REVSTAT – Statistical Journal
Volume 12, Number 1, March 2014, 85–99

MODELING WITHOUT A GOLD STANDARD: STRATIFICATION WITH STRATUM-DEPENDENT PARAMETERS

Authors: Francisco Louzada
– Department of Applied Mathematics and Statistics,
Institute of Mathematical Science and Computing,
University of São Paulo, São Carlos, Brazil
louzada@icmc.usp.br

Gilberto de Araujo Pereira
– Department of Nursing,
Federal University of Triângulo Mineiro, Uberaba, Brazil
pereira_gilberto@yahoo.com.br

Márcia M. Ferreira-Silva
– Federal University of Triângulo Mineiro, Uberaba, Brazil
marcia.mferreira@yahoo.com.br

Valdirene de Fátima Barbosa
– Research Group for Blood Transfusion Security,
Federal University of Triângulo Mineiro, Uberaba, Brazil
valdirene_fbarbosa@yahoo.com.br

Helio de Moraes-Souza
– Department of Medical Clinic,
Federal University of Triângulo Mineiro, Uberaba, Brazil
helio.moraes@dcm.uftm.edu.br

Gleici S. Castro Perdona
– Departament of Social Medicine,
University of São Paulo, Ribeirão Preto, Brazil
pgleici@fmrp.usp.br

Abstract:

- Bayesian latent-class models have been widely applied for assessing the performance of diagnostic tests in the absence of a gold standard. We provide a short discussion on identifiability issues appearing under the absence of a gold standard, and construct an extension of the well-known Hui–Walter stratification model which allows for stratum-dependent parameters. We illustrate our approach using a Chagas disease case study on blood donors from Brazil.

Key-Words:

- *absence of a gold standard; diagnostic test; identifiability; sample size; stratification.*

AMS Subject Classification:

- 49A05, 78B26.

1. INTRODUCTION

In the area of diagnostic medicine, it is common that the medical practitioner considers one or more complementary diagnostic tests for decision-making and detailed clinical analysis. Within this context, it is important that the physician knows the parameters of the test to be used, such as sensitivity and/or specificity, false-positive and/or false-negative rates, and positive and/or negative predictive values. The modeling structure for this estimation problem is relatively simple and straightforward when the subjects being investigated are submitted to the so-called gold standard test for confirmation, as they are usually 100% sensitive and specific (Kraemer, 1992).

However, in many practical situations no patient under investigation is submitted to a confirmatory test (Joseph *et al.*, 1995), either due to the lack of such a test or its high invasiveness, or to the high cost of its large scale implementation, or to the presence of subgroups with different prevalence rates (Hui and Walter, 1980).

Our main objectives here are to provide a short discussion on identifiability issues appearing under the absence of a gold standard, and to construct an extension of the Hui–Walter stratification model which allows for stratum-dependedent performance parameters. In the next section we discuss the modeling concepts and the inference techniques. In Section 3 we report details on numerical experiments, and we provide an illustration to Chagas disease data in Section 4.

2. MODELING WITHOUT A GOLD STANDARD

2.1. Absence of gold standard

In the case where the health condition of a subject (D) cannot be verified, due to the absence of a gold standard, the likelihood for a random sample of n subjects, can be written as

$$\mathscr{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{l=1}^{L}\left\{\xi\,\mathrm{se}_l^{t_{i,l}}(1-\mathrm{se}_l)^{1-t_{i,l}} + (1-\xi)\,\mathrm{sp}_l^{1-t_{i,l}}(1-\mathrm{sp}_l)^{t_{i,l}}\right\}, \tag{2.1}$$

where, $\boldsymbol{\theta} = (\xi, \mathbf{se}, \mathbf{sp})^{\mathrm{T}}$, with ξ denoting the disease prevalence, and

$$\mathbf{se} = (\mathrm{se}_1, ..., \mathrm{se}_L)^{\mathrm{T}}\,, \qquad \mathbf{sp} = (\mathrm{sp}_1, ..., \mathrm{sp}_L)^{\mathrm{T}}\,,$$

Here se_l and sp_l are respectively the sensitivity and specificity of the lth test, and $t_{i,l}$ is the outcome of the lth diagnostic test on the ith subject (0: negative, 1: positive). In this model there are $2L+1$ parameters to be estimated, and $2^L - 1$ degrees of freedom.

