Research Article

Statistics in Medicine

Received 30 March 2009, Accepted 3 June 2010 Published online 25 August 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4018

A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies

Ying Lu,[a] Nandini Dendukuri,[b,c*†] Ian Schiller[d] and Lawrence Joseph[e,f]

Verification bias arises in diagnostic test evaluation studies when the results from a first test are verified by a reference test only in a non-representative subsample of the original study subjects. This occurs, for example, when inclusion probabilities for the subsample depend on first-stage results and/or on a covariate related to disease status. Reference standard bias arises when the reference test itself has imperfect sensitivity and specificity, but this information is ignored in the analysis. Reference standard bias typically results in underestimation of the sensitivity and specificity of the test under evaluation, since subjects that are correctly diagnosed by the test can be considered as misdiagnosed owing to the imperfections in the reference standard. In this paper, we describe a Bayesian approach for simultaneously addressing both verification and reference standard bias. Our models consider two types of verification bias, first when subjects are selected for verification based on initial test results alone, and then when selection is based on initial test results and a covariate. We also present a model that adjusts for a third potential bias that arises when tests are analyzed assuming conditional independence between tests, but some dependence exists between the initial test and the reference test. We examine the properties of our models using simulated data, and then apply them to a study of a screening test for dementia, providing bias-adjusted estimates of the sensitivity and specificity.

Keywords: Bayesian; verification bias; reference standard bias; diagnostic tests; non-identifiability

1. Introduction

Verification bias is a frequent concern in studies designed to estimate the properties of diagnostic tests. Suppose that such a study is to be carried out in two stages, a first stage where the diagnostic test being evaluated is given to all participants, and a second stage where a reference test with known properties is given, but only to a subset of subjects from the first stage. This situation may occur, for example, when the second-stage test is expensive or invasive, and so its use is to be minimized. If the subsample is randomly selected from the initial sample there will usually be no verification bias. However, if the second-phase sampling probabilities depend either on the diagnostic test results from the first phase and/or on a covariate which may be related to the subject's true disease status, then bias may occur.

To illustrate the problem, consider a study where data were collected to evaluate a screening test for dementia among subjects aged 65 years or older [1, 2]. The study followed the two-stage design illustrated in Table I. In Stage I all patients were evaluated by The Community Screening Instrument for Dementia [3]. To create this table, we simplified the original data set, wherein patients with a 'Very Poor' or 'Poor' rating were grouped together into the screen positive category, whereas those rated 'Intermediate' or 'Good' formed the screen negative category. This reflects clinical reality, where a diagnosis must be made for each subject for purposes of treatment or further follow-up. Patients were selected

[a]*# 20E 51/2-110 Soi Sukhumvit 49, Sukhumvit Road, Klongton-Nua, Wattana, Bangkok 10110, Thailand*
[b]*Technology Assessment Unit, McGill University Health Centre, 687 Pine Avenue West, R4.09, Montreal, Que., Canada H3A 1A1*
[c] *Department of Medicine, McGill University Health Centre, 687 Pine Avenue West, A3.09, Montreal, Que., Canada H3A 1A1*
[d]*Division of Clinical Epidemiology, McGill University Health Center, 687 Pine Avenue West R4.05, Montreal, Que., Canada H3A 1A1*
[e]*Division of Clinical Epidemiology, McGill University Health Center, 687 Pine Avenue West, V Building, Room V2.10, Montreal, Que., Canada H3A 1A1*
[f]*Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montreal, Que., Canada H3A 1A2*
[*]*Correspondence to: Nandini Dendukuri, Department of Medicine, McGill University Health Centre, 687 Pine Avenue West, A3.09, Montreal, Que., Canada H3A 1A1.*
[†]*E-mail: nandini.dendukuri@mcgill.ca*

Table I. Data from Rodenberg and Zhou (2000) on 2212 patients screened for dementia. The screening test was The Community Screening Instrument for Dementia, and the reference clinical diagnosis was based on the Cambridge Mental Disorders of the Elderly Examination and the Consortium to Establish a Registry for Alzheimer's Disease.

		Greater than 75 years old		Less than 75 years old	
Stage		Screening test +	Screening test −	Screening test +	Screening test −
Stage I	Total	255	650	174	1133
	Unverified by reference test	99	578	78	1106
	Verified by reference test	156	72	96	27
Stage II	Reference test +	54	1	10	0
	Reference test −	102	71	86	27

for Stage II of the study based on their performance on the screening test, with a screen positive subject more likely to be selected for further testing than a screen negative subject, and on the covariate age, with older subjects more likely to be selected. The reference standard clinical diagnosis of dementia was made following an interview and clinical exam, in part based on the Cambridge Mental Disorders of the Elderly Examination and on a structured cognitive evaluation from the Consortium to Establish a Registry for Alzheimer's Disease (CERAD).

