence will result in a model that lacks identifiability in a way that cannot be handled by adding more tests or dividing the sample into more populations.

Keywords: Hui–Walter paradigm; Latent class analysis; Bayesian analysis; Maximum likelihood estimation

* Corresponding author.
E-mail address: nt@dina.kvl.dk (N. Toft).

0167-5877/$ – see front matter
doi:10.1016/j.prevetmed.2005.01.006

1. Introduction

The performance of a diagnostic test can be evaluated in several ways. The traditional method is to establish a group of animals with known disease status (i.e. classify the animals as positive or negative by a ''gold standard'' method). The sensitivity (Se) of the test (true positive fraction) is then calculated using the diseased animals, and the specificity (Sp) of the test (true negative fraction) using the animals classified by the gold standard test as disease-free. The problem with this approach is that the availability of a gold standard method with perfect sensitivity and specificity is somewhat questionable for a large group of diseases/infections. However, several other approaches have been developed for evaluation of tests in absence of a gold standard, see, e.g. Enøe et al. (2000) for a review of existing methods.

The class of models where the disease status of the individuals is unknown are traditionally referred to as latent class models as the disease status is latent: existing but not presently evident or realized. Use of maximum likelihood estimation in latent class models of test accuracy was first described in Hui and Walter (1980). The assumptions made by Hui and Walter were: (i) the tested individuals are divided into two or more populations with different disease prevalences; (ii) the tests have the same properties in all populations; (iii) the tests are conditionally independent given the true (but latent) disease state. These assumptions have become known as the Hui–Walter paradigm, which data must adhere to in latent class model analysis. From a practical point of view each of the assumptions of the Hui–Walter paradigm can be quite inconvenient. Thus, several approaches to circumvent these requirements have been suggested.

A Bayesian approach to avoid stratifying the population for evaluation of one or two tests was proposed by Joseph et al. (1995). However, as pointed out by Andersen (1997) and recently emphasized by Johnson et al. (2001), even the Bayesian approach requires identifiability of the problem, i.e. data must be able to provide estimates for all the parameters, hence requiring that the degrees of freedom (d.f.) in data is at least equal to the number of parameters. The motivation for avoiding the split into two or more populations stems for large part from the perceived difficulty of obtaining a reasonable split into populations, where a difference in prevalence can be obtained while preserving the assumption that the test properties are the same in each population.

The assumption of conditional independence is often very hard to justify, especially if the tests are based on the same kind of phenomenon, e.g. detection of antibodies, etc. Vacek (1985) demonstrates that failing to allow for conditional dependence will introduce bias in the estimates, in the sense that ignoring a positive correlation will overestimate the test properties while ignoring a negative correlation will underestimate the test properties. Several models that allow for conditional dependence have been proposed using either maximum likelihood or Bayesian estimation, e.g. Qu et al. (1996), Black and Craig (2002), Dendukuri and Joseph (2001) and Yang and Becker (1997).

In this paper, we address the Hui–Walter paradigm once more to explore the importance of the assumptions. We will deal with latent class models on different levels, i.e. cover some general considerations as well as elements that are specific to the individual methods, i.e. particular features of Bayesian methods and maximum likelihood methods, respectively.

To ensure that we explore the assumptions under realistic conditions, we will use examples of test properties similar to those reported for paratuberculosis in cattle in Nielsen et al. (2001). Sample sizes and disease prevalences for no-gold standard estimations are in vicinity of those reported in the studies of Nielsen et al. (2002) and Enøe et al. (2001). In order to have full control of the dataset used for illustration of our various points, we will use simulated data so that the results obtained from latent class analysis can be directly compared to those obtained from traditional methods on the same dataset when the disease status is known.

The rest of this paper is structured as follows: Section 2 briefly reviews why the Hui–Walter assumptions (in a generalized version) are sufficient to allow estimation of sensitivity and specificity in absence of a ''gold standard''. In Section 3, we study the effect of varying the difference in disease prevalence between groups. Section 4 demonstrates how violation of the assumption regarding constant sensitivity and specificity across the different populations affect estimates based on that assumption. The conditional independence is addressed in Section 5. A general discussion concludes the paper.

2. The Hui–Walter assumptions

Whether one uses maximum likelihood estimation or adopts a Bayesian approach to obtain the estimates of test accuracy in absence of a gold standard there are some common assumptions/implications that we need to outline here.

In Section 1, we state that the Hui–Walter paradigm involves a split of the population into two or more populations. Actually, Hui and Walter state that any combination of a number of tests (R) (with constant Se and Sp in all populations) and populations (S) will do as long as $S \geq R/(2^{R-1} - 1)$, i.e. $R = 3$ tests and $S = 1$ population will suffice. However, when equality holds, i.e. two tests, two populations or three tests, one population, we are facing a situation where the degrees of freedom in the data exactly matches the number of parameters. Thus, we are not exactly estimating the parameters, merely rewriting data. As if one were to calculate the mean and standard deviation from a sample of two observations. To adhere to good statistical practice, one should seek to have at least two tests, three populations or four tests, one population. However, throughout our examples, we will use a two tests, two populations scenario to comply with a situation similar to reported studies.