A popular approach for modeling the performance of diagnostic tests, under the absence of a gold standard, is to consider latent classes. In this setting, the health condition y_i of the ith subject (healthy or diseased) can be modeled through a Bernoulli random variable, Y, with probability of success,

$$\tau_i = \frac{\xi \prod_{l=1}^{L} \text{se}_l^{t_{i,l}} (1-\text{se}_l)^{1-t_{i,l}}}{\xi \prod_{l=1}^{L} \text{se}_l^{t_{i,l}} (1-\text{se}_l)^{1-t_{i,l}} + (1-\xi) \prod_{l=1}^{L} \text{sp}_l^{1-t_{i,l}} (1-\text{sp}_l)^{t_{i,l}}} . \tag{2.2}$$

By combining the likelihood of the incomplete data (2.1) with the likelihood of the latent variable, Y, we can write the augmented likelihood (Dempster *et al.*, 1977; Tanner and Wong, 1987) for the case where L diagnostic tests are conducted, as

$$\mathscr{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{l=1}^{L} \left[\left\{ \xi \, \text{se}_l^{t_{i,l}} (1-\text{se}_l)^{1-t_{i,l}} \right\}^{y_i} \left\{ (1-\xi) \, \text{sp}_l^{1-t_{i,l}} (1-\text{sp}_l)^{t_{i,l}} \right\}^{1-y_i} \right], \tag{2.3}$$

where y_i is the unobservable health condition of the ith subject (0: healthy; 1: diseased), which is modeled through a Bernoulli distribution with probability of success τ_i as given in (2.2). Estimation can then be conducted through numeric methods, such as the Expectation-Maximization algorithm (EM) (Dempster *et al.*, 1977), in the frequentist context, and Gibbs sampling (Gelfand and Smith, 1990) or a Metropolis–Hastings algorithm (Chib and Greenberg, 1995), in the Bayesian context.

According to Swartz *et al.* (2004), a primary difficulty regarding latent-class models is related to identifiability issues, and one of the practical lessons obtained by using them is that this issue becomes relatively less important as the dimension of the model increases.

2.2. Identifiability

The modeling approach discussed in §2.1 has been widely applied in the literature, for the case where the model obeys what we will call throughout as the *basic identifiability condition*,

$$\text{df} \geq p \, , \tag{2.4}$$

where df is the number of degrees of freedom, and p is the number of parameters to be estimated. For example, for the latent-class model in §2.1 to obey the basic identifiability condition, $\text{df} = 2^L - 1 \geq 2L + 1 = p$, a minimum of three conditionally independent tests is required.

Several procedures for assessing identifiability have been documented in the literature. For example, Goodman (1974) discusses a Jacobian-based criterion, whereas Garret and Zeger (2000) proposes a graphical method to assess weak identifiability, which is based on the idea that weak identifiability is associated with smaller sample sizes relatively to the number of latent classes, case in which the number of subjects may be insufficient to assign an element to each class.

The Bayesian approach offers here an important advantage: Although a certain model may not be identifiable, it is always valid as data can be suitably described from both its identifiable parameters and prior information (Lindley, 1971); this point is reinforced by Neath and Samaniego (1997), who support the view that Bayesian analysis may yield reasonable answers even for nonindentifiable models.

2.2.1. Hui–Walter stratification

To reestablish the basic identifiability condition many approaches have been considered, such as the introduction of constraints on the parameter space (Walter and Irwig, 1988), the choice of informative priors according to well defined criteria (Gustafson, 2005), or stratification-based approaches (Hui and Walter, 1980). These latter approaches are known as the Hui–Walter paradigm, and will be of particular interest for the remainder of this article; the Hui–Walter stratification paradigm has been widely discussed in the literature, and it has been modeled through a wealth of Bayesian and frequentist approaches (Singer *et al.*, 1998; Johnson *et al.* 2001; Nielsen *et al.*, 2002; Gustafson, 2005; Gardner, 2004; Toft *et al.*, 2005; Branscum *et al.*, 2005; Bertrand *et al.*, 2005; Toft *et al.*, 2007, among others).

The Hui–Walter stratification model is based on stratum-dependent disease prevalence rates, although it uses equal performance parameters across strata. Stratification increases the number of parameters to $2^L + V$ and the number of degrees of freedom to $2^L V - V$; hence, if the population is divided into two strata ($V = 2$), a minimum of two conditionally independent tests ($L = 2$) is sufficient to obey the basic condition for identifiability (2.4). As a byproduct, stratification also allows us estimate specific disease prevalence rates in each homogeneous subpopulation.

For the absence of gold standard, the likelihood of the Hui–Walter stratifi-

cation model can be written as

$$\mathscr{L}(\boldsymbol{\theta}) = \prod_{v=1}^{V}\prod_{i=1}^{n_v}\left[\left\{\xi_v\prod_{l=1}^{L}\mathrm{se}_l^{t_{i,l}}(1-\mathrm{se}_l)^{1-t_{i,l}}\right\}^{y_{i,v}} \times\left\{(1-\xi_v)\prod_{l=1}^{L}\mathrm{sp}_l^{1-t_{i,l}}(1-\mathrm{sp}_l)^{t_{i,l}}\right\}^{1-y_{i,v}}\right],$$

where n_v and ξ_v are respectively the number of subjects and the prevalence rate in the vth stratum, whereas $y_{i,v}$ is the unobservable health condition of the ith subject in the vth stratum, modeled through a Bernoulli distribution with probability of success $\tau_{i,v}$.