Previous analyses of these data [2] assumed the reference clinical diagnosis to be a gold standard test, with sensitivity and specificity both equalling 100 per cent. There is much evidence in the literature, however, that the clinical diagnosis of dementia is far from perfect. For example, a recent review of studies comparing clinical diagnosis to postmortem neuropathology reported that the sensitivity of clinical diagnosis ranged from 39 to 98 per cent, while the specificity ranged from 33 to 100 per cent. Similar results have also been reported in other studies [4, 5]. Since it is well known that not adjusting for an imperfect reference test can very strongly bias estimates of sensitivity and specificity even in the absence of verification bias [6], it is important to adjust for this additional factor. To our knowledge, no model has been developed that simultaneously adjusts for verification bias and the lack of a gold standard reference test.

Verification bias was first described over two decades ago, and various solutions have appeared in the literature, mainly from a frequentist point of view [7–9]. While these methods are useful, they are subject to serious limitations. For example, one typically must assume that the second-stage test is either a perfect gold standard or has exactly known test properties [10], neither assumption being realistic in the vast majority of situations. This is because if one tries to incorporate all uncertainties into the model, including unknown properties for the reference standard test, then the model becomes non-identifiable, there being more parameters to estimate than there are degrees of freedom in the data.

Recently, two articles addressed the problem using Bayesian approaches [11, 12], but both models assumed that the reference test was a perfect gold standard. Further, they proposed models where the verification probabilities depend on the results of both the test under evaluation and the reference test, in addition to covariates. Martinez *et al.* [11] used an empirical Bayes approach where the prior distribution was determined in part by maximum likelihood estimates obtained from a frequentist solution. Buzoianu and Kadane's model [12] used a missing data approach to impute unobserved reference test results for the unverified subjects with prior distributions based on expert opinion. The model described in this paper reduces to that of Buzoianu and Kadane if the second-stage test is assumed to be a perfect reference standard and verification bias is ignorable.

In this paper, we model the sensitivity and specificity of the reference test as additional unknown parameters. Further, we allow the verification probabilities to depend only on variables known at Stage I, depending on the reference test only indirectly, via a covariate. A Bayesian approach is used for inference, and as has often been suggested in the diagnostic testing literature [6, 13–15], posterior inferences are obtained by placing informative prior distributions on a minimum number of parameters.

In Section 2 we describe the most basic situation where the verification probability is a function of the result of the first-stage screening test only, deferring discussion of adding a covariate to the model to Section 3. Whether there is a covariate or not, we allow the reference test to be imperfect, and simultaneously estimate all unknown parameters. In Section 4 we investigate the performance of our methods in simulated data sets to illustrate the effect of various design variables and the prior distribution on the estimated parameters. We return to the problem of estimating the properties of our screening test for dementia in Section 5. We use this example to compare estimates derived from models assuming conditional independence between tests or not. We provide some concluding remarks in Section 6.

2. A Bayesian approach to correcting for verification bias

Suppose that a study is designed to estimate the properties of a diagnostic test (T_1) for a certain disease or condition, and that this test is to be compared with another possibly imperfect reference test (T_2). At a first stage, T_1 is

Table II. A two-stage design for diagnostic test evaluation.

Stage of the study		$T_1=1$	$T_1=0$
Stage I	Total	t_1	t_0
	Unverified on T_2	u_1	u_0
	Verified on T_2	t_1-u_1	t_0-u_0
Stage II	$T_2=1$	d_{11}	d_{01}
	$T_2=0$	d_{10}	d_{00}

One observes t_1 and t_0 positive and negative subjects on test T_1 at Stage I, but only t_1-u_1 and t_0-u_0 positive and negative T_1 subjects, respectively, are selected to be further evaluated by test T_2 at Stage II of the design. The data d_{ij}, i, $j=0,1$ represent the numbers of subjects who test positive (1) or negative (0) on each of the two tests.

applied to a random sample of subjects from the target population. Subsequently, a random sample of these subjects is verified by T_2, with the sampling probabilities possibly dependent on the results of T_1. Let V denote a verification indicator such that $V=1$ if the subject is verified by T_2, and $V=0$ otherwise. Unless the selection probabilities for second-stage subject verification do not depend on T_1, i.e. $P(V=1|T_1)=P(V=1)$, then there is the possibility of verification bias.