Whenever the number of tests is increased, the assumption that these tests are conditionally independent given disease status becomes harder to maintain. Conditional independence between two tests given disease status, implies that when the disease status of a test subject is known, the probability of a test result is unaffected by knowledge of the outcome of the other test. To illustrate the meaning of conditional independence, consider a test with Se = 0.9. If used in a population of 1000 truly diseased animals, we would expect 100 animals to give a false-negative test result. If a second test with Se = 0.7 is used to test the 100 animals that tested negative on the first test, then conditional independence between the two tests implies that 70 out of the 100 animals are expected to test positive using the second test. If, on the other hand, we assume that there is a positive conditional dependence (which is probably the most intuitive to interpret) between the two tests, then

the sensitivity of the second test when applied to the false-negatives of the first test will be less than 0.7 as fewer test-positives of the second test are expected, e.g. only 10 out of 100, when the first test is negative.

For simplicity, we use only scenarios in which two tests are involved, this implies that data can be organized in 2×2 tables where each cell holds the count of tested individuals with a given combination of the two tests in a subpopulation, thus one such table is obtained for each subpopulation enforced by the split of the population. Assuming conditional independence between the two tests makes it possible to treat the data in such a table as if it were obtained from two latent 2×2 tables, where each test is evaluated against the true disease status. In Table 1, this situation is shown with the cell probabilities for each of the cells given. Note that in each of the latent tables, the cell probabilities sum to one. The cell probabilities in the table of observed data are calculated as the sum of the probabilities of observing a certain result, e.g. both tests positive, given the true disease status is positive and the probability of both tests positive given the true disease status is positive.

Table 1
The cell probabilities in the structure of the latent class model under the assumptions from Hui and Walter (1980)

Latent data

		"Gold standard"	
		Diseased	Non-diseased
Test 1	+	$Se_1 p_1$	$(1 - Sp_1)(1 - p_1)$
	-	$(1 - Se_1)p_1$	$Sp_1(1 - p_1)$
		p_1	$1 - p_1$

		"Gold standard"	
		Diseased	Non-diseased
Test 2	+	$Se_2 p_1$	$(1 - Sp_2)(1 - p_1)$
	-	$(1 - Se_2)p_1$	$Sp_2(1 - p_1)$
		p_1	$1 - p_1$

Observed data

		Test 2	
		+	-
Test 1	+	$Se_1 Se_2 p_1 + (1 - Sp_1)(1 - Sp_2)(1 - p_1)$	$Se_1(1 - Se_2)p_1 + (1 - Sp_1)Sp_2(1 - p_1)$
	-	$(1 - Se_1)Se_2 p_1 + Sp_1(1 - Sp_2)(1 - p_1)$	$(1 - Se_1)(1 - Se_2)p_1 + Sp_1 Sp_2(1 - p_1)$

Only one population is shown to simplify the presentation, the assumption of conditional independence between tests allows data to be organized in the two independent latent class 2×2 tables with positive (+) and negative (−) results of test 1 (and 2) evaluated against a "gold standard". The indices on Se and Sp refer to tests 1 and 2, respectively, whereas p_1 indicates the prevalence of population 1, for population 2 (not shown) the equations would be similar but with p_1 replaced by p_2.

An important property of the way the latent data in Table 1 can be organized is that each of the two 2×2 tables only are used for estimating three parameters, the prevalence (which is the same in both tables) and the sensitivity and specificity of one of the tests. This implies that to estimate the sensitivity and specificity of the two tests as well as the disease prevalence of each population, it is sufficient to ensure that the observed data have enough degrees of freedom to allow estimation of the parameters. Adding a population with a different prevalence only increases the number of parameters by one, but increase the degrees of freedom in the data by three. Hence, two populations with different disease prevalences and two tests with constant test properties are sufficient, but as already stated three populations and two tests would be more appropriate.

3. The assumption of different disease prevalences

It follows from Table 1 that the disease prevalence within the populations must differ (i.e. $p_1 \neq p_2$) in order to increase the degrees of freedom in the data. If the disease prevalences in two populations are the same, then these should be treated as one and the observations combined. Of course, the disease prevalences of different populations are rarely known in advance, hence to stratify the population into (sub)populations with differences in disease prevalence is actually a non-trivial issue, but we will ignore this aspect for now. However, it is important that the stratifier does not interact with the test, and thus does not violate the assumption of constant Se and Sp. Thus, e.g. age is a poor stratifier when evaluating tests for chronic diseases like paratuberculosis. It is questionable whether or not the sensitivity and specificity of a test based on detecting antibodies are the same for heifers and multi-parous cows (this of course implies that a method which allows for inclusion of possible covariates such as age should be used in test evaluation, but we will also ignore that aspect for now).

Here, we consider two tests and two populations with varying prevalences. We consider a situation where $(Se_1, Sp_1) = (0.7, 0.99)$ and $(Se_2, Sp_2) = (0.75, 0.95)$ and the total number of animals tested is 2400 with 1200 in each population.