Toft *et al.* (2005) pointed out some potential pitfalls of the Hui–Walter paradigm, particularly regarding the accuracy of estimates, which are strongly influenced by the magnitude of the difference in disease prevalence rates between strata, suggesting that the greater the difference of the prevalence rates the higher the estimation accuracy (smaller amplitude of 95% credibility interval) in the case of two tests ($L = 2$) and two strata ($V = 2$). Moreover, sensitivity and specificity may be overestimated.

2.2.2. Extended stratification

Since in most practical situations it is rather challenging to find a stratification factor in which both sensitivity and specificity of the tests are kept similar across strata, here we propose an extension of the Hui–Walter model which assumes that not only prevalences (ξ)—but also sensitivities and specificities—are stratum-dependent. Specifically, our setting is the following: We assume that L diagnostic tests are conducted—none of which being a gold standard—and we assume that the population is divided into V strata, with stratum-dependent prevalences

$$\left\{\xi_v = P_v(D=1)\colon\ v = 1, ..., V\right\},$$

and with stratum-dependent performance parameters,

$$\left\{(\mathrm{se}_{l,v}, \mathrm{sp}_{l,v})\colon\ l = 1, ..., L;\ v = 1, ..., V\right\}.$$

The unobservable health condition of a subject in the vth stratum, Y_v, can be modeled through a Bernoulli distribution, with probability of success τ_v. With our extension of the Hui–Walter model, the number of parameters increases to $2LV + V$, whereas the number of degrees of freedom remains unchanged ($\mathrm{df} = 2^L V - V$). This means that, for example, for a population stratified into two strata ($V = 2$), at least three tests need to be conducted ($L \geq 3$), so that the model obeys the basic condition for identifiability (2.4). (Compare with §2.2.1.)

The augmented data likelihood of the latent class model, for the general case of L conditionally independent tests and V strata, can be written as

$$\mathscr{L}(\boldsymbol{\theta}) = \prod_{v=1}^{V}\prod_{i=1}^{n_v}\left[\left\{\xi_v \prod_{l=1}^{L} \mathrm{se}_{l,v}^{t_{i,l,v}}(1-\mathrm{se}_{l,v})^{1-t_{i,l,v}}\right\}^{y_{i,v}} \times\left\{(1-\xi_v)\prod_{l=1}^{L}\mathrm{sp}_{l,v}^{1-t_{i,l,v}}(1-\mathrm{sp}_{l,v})^{t_{i,l,v}}\right\}^{1-y_{iv}}\right], \tag{2.5}$$

where, $\boldsymbol{\theta} = (\boldsymbol{\xi}, \mathbf{se}_1, ..., \mathbf{se}_V, \mathbf{sp}_1, ..., \mathbf{sp}_V)^{\mathrm{T}}$ with

$$\boldsymbol{\xi} = (\xi_1, ..., \xi_V)^{\mathrm{T}}, \qquad \mathbf{se}_l = (\mathrm{se}_{1,v}, ..., \mathrm{se}_{L,v})^{\mathrm{T}}, \qquad \mathbf{sp}_l = (\mathrm{sp}_{1,v}, ..., \mathrm{sp}_{L,v})^{\mathrm{T}},$$

for $v = 1, ..., V$. Here, ξ_v is prevalence rate in the vth stratum, whereas $\mathrm{se}_{l,v}$ and $\mathrm{sp}_{l,v}$ are the sensibility and specificity of lth test in the vth stratum, respectively; in addition, $t_{i,l,v}$ is the lth test result for the ith subject in the vth stratum, and $y_{i,v}$ is the unobservable health condition of the ith subject in the vth stratum, which is modeled through a Bernoulli with success probability,

$$\tau_{i,v} = \frac{\xi_v \prod_{l=1}^{L} \mathrm{se}_{l,v}^{t_{i,l,v}}(1-\mathrm{se}_{l,v})^{1-t_{i,l,v}}}{\xi_v \prod_{l=1}^{L} \mathrm{se}_{l,v}^{t_{i,l,v}}(1-\mathrm{se}_{l,v})^{1-t_{i,l,v}} + (1-\xi_v)\prod_{l=1}^{L}\mathrm{sp}_{l,v}^{1-t_{i,l,v}}(1-\mathrm{sp}_{l,v})^{t_{i,l,v}}},$$

for $i = 1, ..., n_v$ and $v = 1, ..., V$.