Let p_{11} be the probability of testing positive on T_1, and assume that the probability of verification for each subject depends only on their results on T_1. Let p_{V1} be the probability of being verified among individuals who are positive on T_1, and p_{V0} be the probability of being verified among individuals who are negative on T_1. Let p_{21} denote the probability of being positive on T_2 given T_1 is positive and the subject is selected for verification. Similarly, let p_{20} denote the probability of testing positive on T_2 given T_1 is negative and the subject is selected for verification. The statistical model may be summarized using a hierarchical notation as

$$T_1 \sim \text{Bernoulli}\ (p_{11}),$$

$$V|T_1=j \sim \text{Bernoulli}\ (p_{Vj}), \quad j=0,1 \quad \text{and}$$

$$T_2|V=1,T_1=j \sim \text{Bernoulli}(p_{2j}), \quad j=0,1.$$

The observed data can be summarized as described in Table II. Let the number of subjects testing positive ($T_1=1$) or negative ($T_1=0$) on the first test be given by t_1 and t_0, respectively. The t_j subjects, $j=0,1$, can be further subdivided into d_{j1} who were verified and tested positive on T_2, d_{j0} who were verified and tested negative on T_2, and u_j who were not verified, so that $d_{j1}+d_{j0}+u_j=t_j$.

The full likelihood function can be written as the product of the likelihood contribution at each stage, giving

$$\begin{aligned} &L(p_{11},p_{V1},p_{V0},p_{21},p_{20}|t_1,t_0,u_1,u_0,d_{11},d_{10},d_{01},d_{00}) \\ &\propto p_{11}^{t_1}(1-p_{11})^{t_0}\times\{p_{V1}^{(d_{11}+d_{10})}(1-p_{V1})^{u_1}\times p_{V0}^{(d_{01}+d_{00})}(1-p_{V0})^{u_0}\}\times\{p_{21}^{d_{11}}(1-p_{21})^{d_{10}}\times p_{20}^{d_{01}}(1-p_{20})^{d_{00}}\}. \end{aligned} \quad (1)$$

The probabilities of testing positive or negative on each test at Stage I and Stage II can be expressed as functions of the prevalence of the condition of interest (π), and the properties (sensitivity S_i and specificity C_i) of each test i, $i=1,2$. For the Stage I sample we have

$$\begin{aligned} p_{11}=P(T_1=1)&=P(D=1)P(T_1=1|D=1)+P(D=0)P(T_1=1|D=0) \\ &=\pi S_1+(1-\pi)(1-C_1), \end{aligned}$$

where D is a binary variable denoting the true (unobserved, or latent) disease status.

We assume that at the second-stage subjects are selected at random from within subsets denoted by $T_1=1$ and $T_1=0$. Since subjects at the second stage were selected randomly, $P(T_2=1|T_1=1,V=1)=P(T_2=1|T_1=1)$ and $P(T_2=1|T_1=0,V=1)=P(T_2=1|T_1=0)$. This implies that the Stage II sample provides unbiased estimates of the predictive values $P(T_2=1|T_1)$ even in the presence of verification bias, which can affect other parameters.

We first assume that T_1 and T_2 are conditionally independent given the true disease status, but later remove this assumption to model possible dependence between T_1 and T_2. Conditional independence implies

$$p_{21}=P(T_2=1|T_1=1)$$

$$= P(T_2=1, D=1|T_1=1)+P(T_2=1, D=0|T_1=1)$$
$$= P(T_2=1|D=1)P(D=1|T_1=1)+P(T_2=1|D=0)P(D=0|T_1=1)$$
$$= S_2\frac{\pi S_1}{\pi S_1+(1-\pi)(1-C_1)}+(1-C_2)\frac{(1-\pi)(1-C_1)}{\pi S_1+(1-\pi)(1-C_1)}. \quad (2)$$

Similarly,

$$p_{20}=P(T_2=1|T_1=0)=S_2\frac{\pi(1-S_1)}{\pi(1-S_1)+(1-\pi)C_1}+(1-C_2)\frac{(1-\pi)C_1}{\pi(1-S_1)+(1-\pi)C_1}. \quad (3)$$

Most of the literature to date [7, 9, 11, 12] assumes that T_2 is a gold standard test with $S_2=C_2=1$. Then equations (2) and (3) reduce to

$$p_{21}=P(T_2=1|T_1=1)=\frac{\pi S_1}{\pi S_1+(1-\pi)(1-C_1)} \quad \text{and} \quad (4)$$
$$p_{20}=P(T_2=1|T_1=0)=\frac{\pi(1-S_1)}{\pi(1-S_1)+(1-\pi)C_1},$$

respectively. However, one could argue that this assumption almost never holds in practice, as few if any diagnostic tests are error-free, and even theoretically perfect tests are rendered imperfect by human error in application and administration. Accurate estimation and full correction for verification bias therefore depend on taking these imperfections into account.

The amount of prior information required for reasonable estimation is determined in large part by the number of degrees of freedom in the observed data. At Stage I we observe the number of positive results on T_1, a binomial random variable contributing 1 degree of freedom. At the verification stage we observe two binomial variables contributing 1 degree of freedom each. Finally, at Stage II, within each of the groups of individuals who are positive or negative on T_1 we observe the number of positive results on T_2. Thus we observe results of two further binomial variables each adding 1 degree of freedom, for a total of 5. However, since the parameters of interest $(\pi, S_1, C_1, S_2, C_2)$ are involved in Stages I and II only, the two degrees of freedom from the verification stage do not contribute useable information towards estimating these parameters, hence the number of degrees of freedom available for this purpose is 3. This renders our model non-identifiable, meaning that in practice substantive prior inputs are required for at least two of these parameters in order to derive reasonable estimates [6, 15]. While considerable care needs to be exercised in selecting prior distributions, and one will usually want to check the robustness of the final estimates across a range of prior choices, this approach seems preferable to falsely assuming T_2 to be a perfect gold standard, which will almost always lead to biased estimates.