The formulas originally derived in Hui and Walter (1980) are exact solutions to the case of two tests, two populations. Thus, ML estimates based on these only require that the prevalences differ. However, Hui and Walter (1980) also provide exact solutions to the information matrix which provides estimates of the asymptotic covariance matrix of the parameters. Using these, it is possible to explore the precision on the ML estimates when varying the difference between prevalences while fixing all other elements. In Fig. 1, the standard error of the ML estimates of Se_1 and Sp_1 are plotted against the difference in prevalence between the populations for the case with 1200 animals in each population as well as 200 in each population to compare with sample sizes used in other studies. The figure is constructed using the above assumptions, hence the prevalences of the two populations are centered around 50%. Thus, for a difference of, say 20% point, the prevalence of population 1 is 40% and population 2 is 60%. The increase in precision (or decrease in standard error) is more evident for the specificity because the estimate ($Sp_1 = 0.99$) implies less variation than the estimate of sensitivity ($Se_2 = 0.7$) when combined with test 2 ($Se_2 = 0.75$; $Sp_2 = 0.95$). There is a potential reduction in the standard

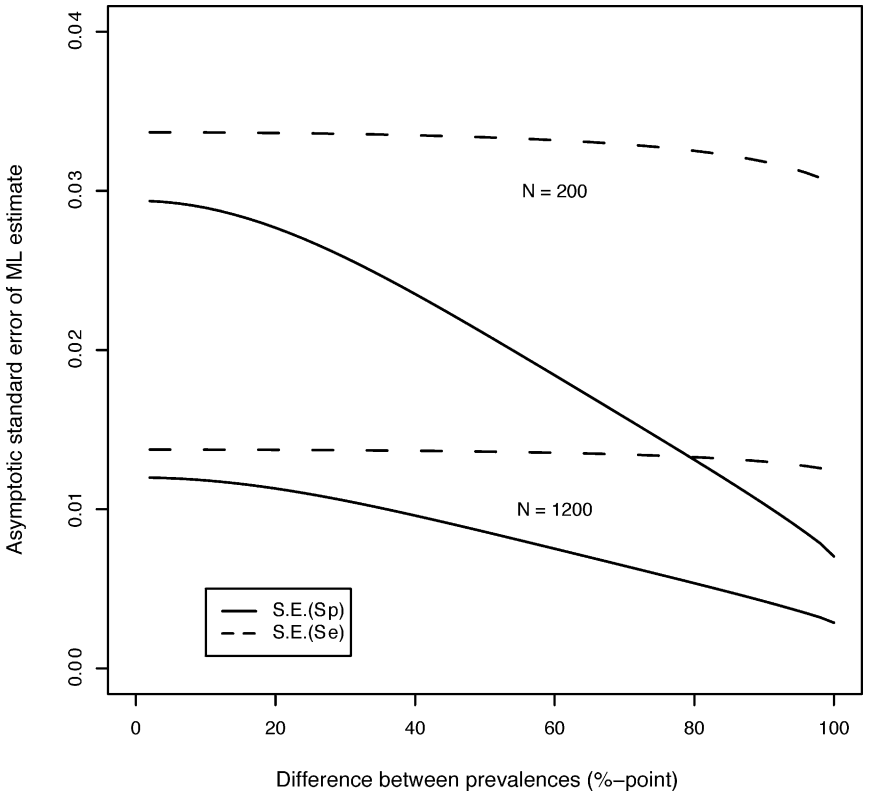

Fig. 1. The standard error of the estimates of sensitivity (Se = 0.7) and specificity (Sp = 0.99) of a test (evaluated in a latent class analysis against a test with Se = 0.75 and Sp = 0.95, using the formulas derived in Hui and Walter (1980)) using 1200 (and 200) diseased and 1200 (and 200) non-diseased animals divided into two populations with increasing difference in prevalence, i.e. the prevalences are centered around 50%.

error of Sp_1 by 76% and a 9% reduction in the standard error of Se_1 by reordering data from the smallest possible difference to the largest possible difference (with 200 animals in each population). If we had knowledge of the disease status, the asymptotic standard error could be calculated using the standard formula, $\sqrt{s(1-s)/n}$, where s is the proportion, n the sample size and it is implicitly assumed that the distribution of the estimate is approximated by a normal distribution. Under these assumptions, using disease status would give a standard error for Se = 0.7 of 0.013 for 1200 diseased animals and 0.032 for 200 diseased animals. The standard errors for Sp = 0.99 are 0.003 and 0.007 for 1200 and 200 non-diseased animals, respectively. Comparing these estimates with the ones obtained using maximum likelihood suggests that the latter might be somewhat optimistic, especially with respect to the sensitivity.

The central limit theorem ensures that for large samples the sampling distribution of a maximum likelihood estimator is approximated by a normal distribution. However, for more realistic sample sizes, it is unclear how good this approximation is. This might impose too narrow bounds on the precision of the estimates. Hence, we have carried out a

simulation study using a Bayesian equivalent to the Hui–Walter model, i.e. two tests, two populations, essentially, the model from Johnson et al. (2001). Prior distributions on the parameters are modelled as non-informative beta-distributions, e.g. beta(1,1) (uniform on the interval [0, 1]).

In Table 2, the posterior mean estimates of the diagnostic values and the associated 95% credible intervals (which are the Bayesian equivalences of confidence intervals (CIs)) are given for scenarios with different prevalences in the populations. As before the only difference between the scenarios is the way the 1200 diseased and 1200 non-diseased animals are organized. Data have been generated by applying the formulas given in Table 1 multiplied with the number of animals in each the of (sub)populations (e.g. 1200). In order to compare the posterior mean estimates and their 95% credible intervals obtained by latent class analysis to the results of a traditional analysis where the disease status is known, the latter have been included in Table 2.