The non-stratified model ($V = 1$) in (2.3), and the Hui–Walter model in (2.2.1) are particular cases of our stratification model with stratum-dependent parameters.

2.2.3. Inference

A fully Bayesian approach is here used for conducting inference. This choice is based on the fact that each parameter in (2.5) is directly interpreted within the context of diagnostic tests, including the availability of expert opinions that can be modeled separately in terms of prior distribution for each parameter. We consider Beta$(1, 1)$ prior distributions for the components of $\boldsymbol{\theta}$, all independent among them; by combining the likelihood (2.5) with the joint prior of $\boldsymbol{\theta}$ we obtain the joint posterior and full conditionals, which can then be used in a Gibbs sampler, and which are given by

$$\begin{aligned} \xi_v \mid \boldsymbol{X}_{\xi_v} &\sim \mathrm{Beta}(\alpha_{\xi_v}, \beta_{\xi_v}), \\ \mathrm{se}_{l,v} \mid \boldsymbol{X}_{\mathrm{se}_{l,v}} &\sim \mathrm{Beta}(\alpha_{\mathrm{se}_{l,v}}, \beta_{\mathrm{se}_{l,v}}), \\ \mathrm{sp}_{l,v} \mid \boldsymbol{X}_{\mathrm{sp}_{l,v}} &\sim \mathrm{Beta}(\alpha_{\mathrm{sp}_{l,v}}, \beta_{\mathrm{sp}_{l,v}}), \end{aligned} \tag{2.6}$$

where, $\boldsymbol{X}_{\xi_v} = \{a_{\xi_v}, b_{\xi_v}, y_{i,v}, n_v\}$,

$$\boldsymbol{X}_{\mathrm{se}_{l,v}} = \{a_{\mathrm{se}_{l,v}}, b_{\mathrm{se}_{l,v}}, t_{i,l,v}, y_{i,v},\} \ , \qquad \boldsymbol{X}_{\mathrm{sp}_{l,v}} = \{a_{\mathrm{sp}_{l,v}}, b_{\mathrm{sp}_{l,v}}, t_{i,l,v}, y_{i,v}\} \ ,$$

and

$$\alpha_{\xi_v} = \sum_{i=1}^{n_v} y_{i,v} + a_{\xi_v} \ , \qquad \beta_{\xi_v} = n_v - \sum_{i=1}^{n_v} y_{i,v} + b_{\xi_v} \ ,$$

$$\alpha_{\mathrm{se}_{l,v}} = \sum_{i=1}^{n_v} t_{i,l,v}\, y_{i,v} + a_{\mathrm{se}_{l,v}} \ , \qquad \beta_{\mathrm{se}_{l,v}} = \sum_{i=1}^{n_v} (1-t_{i,l,v})\, y_{i,v} + b_{\mathrm{se}_{l,v}} \ ,$$

$$\alpha_{\mathrm{sp}_{l,v}} = \sum_{i=1}^{n_v} (1-t_{i,l,v})\,(1-y_{i,v}) + a_{\mathrm{sp}_{l,v}} \ , \qquad \beta_{\mathrm{sp}_{l,v}} = \sum_{i=1}^{n_v} t_{i,l,v}(1-y_{i,v}) + b_{\mathrm{sp}_{l,v}} \, .$$

3. SIMULATION STUDY

We consider a simulation study to compare the performance of our model with the Hui–Walter model. Following Georgiadis *et al.* (2003), we simulate data according to the following steps.

Step 1. Calculate the probabilities for each combination of outcomes of the L tests under investigation in vth stratum, given the health condition of a subject, $D \in \{0, 1\}$, i.e.

$$\begin{aligned} &P_{v|D=1}\Big(T_{1,v} = t_{1,v},\, ...,\, T_{L,v} = t_{L,v} \mid D = 1\Big) \, , \\ &P_{v|D=0}\Big(T_{1,v} = t_{1,v},\, ...,\, T_{L,v} = t_{Lv} \mid D = 0\Big) \, . \end{aligned} \tag{3.1}$$

Step 2. Calculate the amount of $X_{v|D}$ elements for each combination of outcomes of the L tests under investigation in vth stratum, given the health condition of a subject, $D \in \{0, 1\}$,

$$\begin{aligned} \mathrm{E}(X_{v|D}) \ = \ n_v \Big\{ &\xi_v \, P_{v|D=1}\Big(T_{1,v} = t_{1,v},\, ...,\, T_{L,v} = t_{L,v} \mid D = 1\Big) \\ &+ (1-\xi_v)\, P_{v|D=0}\Big(T_{1,v} = t_{1,v},\, ...,\, T_{L,v} = t_{L,v} \mid D = 0\Big)\Big\} \, . \end{aligned} \tag{3.2}$$