While substantive prior information can be input on any two or more parameters, usually S_2 and C_2 are chosen, as the properties of the reference standard should be at least approximately known. In addition, there might sometimes be some knowledge about the prevalence. For example, if the main objective of the study is to estimate the properties of T_1, the design might aim for approximately 50 per cent prevalence, providing equal sample sizes towards estimating sensitivity and specificity. Typically, one would want to run analyses with low information prior distributions on S_1 and C_1, to let the data drive inferences about the main parameters of interest. Two degrees of freedom are available to estimate the verification stage parameters (p_{V1}, p_{V0}), so that non-informative prior distributions can be used for these parameters.

In principle, any joint prior distribution can be used over the set of seven unknown parameters. The Beta (α, β) density is a convenient choice, as it covers the 0,1 range of each parameter, and has a reasonably flexible shape. To determine the values of α and β for a parameter about which some substantive prior knowledge is available, we need information on any two features of the distribution, for example the mean and standard deviation, or the 2.5 and 97.5 per cent quantiles [6]. If a low information prior distribution is needed one can use a Beta $(\alpha=1, \beta=1)$, which corresponds to a uniform prior distribution. For example, we used Beta(1,1) prior distributions for the verification probabilities, p_{V1} and p_{V0}.

Having stated our likelihood function and the form of our prior distributions, posterior densities are derived through Bayes theorem. In this case, there are no simple closed-form formulae, but samples from the marginal posterior distributions for each parameter can be obtained using a Gibbs Sampler as implemented in WinBUGS. Our programs are available from the software section at http://www.medicine.mcgill.ca/epidemiology/dendukuri. For each of the models presented in this paper, we ran 5 Gibbs sampler chains with different initial values. Convergence was determined using the Gelman–Rubin statistic provided by WinBUGS. Once the chains had converged, we used 100 000 iterations of the Gibbs sampler to estimate posterior medians and equal tailed 95 per cent credible intervals for all parameters of interest.

Our model can be extended to accommodate conditional dependence between T_1 and T_2. Using a fixed effects model [13] adds two parameters, representing the covariances between T_1 and T_2 among those who have positive (covp) or negative (covn) latent disease status. Equations (2) and (3) become, respectively,

$$p_{21} = P(T_2=1|T_1=1) = \frac{P(T_2=1, T_1=1)}{P(T_1=1)}$$
$$= \frac{\pi(S_1S_2+\text{covp})+(1-\pi)((1-C_1)(1-C_2)+\text{covn})}{\pi S_1+(1-\pi)(1-C_1)} \quad \text{and} \tag{5}$$
$$p_{20} = P(T_2=1|T_1=0) = \frac{P(T_2=1, T_1=0)}{P(T_1=0)}$$
$$= \frac{\pi((1-S_1)S_2-\text{covp})+(1-\pi)(C_1(1-C_2)-\text{covn})}{\pi(1-S_1)+(1-\pi)C_1}. \tag{6}$$

As in [13] these parameters take on values between 0 (no dependence) and upper limits that are functions of the sensitivity and specificity of the two tests. If the degree of dependence is unknown, uniform prior distributions may be used over these ranges, giving

$$\text{covp} \sim \text{U}(0, \min(S_1, S_2)-S_1S_2)$$
$$\text{covn} \sim \text{U}(0, \min(C_1, C_2)-C_1C_2).$$

Of course, if more information is available concerning the values of covp and covn, narrower prior distributions may be used.

3. Modeling the probability of verification as a function of a covariate

In practice, the probability of disease may depend on one or more covariates, so that in turn, verification decisions may also depend on these covariates. For example, if verification is an expensive or invasive process, then only subjects thought to be at high risk for the disease or condition may be subject to further testing. This risk may in turn depend on a subject's medical history as well as the results of any preliminary tests, especially if the properties of the first test are not well established. Keeping in mind our application to the diagnosis of dementia of Section 5, and for ease of exposition, here we assume a single dichotomous covariate, X, but the methods can be easily extended to continuous covariates, or to the situation when X is a vector of several factors that can influence the verification probabilities. We assume that verification probabilities are related to the result of T_1 and the covariate X through the logistic regression model

$$P(V=1|X, T_1) = \frac{\exp(\beta_0+\beta_1X+\beta_2T_1)}{1+\exp(\beta_0+\beta_1X+\beta_2T_1)}$$

where β_0, β_1 and β_2 are unknown parameters. Similarly, we assume that the prevalence is related to the covariate X so that

$$P(D=1|X) = \pi_X = \frac{\exp(\alpha_0+\alpha_1X)}{1+\exp(\alpha_0+\alpha_1X)}.$$