Consider, for example, the sensitivity of test 1 represented in the third column of Table 2. The width of the 95% CI is increasing from 6.2 when the difference in prevalence is 80% point to 9.4, 14.8 and 26.8 in the 40, 20 and 10% point difference, respectively. Thus, the value of full information, i.e. applying a perfect test as well as the two tests under evaluation (corresponding to a width of 5.2) depends on the difference in prevalences. This pattern is seen in the posterior estimates and 95% CIs of the other diagnostic values as well. Furthermore, the distributions of the specificities seem to skew towards the left, while the distributions of the sensitivities skew to the right as the difference in prevalence decrease (most obvious in the 10% point difference case, i.e. when the prevalences are 45 and 55%, respectively). It can even be questioned whether or not the estimates for that scenario still are reasonable compared to the ''true'' values. The width of the 95% credible intervals are in strong contrast to the width of the 95% confidence intervals imposed by the asymptotic standard error of the ML-estimates from Fig. 1 which seem to imply a width of $2 \times 1.96 \times 0.014 = 0.055$, i.e. 5.5% point in the worst case.

Now, consider a case where the difference in prevalence is even less than 10% point. In Fig. 2, we have plotted the posterior distributions of the six parameters from the two tests, two populations case where the simulated disease prevalences are 3 and 10%, respectively in the populations (a situation similar to the one in Nielsen et al. (2002)). Because of the

Table 2
The posterior mean estimates and their 95% credible intervals of sensitivity and specificity obtained by Bayesian latent class analysis using 1200 diseased and 1200 non-diseased animals organized in two populations with 80, 40, 20 and 10% point difference in prevalences

Difference in prevalence	Test 1		Test 2	
	Se (95% CI)	Sp (95% CI)	Se (95% CI)	Sp (95% CI)
80	69.9 (66.6; 73.2)	98.9 (97.6; 99.9)	74.9 (71.8; 78.0)	95.0 (93.2; 96.7)
40	70.2 (65.5; 74.9)	98.2 (95.4; 99.9)	75.7 (72.0; 79.6)	94.6 (90.9; 98.6)
20	71.5 (65.1; 79.9)	96.8 (91.8; 99.9)	77.2 (72.4; 83.9)	93.6 (87.2; 99.3)
10	74.9 (65.1; 91.9)	94.3 (86.0; 99.7)	80.9 (72.9; 95.5)	91.4 (81.2; 99.4)
Known disease status	70.0 (67.3; 72.6)	99.0 (98.3; 99.4)	75.0 (72.5; 77.4)	95.0 (93.7; 96.1)

For comparison, the posterior mean estimates and 95% credible intervals of the sensitivity and specificity estimated using 1200 diseased and 1200 non-diseased animals with known disease status is given.