For the structure of conditional independence we have conditional probabilities (3.1) given by

$$\begin{aligned} P_{v|D=1}\Big(T_{1v} = t_{1,v},\, ...,\, T_{L,v} = t_{L,v} \mid D = 1\Big) \ &= \ \prod_{l=1}^{L} \mathrm{se}_{l,v}^{\,t_{i,l,v}} (1-\mathrm{se}_{l,v})^{1-t_{i,l,v}} \, , \\ P_{v|D=0}\Big(T_{1v} = t_{1,v},\, ...,\, T_{L,v} = t_{L,v} \mid D = 0\Big) \ &= \ \prod_{l=1}^{L} \mathrm{sp}_{l,v}^{\,1-t_{i,l,v}} (1-\mathrm{sp}_{l,v})^{t_{i,l,v}} \, . \end{aligned} \tag{3.3}$$

Table 1: Settings under which data were simulated; here ξ denotes prevalence, whereas 'se' and 'sp' denote sensitivity and specificity. Data have been simulated with the following sample sizes: $n = 50, 100, 500, 1000$.

`Configuration (CONF)`	Stratum (v)								
	1			2			3		
	ξ_1	$\text{se}_{l,1}$	$\text{sp}_{l,1}$	ξ_2	$\text{se}_{l,2}$	$\text{sp}_{l,2}$	ξ_3	$\text{se}_{l,3}$	$\text{sp}_{l,3}$
I	0.30	0.93	0.99	0.70	0.99	0.93	0.50	0.95	0.95
II	0.35	0.93	0.99	0.65	0.99	0.93	0.50	0.95	0.95
III	0.40	0.93	0.99	0.60	0.99	0.93	0.50	0.95	0.95

We have compared the performance of two particular cases of our model: MODEL I (Hui–Walter stratification) and MODEL II (Hui–Walter extended stratification).

Table 2: AIC, BIC, and DIC for MODEL I and MODEL II, according to the settings in Table 1.

	`Configuration (CONF)`	n	AIC	BIC	DIC
MODEL I	I	50	1605.4	1626.5	2018.6
		100	3701.5	3727.4	4555.8
		500	24551.4	24588.6	28832.8
		1000	54053.8	54095.9	62572.0
	II	50	1587.0	1608.0	1955.9
		100	2471.8	2497.7	2583.8
		500	23544.8	23581.9	2743.7
		1000	51669.2	51711.3	59633.8
	III	50	1499.6	1520.7	1835.6
		100	3389.0	3414.9	4083.2
		500	21695.6	21732.7	25094.1
		1000	47547.6	4759.6	54538.5
MODEL II	I	50	1065.2	1146.5	2524.1
		100	2565.6	2665.6	6393.4
		500	11676.3	11819.7	15908.6
		1000	27800.0	27962.1	36265.5
	II	50	1094.4	1175.7	1395.8
		100	2578.1	2678.1	3279.6
		500	17765.8	17909.2	21554.1
		1000	40106.1	40268.3	47881.8
	III	50	1169.7	1251.0	1451.0
		100	2660.1	2760.1	3298.8
		500	17623.7	17767.2	21019.9
		1000	39340.7	39502.8	46184.2

Two MCMC parallel chains of 50.000 iterations were generated from posterior conditionals (2.6), discarding the first 5.000 iterations (burn-in) of each chain; after thining, we were left with a posterior sample of size $n = 2.000$. The convergence of posterior conditionals (2.6) to the posterior marginals of $\boldsymbol{\theta}$, was monitored by using the potential scale reduction factor (R) (Gelman and Rubin, 1992), and the posterior marginals were graphically evaluated in terms of symmetry, unimodality, and variability of estimates based on the amplitude of 95% credibility interval and mean standard errors. The AIC, BIC, and DIC criteria were used to evaluate the performance of the models (Iliopoulos *et al.*, 2007), and according to these criteria our model (MODEL II) overperforms MODEL I; see Table 2.

We observe estimates with smaller standard error as we increase the sample size and/or absolute mean difference in disease prevalence rates between the strata, with slightly smaller rates of sensitivity ($\text{se}_{l,v}$) and specificity ($\text{sp}_{l,v}$) being found in more prevalent and less prevalent strata, respectively; see Figure 2.

Despite presenting a slightly larger standard error to that of MODEL I, our model (MODEL II) had stationary marginals and estimates very close to the true ones; in addition, we note that sensitivity and specificity are always overestimated with MODEL I; see Figure 3.