Let t_{1x} and t_{0x} denote the number of individuals who were positive and negative, respectively, on T_1, within the subgroup $X=x$, $x=0, 1$. Let d_{ijx} denote the number of individuals with results $T_1=i$ and $T_2=j$ within the subgroup $X=x$, $x=0, 1$. With a dichotomous covariate, the data format remains similar to that shown in Table II, except that such a table can now be constructed for each value of the covariate, i.e. we have observed data consisting of t_{1x}, t_{0x}, d_{11x}, d_{10x}, d_{01x} and d_{00x}, for $X=x$, $x=0, 1$.

The likelihood function of the observed data now becomes

$$L(p_{111}, p_{110}, p_{211}, p_{210}, p_{201}, p_{200}, \beta_0, \beta_1, \beta_2, |t_{1x}, t_{0x}, u_{1x}, u_{0x}, d_{11x}, d_{10x}, d_{01x}, d_{00x}, x=0, 1)$$
$$\propto \prod_{X=0}^{1} p_{11x}^{t_{1x}}(1-p_{11x})^{t_{0x}} \times p_{21x}^{d_{11x}}(1-p_{21x})^{d_{10x}} \times p_{20x}^{d_{01x}}(1-p_{20x})^{d_{00x}}$$

$$\times\left\{\frac{\exp(\beta_0+\beta_1X+\beta_2)}{1+\exp(\beta_0+\beta_1X+\beta_2)}\right\}^{d_{11x}+d_{10x}}\left\{\frac{1}{1+exp(\beta_0+\beta_1X+\beta_2)}\right\}^{u_{1x}},$$

$$\times\left\{\frac{\exp(\beta_0+\beta_1X)}{1+\exp(\beta_0+\beta_1X)}\right\}^{d_{01x}+d_{00x}}\left\{\frac{1}{1+\exp(\beta_0+\beta_1X)}\right\}^{u_{0x}} \tag{7}$$

where p_{11x}, p_{21x} and p_{20x}, $x=0,1$, can each be expressed in terms of S_1, C_1, S_2, C_2 and the parameters used to model the prevalence, α_0 and α_1, such that

$$\begin{aligned}p_{11x} &= P(T_1=1|X)\\ &= \frac{\exp(\alpha_0+\alpha_1X)S_1+(1-C_1)}{1+\exp(\alpha_0+\alpha_1X)},\end{aligned}$$

$$\begin{aligned}p_{21x} &= P(T_2=1|T_1=1,V=1,X)\\ &= \frac{\exp(\alpha_0+\alpha_1X)S_1S_2}{\exp(\alpha_0+\alpha_1X)S_1+(1-C_1)}+\frac{(1-C_1)(1-C_2)}{\exp(\alpha_0+\alpha_1X)S_1+(1-C_1)}\end{aligned}\quad\text{and}$$

$$\begin{aligned}p_{20x} &= P(T_2=1|T_1=0,V=1,X)\\ &= \frac{\exp(\alpha_0+\alpha_1X)(1-S_1)S_2}{\exp(\alpha_0+\alpha_1X)(1-S_1)+C_1}+\frac{C_1(1-C_2)}{\exp(\alpha_0+\alpha_1X)(1-S_1)+C_1}.\end{aligned}$$

As in Section 2, the expressions above can be modified in a straightforward manner to model conditional dependence between T_1 and T_2.

We again use Beta prior distributions for the sensitivity and specificity parameters. For the logistic regression parameters α_0, α_1, β_0, β_1 and β_2 we use Normal(μ,σ^2) prior distributions. A suitable low information prior distribution could have $\mu=0$ and a sufficiently large value such as $\sigma=10$. Even though such priors are not nearly uniform, as they place more weight on the extremes of the (0,1) interval, they are very weak compared with the information in the data. Other choices of prior distributions that induce a more uniform prior on the probability scale have been described in [12].

If prior information is available about the prevalence, it can be transformed into prior information on α_0 and α_1. For example, suppose the range of values of π_x, $X=0,1$ is given by (l_x,u_x) and that π_1 is believed to be greater than π_0. We can estimate the lower bound of the range of α_0 by $\alpha_{0l}=\text{logit}(l_0)$ and the upper bound by $\alpha_{0u}=\text{logit}(u_0)$. Similarly, for α_1 we could estimate the lower bound by $\alpha_{1l}=\text{logit}(l_1)-\text{logit}(u_0)$ and the upper bound by $\alpha_{1u}=\text{logit}(u_1)-\text{logit}(l_0)$. These ranges can be converted into estimates for the prior mean and standard deviation of α_0 and α_1. For example, the prior mean can be set equal to the centre of the range, and the prior standard deviation can be set equal to one quarter of the range.