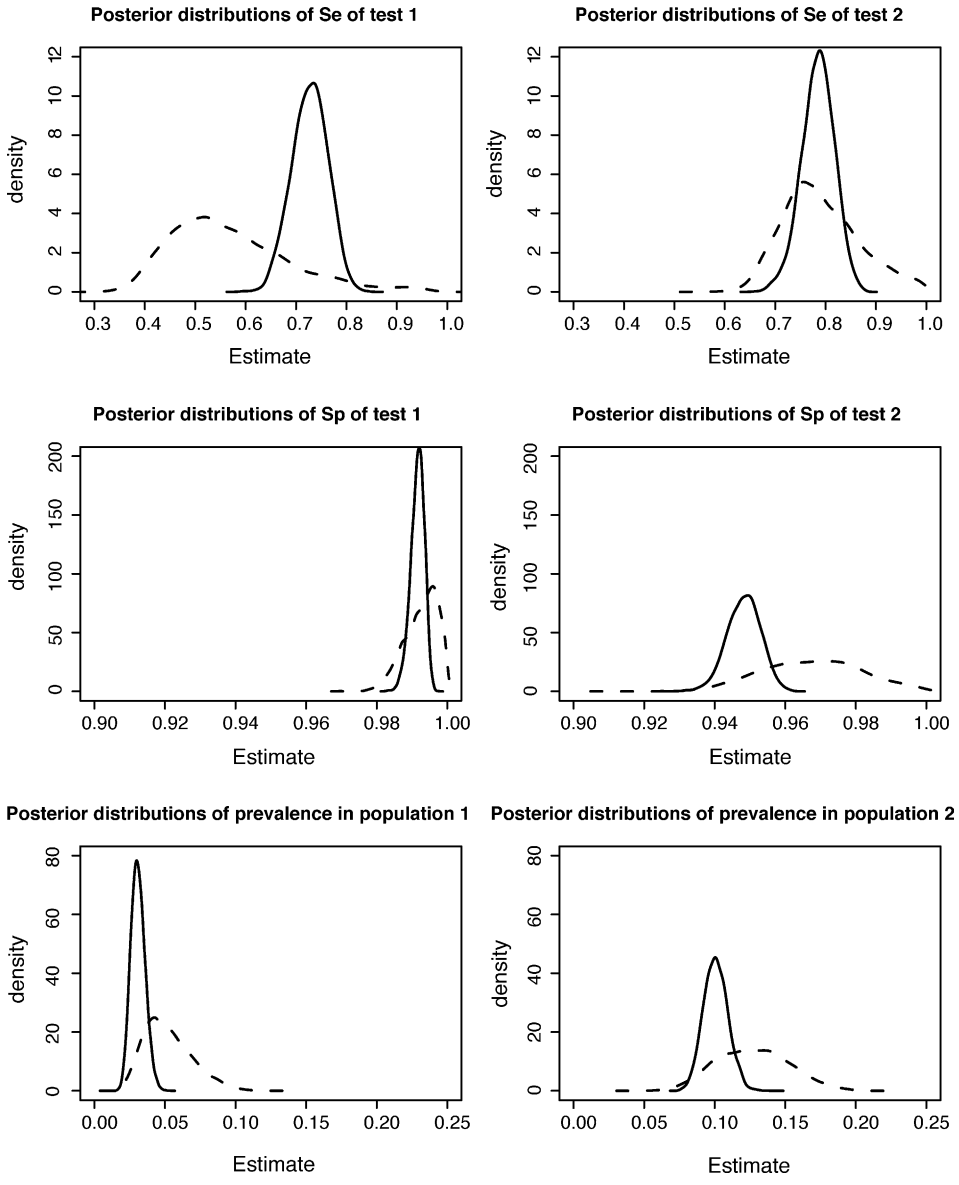

Fig. 2. Posterior distributions for sensitivity, specificity and disease prevalence for the two tests, two populations (of 1200 animals each) scenario with population prevalences simulated as 3 and 10%. Solid lines represent the posterior distributions obtained using the true disease status (e.g. a "gold standard" test as well as the two tests under evaluation), the latent class analysis distributions are shown as dashed lines.

rather low overall disease prevalence the sensitivities of the tests are evaluated using approximately 160 animals. This can be seen to affect the precision of the estimates of sensitivity in Fig. 2, so that also the posterior distributions when using the true disease status are rather wide. The positive side is that the specificity will be based on a large number of animals, hence the precision of these estimates is less influenced by the absence of information on disease status. The overall impression, however, is that there is reason for concern regarding the general validity of the estimates in absence of a gold standard under these conditions. It is of course unfortunate that the sensitivity of the available tests are rather poor since this only adds to the variation.

4. The assumption of constant Se and Sp

From the latent structure in Table 1, it is obvious that if the test sensitivity and specificity vary across populations there is no gain from splitting the population to obtain an identifiable model with degrees of freedom matching the number of parameters. Adding a new population will add three degrees of freedom to data, but three to five new parameters depending on whether both tests vary or not. Adding more tests, however, might do the trick as discussed in Section 2, but we will refrain from pursuing this option for reasons that we shall elaborate on in the next section.

A split based on biological factors such as age or sex is often accused of violating the assumption of constant Se and Sp between populations, due to factors such as increasing cross-reactions with age, etc. Hence, it is often preferred to use a geographic split using, zip-code, veterinary practices, etc. as stratifiers. It is unclear to the authors why it is generally believed that this will cause less problems. Apart from the obvious chance of region specific cross-reactions due to specific antigens, the major problem still remains: is this split providing populations with different disease prevalences?

Rather than elaborating more on these elements, we provide an example in Fig. 3 where the assumption of constant sensitivity is violated for one of the tests. We use the example from the previous section, with 1200 diseased and 1200 non-diseased animals organized in two populations with disease prevalences of 10 and 90%, respectively. We will assume that Se_1 differs between populations, so that the low prevalence population has a sensitivity of 50% and the high prevalence populations a sensitivity of 90% which yield a combined sensitivity of $0.1 \times 50\% + 0.9 \times 90\% = 86\%$ if the test is evaluated on the full data set using the true disease status. The posterior distributions of Fig. 3 reveal some rather interesting properties that in retrospect become obvious. The posterior distribution of Se_1 using the latent class analysis is centered around 0.90, i.e. the sensitivity in the population with high prevalence. Thus, the latent class analysis does not give a mixture of the two sensitivities but relies on the estimate supported by most data. In this case, the high prevalence of population 2 makes that sensitivity the ‘‘true’’ one. The consequence of violating the assumption can been seen to influence the other parameters as well, in the present case particularly the specificity of test 2 and the prevalence of population 1.

Although the example might be somewhat artificial, it raises some concerns about a number of issues. Suppose that we are actually interested in a test for screening purposes. Then, we might have a disease prevalence of 10% or even lower in the target population.