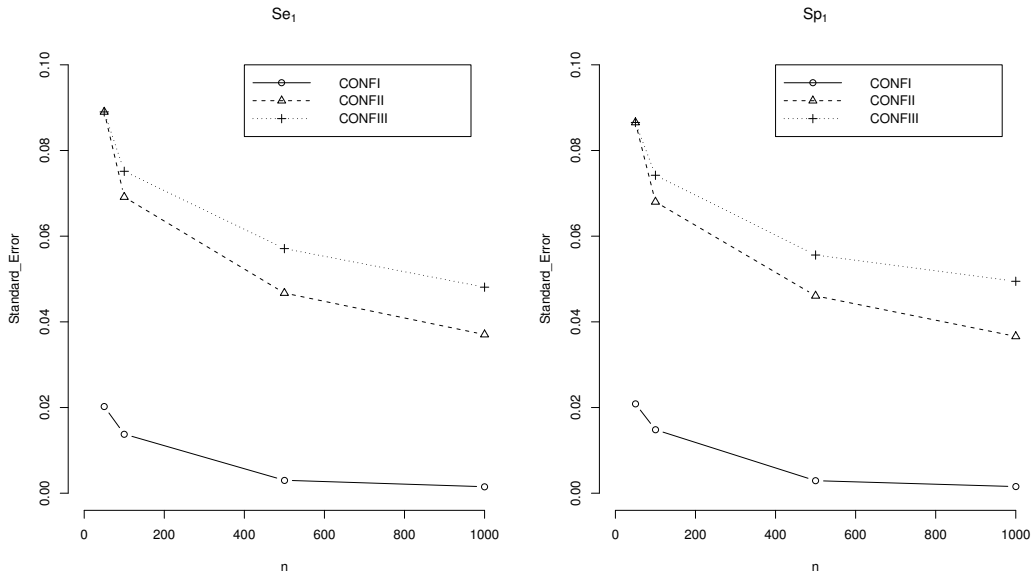

Figure 1: Standard error ($\times 10^{-2}$) to the sensitivities and specificities of the first test in MODEL I according to the settings in Table 1.

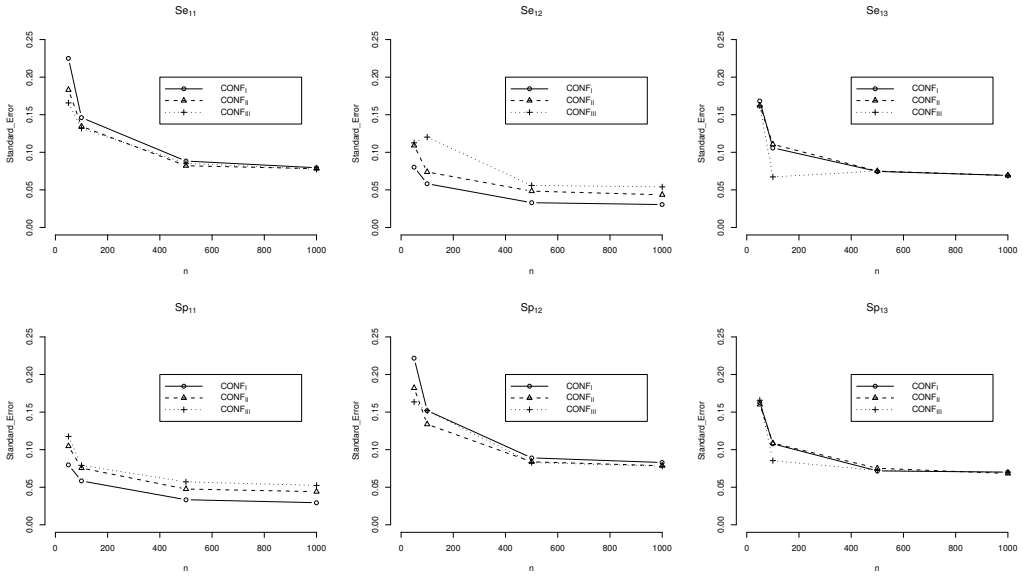

Figure 2: Standard error ($\times 10^{-2}$) to the sensitivities and specificities of the first test in MODEL II according to the settings in Table 1.