The addition of a covariate to the model means essentially doubling the amount of data available, while only slightly increasing the numbers of parameters to estimate, so that the model will usually be identifiable even when T_2 is not a gold standard test. The number of degrees of freedom available for estimating the six parameters of interest, $(\alpha_0,\alpha_1,S_1,C_1,S_2,C_2)$ now increases from 3 to 6. This is similar to the well-known two-population situation described by Hui and Walter [16], and considered from a Bayesian viewpoint by Johnson *et al.* [17]. This allows the use of non-informative prior distributions across all parameters, if desired. However, as the prevalence in the two strata becomes similar (i.e. as α_1 approaches 0), the problem approaches non-identifiability [15], essentially because there really are not two different populations, as the covariate has no effect on the prevalence. In this case, one must ignore the covariate information and the problem collapses to that described in Section 2, where inference requires informative prior distributions on a minimum of two parameters. If, instead, the effect of the covariate on verification probability is negligible, i.e. parameter β_1 approaches 0, the problem remains identifiable as long as the prevalence remains a function of X.

4. Application to simulated data sets

To illustrate the performance of the models described in Sections 2 and 3, we applied them to a series of simulated data sets. Of course, since our models contain a large number of parameters and the study designs can be associated with a wide range of prevalence values, sample sizes, verification rates and test properties, it is impossible to investigate the performance of these models across all possible scenarios. We therefore selected a range of situations that would typically arise in practice.

In particular, we examined the results of fitting the conditional independence model in Section 2 to data sets that were generated using relatively high and relatively low values for each of the following parameters:

(i) the prevalence: $\pi=0.1$ or 0.4,
(ii) the probability of verification among those testing negative at the first stage: $P(V=1|T_1=0)=0.1$ or 0.6
(iii) the sensitivity of the reference test: $S_2=0.3$ or 0.9
(iv) the specificity of the reference test: $C_2=0.3$ or 0.9
(v) the sample size at Stage I: $N=t_1+t_0=200$, 2000 or 20 000.

For all data sets we selected parameter values such that $S_2+C_2>1$ to ensure a meaningful reference test. We set $S_1=C_1=0.7$ and $P(V=1|T_1=1)=0.9$.

We also examined the consequence of correctly and incorrectly specifying prior information for the model without a covariate. For data sets generated with $S_2=C_2=0.9$, $\pi=0.4$ and $P(V=1|T_1=0)=0.6$, we considered the following situations: (i) treating T_2 as a gold standard, (ii) using prior distributions over S_2 and C_2 whose 95 per cent credible interval covered the true value but was not centred on the correct values, (iii) using prior distributions for S_2 and C_2 that were indeed centred on the true values.

To study the result of fitting the model in the presence of a covariate, we generated data sets assuming that both sensitivity and specificity of the reference test were high (0.9). We let the covariate have a population distribution $P(X=1)=0.3$, and the parameters relating the prevalence to the covariate to have values $\alpha_0=-0.2$ and $\alpha_1=2$. This implies a prevalence in the groups $X=0$ and $X=1$ of 0.45 and 0.86, respectively. We let the probability of verification be dependent on the covariate and the test result, with $\text{logit}(P(V=1|T_1,X))=-0.7-0.69X+2.78T_1$. This implies probabilities of verification in the groups $(X=1,T_1=0)$, $(X=0,T_1=0)$, $(X=1,T_1=1)$ and $(X=0,T_1=1)$ of 0.20, 0.33, 0.80 and 0.89, respectively. We used $\text{N}(\mu=0,\sigma=10)$ priors for the parameters $\beta_0,\beta_1,\beta_2,\alpha_0$ and α_1. We also ran models where α_1 and β_1 were set equal to 0 to study the impact of ignoring adjustment for a covariate that influences the verification probability.

Throughout, uniform Beta(1, 1) prior distributions were used for π, S_1 and C_1. We chose different prior densities for the properties of T_2, depending on whether the true sensitivity and specificity values were high (0.9) or low (0.3), and on whether we wanted to centre the prior density at these true values or not. When the true value was 0.3 we used a Beta(24.9, 58.1) prior density implying a 95 per cent prior range of (0.2, 0.4). When the true value was 0.9 we used Beta(31.5, 3.5) or Beta(42.5, 7.5) densities for centred and non-centred cases, respectively, implying 95 per cent prior ranges of (0.8, 1.0) and (0.75, 0.95), respectively.

In non-identifiable problems, increasing the sample size will often not decrease the width of the posterior credible interval below a certain limiting value [18]. To study the impact of sample size on precision of the posterior estimates in our models, we simulated data sets of size 20 000 for the scenario where the sensitivity and specificity of the reference were high (0.9), verification among $T_1=0$ was high (0.6) and the prevalence was high (0.4).