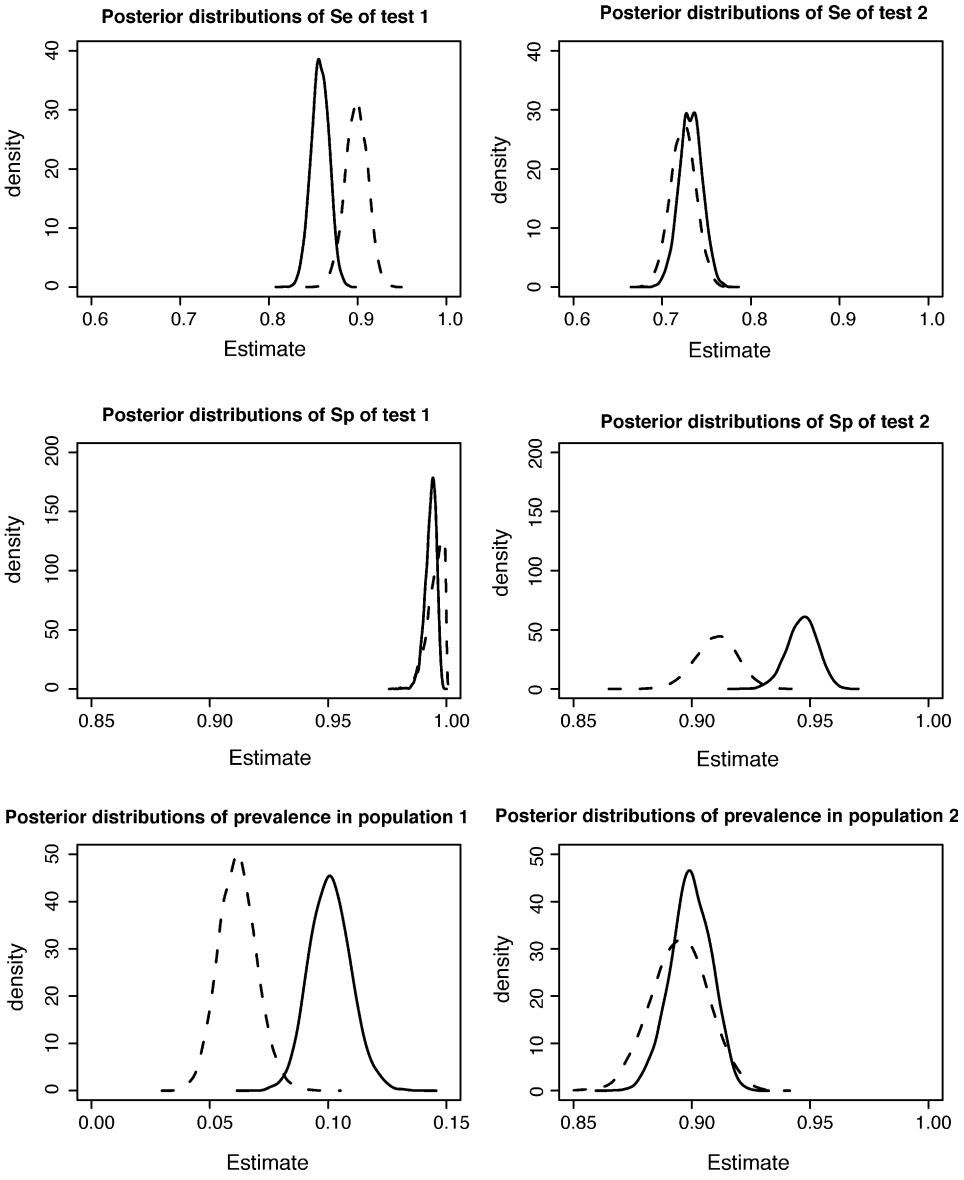

Fig. 3. Posterior distributions for sensitivity, specificity and disease prevalence for the two tests, two populations (of 1200 animals each) scenario with population prevalences simulated as 10 and 90%. The assumption of constant sensitivity across populations is violated for test 1, i.e. $Se_1 = 0.5$ for the low prevalence population (1) and $Se_2 = 0.9$ for the high prevalence population (2). Other parameters are kept constant in accordance with the example used throughout this paper. Solid lines represent the posterior distributions obtained using the true disease status, the latent class analysis distributions are shown as dashed lines.

Still, to evaluate the test, one uses a sample from the population, but also a selected sample of slaughtered animals that showed clinical symptoms at slaughter. It is actually not that unreasonable to assume that in such a population, the antibody responses, etc. are more pronounced, thus giving a better sensitivity of the test. Clearly, in our case this gives problems. But there is another consequence which is just as interesting. In traditional analysis, clinical symptoms at slaughter could serve as a ''gold standard'' to evaluate the tests against, hence introduce serious selection bias in the sample used for evaluation of the test in a traditional approach. This would yield an estimate of Se and Sp that are incorrect for the general population with a wider range of disease manifestations. Thus, when the test is used in the general population with the assumption that the test have the same properties as in the population used for evaluation, then the conclusions made from using the test could be compromised. Using the result from a traditional test evaluation as prior information in a latent class analysis might also lead to misleading conclusions due to this selection bias.

5. The assumption of conditional independence

If the assumption of conditional independence between tests cannot be justified, then the two latent 2×2 tables in Table 1 do not suffice to derive the joint probability density function of the combined test results. Whenever conditional independence cannot be assumed, there is a need to decide how to model the conditional dependence between tests. Here, we will adopt the model suggested by Vacek (1985) where conditional dependence of the two tests is modelled using a covariance when the true state is diseased and a covariance when the true state is non-diseased. This model or a re-parameterization of the model is essentially the same used in most reported studies. However, it is important to realize that while the model without conditional dependence is unique, there exists several different models for conditional dependence. Thus, the choice of this specific model is a further assumption.

To estimate such conditional dependence between the sensitivity (and specificity) of two tests, the latent data needs to be divided into a diseased population and a non-diseased population. Based on these the sensitivity, specificity, prevalence and covariances between the two tests (γ_{Se} and γ_{Sp}) can be found as in Gardner et al. (2000). This structure as well as the cell probabilities are given in Table 3. Here, both tables are needed to sum the probabilities to one, implying that there is dependence between the latent tables. It is worth to note that it is possible to obtain upper and lower bounds on the covariances (γ_{Se} and γ_{Sp}), using the cell probabilities in the latent tables of Table 3. Unfortunately, there is a problem with this latent structure. Each of the two 2×2 tables requires four parameters to be specified: one for each of the sensitivities (specificities), one for the prevalence and one for the covariance between sensitivities (specificities). This constitutes a non-identifiable problem, since in a 2×2 table only three parameters can be estimated. This has the implication, that regardless of how many populations the population is divided into, the lack of identifiability in the individual latent structures will imply that estimates of a model with assumed conditional dependence in this form, cannot be reliably obtained from the data. The problem is the same if one seeks to increase the degrees of freedom by adding more tests rather than adding more populations.