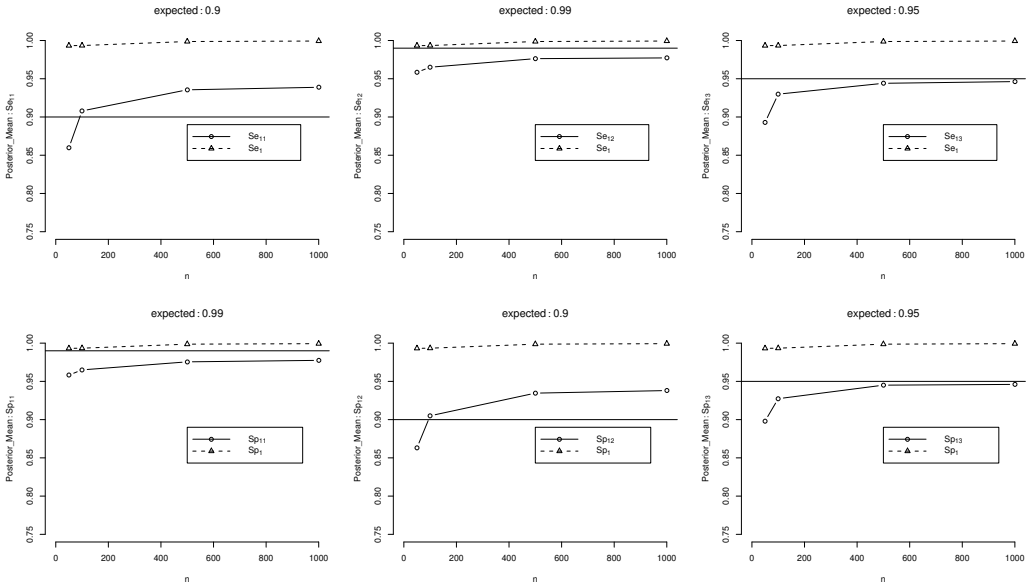

Figure 3: Posterior mean to the sensitivities and specificities of the first test in MODEL I and MODEL II according to the settings in Table 1.

4. ILLUSTRATION ON CHAGAS DISEASE DATA

We now consider an illustration using a Chagas disease case study in Brazil. The data were gathered from 238 blood donors attending a blood center in the region of Triângulo Mineiro, Brazil, who were randomly selected from two groups with different prevalences. Stratum I consists of 29 samples from blood bank donors with positive serology in three conventional serological reactions for Chagas' disease (positive control), and 30 blood samples with five or more negative donations (negative control). Stratum II consists of 179 samples from blood bank donors collected between 2005 and 2008, whose values were low, or within the region denominated 'gray zone' $\pm$ 20% of the reactivity threshold (undetermined serology). Several commercially available kits have been used to determine the diagnostic performance of the four tests, namely: One immunoblotting `TB` (TESA-blot), and three ELISA-based tests, viz.: ELISA Wienner total extract from the subclass `IgG1`, `E-BIO` (ELISA BioMérieux) and `E-WIE` (ELISA Winner recombinant).

Table 3: Results of four serological tests in two subgroups of blood donors.

Test				Group	
IgG1	E-BIO	E-WIE	TB	Control	Inc.
−	−	−	−	30	78
−	−	−	+	0	1
−	−	+	−	0	13
+	−	−	−	0	11
+	−	+	−	0	18
+	−	+	+	0	1
+	+	+	+	29	57
Total:				59	179

`IgG1`: ELISA Wienner total extract from the subclass `IgG1`;
`E-BIO`: ELISA BioMérieux;
`E-WIE`: ELISA Winner recombinant;
`TB`: *Imunoblotting* TESA-blot;
`Control`: negative and positive serology;
`Inc.`: inconclusive in screening serology;
'−': negative result;
'+': positive result.

In Table 4 we report the AIC, BIC, and DIC, for MODEL I and MODEL II; similarly to the simulation study in §3, we observe here that our model overperforms the Hui–Walter model.

Table 4: AIC, BIC, and DIC for MODEL I (Hui–Walter stratification) and MODEL II (Hui–Walter extended stratification); for purposes of presentation each of the entries was multiplied by $\times 10^{-4}$.

MODEL I				MODEL II			
p	DIC	BIC	AIC	p	DIC	BIC	AIC
10	67.4	52.0	52.0	18	36.9	31.4	31.4

In Table 5 we present the estimates obtained from the application of our model by using the group serology strata defined above.

Table 5: Estimates obtained from the application of our model by using group serology strata (Stratum I and Stratum II).

	Test	`Control`			`Inc.`		
		Mean	2.5 Pc	97.5 Pc	Mean	2.5 Pc	97.5 Pc
Sensitivity	`IgG1`	96.94	89.52	99.91	98.39	94.02	99.97
	`E-BIO`	96.88	88.52	99.90	96.64	90.97	99.71
	`E-WIE`	96.73	88.14	99.92	98.30	93.31	99.95
	`TB`	96.66	88.47	99.94	98.14	93.07	99.95
Specificity	`IgG1`	96.92	89.93	99.93	79.21	73.03	85.35
	`E-BIO`	96.92	88.84	99.90	99.34	97.53	99.98
	`E-WIE`	96.92	89.16	99.91	79.19	72.47	85.35
	`TB`	96.95	90.05	99.91	98.63	96.40	99.83
Prevalence		49.30	36.22	61.74	28.19	22.64	34.09

Pc: percentile;
`IgG1`: ELISA Wienner total extract from the subclass `IgG1`;
`E-BIO`: ELISA BioMérieux;
`E-WIE`: ELISA Winner recombinant;
`TB`: *Imunoblotting* TESA-blot;
`Control`: negative and positive serology;
`Inc.`: inconclusive in screening serology.