The results of applying our models to the simulated data sets are given in Tables III and IV. Posterior distributions for the verification probabilities (i.e. p_{V1} and p_{V0} or β_0, β_1 and β_2) were centred over the true values and had high precision for all models (data not shown). We make the following general observations:

1. Parameter C_1 was always estimated with greater precision compared with S_1. This was because the prevalence was less than 0.5 in all scenarios, so that more truly negative subjects were available to provide data on the specificity compared with fewer truly positive subjects providing data for the sensitivity. With higher prevalences, the precision of the estimates of S_1 increased, and estimates for C_1 were less accurate.
2. A higher verification probability among those testing negative on T_1 resulted in a higher precision of the estimates of π and S_1 only in a few scenarios where the sample size was 2000 and both sensitivity and specificity of the reference test were high (Table III).
3. Incorrectly assuming that the reference test was a gold standard resulted in overestimation of the prevalence and underestimation of S_1 (Table IV), compared with the situation where the prior distribution was centred over the correct values, $S_2=0.9$ and $C_2=0.9$ (Table III). The credible intervals were artificially narrow, excluding the true value of the parameter (Table IV).
4. An increase in sample size from $N=200$ to $N=2000$ was associated with a narrowing of the credible intervals for π, S_1 and C_1, particularly when $\pi=0.4$ (Table III). The median values of π, S_1 and C_1 were also more likely to be closer to the true values. The very wide credible intervals for all parameters when $N=200$ or $P(V=1|T_1=0)=0.1$ suggest that a large Stage I sample size and a moderate probability of verification among subjects testing negative on T_1 may be required to obtain reasonable precision.
5. Even when the prior credible intervals over S_2 and C_2 were not centred on the true values, the true values of π, S_1 and C_1 were captured within their respective posterior credible intervals (Table IV). This would suggest that in the situations considered the models we propose are robust to a slight mis-specification of the prior distribution.

Table III. Posterior medians and 95 per cent posterior credible intervals when applying the model to simulated data, with prior credible intervals over S_2 and C_2 centred on true values. The true values for S_1 and C_1 are both 0.7.