Table 3
The cell probabilities in the structure of the latent class model under the assumptions from Hui and Walter (1980), but allowing for conditional dependence modelled as covariances as in Vacek (1985)

Latent data

Diseased group

		Test 2 +	Test 2 -
Test 1	+	$(\mathrm{Se}_1\mathrm{Se}_2 + \gamma_{\mathrm{Se}})p_1$	$(\mathrm{Se}_1(1-\mathrm{Se}_2) - \gamma_{\mathrm{Se}})p_1$
	-	$((1-\mathrm{Se}_1)\mathrm{Se}_2 - \gamma_{\mathrm{Sp}})p_1$	$((1-\mathrm{Se}_1)(1-\mathrm{Se}_2) + \gamma_{\mathrm{Se}})p_1$

Non-diseased group

		Test 2 +	Test 2 -
Test 1	+	$((1-\mathrm{Sp}_1)(1-\mathrm{Sp}_2) + \gamma_{\mathrm{Sp}})(1-p_1)$	$((1-\mathrm{Sp}_1)\mathrm{Sp}_2 - \gamma_{\mathrm{Sp}})(1-p_1)$
	-	$(\mathrm{Sp}_1(1-\mathrm{Sp}_2) - \gamma_{\mathrm{Sp}})(1-p_1)$	$(\mathrm{Sp}_1\mathrm{Sp}_2 + \gamma_{\mathrm{Sp}})(1-p_1)$

Observed data

		Test 2 +	Test 2 -
Test 1	+	$(\mathrm{Se}_1\mathrm{Se}_2 + \gamma_{\mathrm{Se}})p_1$ $+((1-\mathrm{Sp}_1)(1-\mathrm{Sp}_2) + \gamma_{\mathrm{Sp}})(1-p_1)$	$(\mathrm{Se}_1(1-\mathrm{Se}_2) - \gamma_{\mathrm{Se}})p_1$ $+((1-\mathrm{Sp}_1)\mathrm{Sp}_2 - \gamma_{\mathrm{Sp}})(1-p_1)$
	-	$((1-\mathrm{Se}_1)\mathrm{Se}_2 - \gamma_{\mathrm{Se}})p_1$ $+(\mathrm{Sp}_1(1-\mathrm{Sp}_2) - \gamma_{\mathrm{Sp}})(1-p_1)$	$((1-\mathrm{Se}_1)(1-\mathrm{Se}_2) + \gamma_{\mathrm{Se}})p_1$ $+(\mathrm{Sp}_1\mathrm{Sp}_2 + \gamma_{\mathrm{Sp}})(1-p_1)$

Only one population is shown to simplify the presentation, the assumption of conditional dependence between tests requires data to be organized in the two dependent latent class 2 × 2 tables. The indices on Se and Sp refer to tests 1 and 2 respectively, whereas p_1 indicates the prevalence of population 1, for population 2 (not shown) the equations would be similar but with p_1 replaced by p_2.

Despite this problem, several authors have proposed latent class models without the assumption of conditional independence between tests in the absence of a gold standard test. The first paper to address the implications of conditional dependence between tests in the Hui–Walter model was Vacek (1985). Vacek did not estimate the covariances directly, but explored how an assumed fixed proportion of the maximum possible covariance between tests would influence the estimates. Using that approach, she demonstrated how a possible conditional dependence would produce seriously biased estimates if ignored. Another maximum likelihood approach is that of Qu et al. (1996) who uses random effects models to model dependence between multiple tests or readers of X-rays, which is their prime concern. They devise a model for more than four testers, thus avoiding to split the population. It is unclear how this approach handles the lack of identifiability caused by conditional dependence, but the implications of using random effect to model dependence should be noted. Using random effects to model dependence, one must implicitly assume that dependence between tests (or readers) is due to the fact that they are realizations of the same distribution. An assumption which perhaps can be justified for readers of X-rays, but hardly a proper generalization for diagnostic tests in general. Yang and Becker (1997) uses log-linear models to allow models of conditional dependence modelled as pairwise non-independence between tests. Their setup is a four tests, one population situation in which the dependence between two tests is modelled using only one parameter, furthermore they only allow conditional dependence between a subset of the tests, i.e. at least one test is assumed to be conditionally independent of the others. This assumption ensures that enough of the latent structure can be organized according to the setup in Table 1, hence making the estimation of the pairwise association between test possible. It is not obvious how to interpret such a model for conditional dependence in terms of technical/biological test properties.

The Bayesian models of Dendukuri and Joseph (2001), Black and Craig (2002) and Hanson et al. (2005) all adopt prior information, but ignore that data itself only contribute information about a combination of a small subset of the specified parameters (in this case proportion of test positive), while the posterior distribution of other parameters is a direct result of the prior specification (see Neath and Samiengo (1997) for further discussion). The priors used in Black and Craig (2002) for test sensitivity and specificity actually contain more information, than the 200 observations used in the simulations would provide, even if the true disease status was known. Furthermore, Black and Craig (2002) does not split the population into two or more populations but elaborate on the two tests, one population idea of Joseph et al. (1995), which have already been criticized for lack of identifiability. In Hanson et al. (2005), prior distributions are imposed upon the prevalences of the different populations. Where do these prior distributions come from? If one knows the prevalence of a disease in a certain area, one must have a gold standard test available. If one already has good estimates of test properties to use as priors, then why carry out the evaluation?