ACKNOWLEDGMENTS

We thank the Brazilian organizations FAPESP, CAPES and CNPq for financial support.

REFERENCES

Berkvens, D.; Speybroeck, N.; Praet, N.; Adel, A. and Lesaffre, E. (2006). Estimating disease prevalence in a Bayesian framework using probabilistic constraints, *Epidemiology*, **17**, 145–153.

Bertrand, P.; Bénichou, J.; Grenier, P. and Chastang, C. (2005). Hui and Walter's latent-class reference-free approach may be more useful assessing agreement than diagnostic performance, *Journal of Clinical Epidemiology*, **58**, 688–700.

Branscum, A. J.; Gardner, I. A. and Johnson, W. O. (2005). Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling, *Preventive Veterinary Medicine*, **68**, 145–163.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings Algorithm, *The American Statistician*, **49**, 327–335.

Dempster, A. P.; Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Ser.* B, **39**, 1–38.

Georgiadis, M. P.; Johnson, W. O.; Gardner, I. A. and Singh, R. (2003). Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests, *Journal of the Royal Statistical Society, Ser.* C, **52**, 63–76.

Gardner, I. A. (2004). An epidemiologic critique of current microbial risk assessment practices: the importance of prevalence and test acuracy data, *Journal of Food Protection*, **67**, 2000–2007.

Garrett, E. S. and Zeger, S. L. (2000). Latent class model diagnosis, *Biometrics*, **56**, 1055–1067.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of American Statistical Association*, **85**, 398–409.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika*, **61**, 215–231.

Gustafson, P. (2005). The utility of prior informative and stratification for parameter estimation with two screening tests but no gold standard, *Statistics in Medicine*, **24**, 1203–1217.

Hadgu, A.; Dendukuri, N. and Hielden, J. (2005). Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues, *Epidemiology*, **16**, 604–612.

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

Iliopoulos, G.; Kateria, M. and Ntzoufras, I. (2007). Bayesian order-restricted association models for a two-way contingency table, *Computational Statistics and Data Analysis*, **51**, 4643–4655.

Johnson, W. O.; Gastwirth, J. L. and Pearson, L. M. (2001). Screening without a 'gold standard':the Hui–Walter paradigm revisited, *American Journal of Epidemiology*, **153**, 921–924.

Jones, G.; Johnson, W. O.; Hanson, T. E. and Christensen, R. (2009). Identifiability of models for multiple diagnostic testing in the absence of a gold standard, *Biometrics*, **66**, 855–863.

Joseph, L.; Gyorkos, T. W. and Coupal, L. (1995). Bayesian estimation of disease prevalence and parameters for diagnostic tests in the absence of a gold standard, *American Journal of Epidemiology*, **141**, 263–272.

Kraemer, H. C. (1992). *Evaluating Medical Tests*, Lavoisier Librairie, Bervely Hills.

Lindley, D. V. (1971). *Bayesian Statistics: A Review*, Society for Industrial and Applied Mathematics, Philadelphia.

Neath, A. A. and Samaniego, F. J. (1997). On the efficacy of Bayesian inference for nonidentifiable models, *The American Statistician*, **51**, 225–232.

Nielsen, S. S.; Grønbaek, C.; Agger, J. F. and Houe, H. (2002). Maximum-likelihood estimation of sensitivity and specificity of ELISAs and faecal culture for diagnosis of paratuberculosis, *Preventive Veterinary Medicine*, **53**, 191–204.

Singer, R. S.; Boyce, W. M.; Gardner, I. A.; Johnson, W. O. and Fisher, A. S. (1998). Evaluation of bluetongue virus diagnostic tests in free ranging bighorn sheep, *Preventive Veterinary Medicine*, **35**, 265–282.

Swartz, T.; Haitovsky, Y.; Vexler, A. and Yang, T. (2004). Bayesian identifiability and misclassification in multinomial data, *The Canadian Journal of Statistics*, **32**, 1–18.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, **82**, 528–540.

Toft, N.; Jørgensen, E. and Højsgaard, S. (2005). Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard, *Preventive Veterinary Medicine*, **68**, 19–33.

Toft, N.; Innocent, G. T.; Gettinby, G. and Reid, S. W. J. (2007). Assessing the convergence of Markov Chain Monte Carlo methods: An example from evaluation of diagnostic tests in absence of a gold standard *Preventive Veterinary Medicine*, **79**, 244–256.

Walter, S. D. and Irwing, L. M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a Review, *Journal of Clinical Epidemiology*, **41**, 923–937.