Adding prior information about certain parameters in a Bayesian approach is sometimes perceived as equivalent to simply fixing in advance a certain number of parameters in an ML-estimation to obtain an identifiable problem, i.e. fixing the disease prevalences or the properties of one of the tests. However, a major difference between the two scenarios is that in the latter, one does not believe that the fixed parameter is estimated by data. It is

important to realize that the ''robustness'' to misspecification of prior distributions only relates to identifiable problems, where information from observed data will dominate the prior specification. We feel that the elicitation of prior distributions should include the consequences of the prior specification especially regarding the non-identifiable subset of the parameters. Thus in general, the assumption of conditional independence cannot be relaxed, without posing serious problems to the estimation procedure. It seems that the most sensible plan of attack to this problem is the one used by Vacek (1985).

6. Discussion

The main conclusions from this study can be summarized as follows:

The assumption of conditional independence is central in the Hui–Walter paradigm. Adding more tests or more populations will not allow that data support all the parameters necessary to handle conditional dependence when modelled as in the majority of studies, i.e. by allowing different covariances between tests when the true state of the animal is diseased and non-diseased, respectively.

The smaller the difference between disease prevalences in the populations, the less the precision in the estimates. Furthermore, our simulation studies indicate that not only the precision, but also the estimate itself might be affected, when the difference between disease prevalence in the populations decrease. Although sample sizes of 2400 animals might seem impressive, the lack of diseased animals in the population might impose serious problems when estimating the sensitivity of the tests in scenarios with an overall low prevalence of the disease. Furthermore, the estimated asymptotic standard errors are based on large-sample theory, thus for small samples these estimates become questionable.

By simulation, we showed how the lack of constant sensitivity of a test between populations introduced bias towards the estimate supported by the population with the highest disease prevalence.

The combined effect of these conclusions allows us to emphasize what others also have pointed out: when choosing two tests to evaluate against each other using latent class analysis, make sure that the tests are based on different physiological phenomena to obtain the best possible compliance with the conditional independence assumption.

A point which is often ignored when using estimates of certain properties such as the test accuracy is the precision of the estimates. Usually, these are quoted in the original study and ignored thereafter. However, when the precision of an estimate becomes so poor that the 95% CI spans the interval [0.4, 0.9], the validity of that estimate becomes questionable and any inference made from such estimates should be treated with caution. In general, the precision of an estimate should be included in the decision process whenever decisions are made from such estimates. Thus, to obtain reasonable estimates for use in decision making, the latent class analysis should use sample sizes that at least produce estimates with the same precision as those obtained from the traditional analysis (not that the precisions of these estimates generally are impressive as reported in, e.g. Nielsen et al. (2001)). The Bayesian solution to poor precision from lack of data is to use prior information about the test properties and/or disease prevalence and use that in the analysis. However, a good Bayesian analysis includes a judgement of the importance of the prior. For instance, the

results should be reported using different priors, e.g. a non-informative, a skeptical and a optimistic prior. Using this approach, the impact of observed data on the conclusions may be judged, see Spiegelhalter et al. (2000) for elaboration on the use of Bayesian methods in health technology assessment. On top of these concerns, there is still the problem of how to obtain the priors. As the example in Section 4 showed, incorrect information will lead to biased estimates. Thus, if priors are based on results obtained from traditional test evaluations, then extreme care should be taken to ensure that the tests under evaluation perform equally in the populations used for the traditional analysis and the target population designated for latent class analysis.

References

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 (3), 290–291.

Black, M.A., Craig, B.A., 2002. Estimating disease prevalence in the absence of a gold standard. Stat. Med. 21, 2653–2669.

Dendukuri, N., Joseph, L., 2001. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics 57, 158–167.

Enøe, C., Andersen, S., Sørensen, V., Willeberg, P., 2001. Estimation of sensitivity, specificity and predictive values of two serologic tests for the detection of antibodies against *Actinobacillus pleuropneumoniae* serotype 2 in the absence of a reference test (gold standard). Prev. Vet. Med. 51, 227–243.

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

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

Hanson, T., Johnson, W., Gardner, I. Hierarchical models for estimating herd prevalence and test accuracy in the absence of a gold-standard. J. Agric. Biol. Environ. Stat. 8 (2), 223–239.

Hui, S.L., Walter, S.D., 1980. Estimating the error rates of diagnostic tests. Biometrics 36, 167–171.

Johnson, W.O., Gastwirth, J.L., Pearson, L.M., 2001. Screening without a ‘‘gold standard’’: the Hui–Walter paradigm revisited. Am. J. Epidemiol. 153 (9), 921–924.

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. Epidemiol. 141 (3), 263–272.

Neath, A.A., Samiengo, E.J., 1997. On the efficacy of bayesian inference for nonidentifiable models. Am. Stat. 51 (3), 225–232.

Nielsen, S.S., Grønbæk, C., Agger, J.F., Houe, H., 2002. Maximum-likelihood estimation of sensitivity and specificity of ELISAs and faecal culture for diagnosis of paratuberculosis. Prev. Vet. Med. 53, 191–204.

Nielsen, S.S., Nielsen, K.K., Huda, A., Condron, R., Collins, M.T., 2001. Diagnostic techniques for paratuberculosis. Bull. Int. Dairy Fed. 362, 5–17.

Qu, Y., Tan, M., Kutner, M.K., 1996. Random effects models for evaluating accuracy of diagnostic tests. Biometrics 52 (3), 797–810.

Spiegelhalter, D.J., Myles, J.P., Jones, D.R., Abrams, K.R., 2000. Bayesian methods in health technology assessment: a review. Health Technol. Assess. 4 (38), 1–130.

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

Yang, I., Becker, M.P., 1997. Latent variable modeling of diagnostic accuracy. Biometrics 53 (4), 948–958.