True values					Posterior median (95 per cent Credible Interval)				
N	$p(V=1\|T_1=0)$	π	S_2	C_2	π	S_1	C_1	S_2	C_2
200	0.1	0.1	0.3	0.9	0.27 (0.02, 0.84)	0.40 (0.04, 0.96)	0.70 (0.18, 0.91)	0.27 (0.19, 0.39)	0.91 (0.82, 0.98)
200	0.1	0.1	0.9	0.3	0.18 (0.01, 0.79)	0.22 (0.01, 0.92)	0.73 (0.24, 0.86)	0.89 (0.74, 0.97)	0.33 (0.25, 0.42)
200	0.1	0.1	0.9	0.9	0.10 (0.01, 0.28)	0.55 (0.12, 0.97)	0.67 (0.58, 0.75)	0.90 (0.78, 0.97)	0.91 (0.81, 0.97)
200	0.1	0.4	0.3	0.9	0.63 (0.12, 0.97)	0.48 (0.17, 0.92)	0.58 (0.05, 0.95)	0.29 (0.22, 0.39)	0.89 (0.77, 0.97)
200	0.1	0.4	0.9	0.3	0.34 (0.03, 0.87)	0.50 (0.06, 0.96)	0.55 (0.13, 0.88)	0.89 (0.77, 0.97)	0.30 (0.22, 0.39)
200	0.1	0.4	0.9	0.9	0.37 (0.24, 0.56)	0.76 (0.51, 0.98)	0.70 (0.57, 0.81)	0.90 (0.77, 0.97)	0.91 (0.79, 0.97)
200	0.6	0.1	0.3	0.9	0.27 (0.05, 0.57)	0.70 (0.27, 0.98)	0.78 (0.62, 0.97)	0.29 (0.20, 0.39)	0.93 (0.86, 0.98)
200	0.6	0.1	0.9	0.3	0.27 (0.02, 0.84)	0.38 (0.04, 0.94)	0.73 (0.37, 0.94)	0.89 (0.76, 0.97)	0.30 (0.22, 0.39)
200	0.6	0.1	0.9	0.9	0.08 (0.02, 0.18)	0.69 (0.30, 0.98)	0.68 (0.60, 0.75)	0.90 (0.78, 0.97)	0.92 (0.85, 0.97)
200	0.6	0.4	0.3	0.9	0.53 (0.19, 0.92)	0.63 (0.35, 0.96)	0.80 (0.39, 0.99)	0.30 (0.23, 0.40)	0.90 (0.79, 0.97)
200	0.6	0.4	0.9	0.3	0.44 (0.05, 0.92)	0.52 (0.11, 0.94)	0.58 (0.17, 0.93)	0.89 (0.79, 0.97)	0.28 (0.21, 0.38)
200	0.6	0.4	0.9	0.9	0.36 (0.22, 0.50)	0.72 (0.55, 0.94)	0.66 (0.56, 0.77)	0.90 (0.77, 0.97)	0.90 (0.78, 0.97)
2000	0.1	0.1	0.3	0.9	0.20 (0.03, 0.53)	0.60 (0.25, 0.97)	0.72 (0.64, 0.86)	0.28 (0.19, 0.39)	0.92 (0.86, 0.97)
2000	0.1	0.1	0.9	0.3	0.23 (0.01, 0.83)	0.36 (0.04, 0.92)	0.66 (0.30, 0.84)	0.88 (0.74, 0.97)	0.29 (0.24, 0.38)
2000	0.1	0.1	0.9	0.9	0.11 (0.04, 0.20)	0.59 (0.37, 0.96)	0.70 (0.67, 0.73)	0.90 (0.77, 0.97)	0.90 (0.82, 0.97)
2000	0.1	0.4	0.3	0.9	0.47 (0.20, 0.83)	0.69 (0.46, 0.97)	0.74 (0.54, 0.97)	0.28 (0.22, 0.39)	0.90 (0.82, 0.97)
2000	0.1	0.4	0.9	0.3	0.39 (0.24, 0.64)	0.86 (0.63, 0.99)	0.74 (0.61, 0.96)	0.90 (0.80, 0.97)	0.34 (0.26, 0.42)
2000	0.1	0.4	0.9	0.9	0.39 (0.27, 0.51)	0.71 (0.60, 0.91)	0.68 (0.63, 0.75)	0.90 (0.77, 0.97)	0.90 (0.78, 0.97)
2000	0.6	0.1	0.3	0.9	0.20 (0.03, 0.50)	0.54 (0.33, 0.96)	0.72 (0.67, 0.85)	0.28 (0.19, 0.39)	0.91 (0.87, 0.97)
2000	0.6	0.1	0.9	0.3	0.15 (0.01, 0.73)	0.41 (0.05, 0.94)	0.68 (0.48, 0.82)	0.88 (0.73, 0.97)	0.30 (0.26, 0.38)
2000	0.6	0.1	0.9	0.9	0.09 (0.03, 0.18)	0.57 (0.42, 0.95)	0.70 (0.67, 0.72)	0.90 (0.77, 0.97)	0.90 (0.84, 0.97)
2000	0.6	0.4	0.3	0.9	0.44 (0.22, 0.68)	0.74 (0.60, 0.98)	0.73 (0.61, 0.96)	0.29 (0.22, 0.39)	0.91 (0.86, 0.97)
2000	0.6	0.4	0.9	0.3	0.39 (0.18, 0.67)	0.74 (0.56, 0.98)	0.72 (0.62, 0.96)	0.90 (0.82, 0.97)	0.30 (0.25, 0.39)
2000	0.6	0.4	0.9	0.9	0.42 (0.31, 0.52)	0.71 (0.65, 0.85)	0.69 (0.65, 0.76)	0.90 (0.77, 0.97)	0.90 (0.77, 0.97)
20 000	0.6	0.4	0.9	0.9	0.40 (0.30, 0.49)	0.71 (0.66, 0.83)	0.70 (0.68, 0.76)	0.90 (0.78, 0.97)	0.90 (0.79, 0.97)

Of course, this also depends on the precision of the estimates. Our posterior credible intervals tended to be relatively wide, so that it is not surprising that the true values were often inside of these intervals, even with imperfect choice of prior distributions.

6. In the presence of a covariate when $N=2000$, posterior densities were concentrated around the true parameter values even when using low information prior distributions over all parameters (Table IV).
7. The bias due to ignoring the covariate was apparent—informative prior distributions were required to bring the posterior median estimates close to their true values. Even with informative prior distributions the posterior credible intervals remained wider compared with fitting the correct model adjusting for the covariate.
8. When the sample size was 20 000, the posterior credible intervals for π, S_1 and C_1 decreased in width by 30–50 per cent for an identifiable model, that is, those where the gold standard was assumed to be perfect (Table IV). There was a smaller improvement in precision for the non-identifiable models (Table III), demonstrating that increasing the sample size will not necessarily improve the precision of the estimates under these circumstances.

5. Evaluating a screening test for dementia

We now return to the problem of evaluating a screening test for dementia, as discussed in the introduction. The initial analysis [2] assumed that clinical diagnosis is a perfect gold standard test, but as discussed in Section 1, this assumption is unrealistic, with estimates of the sensitivity of clinical diagnosis ranging from 39 to 98 per cent, and specificity estimates ranging from 33 to 100 per cent [19]. There was a strong negative correlation between the sensitivity and the specificity of -0.79 [19] among these studies, showing that raising sensitivity was only at the expense of decreased specificity. We assumed uniform prior distributions over the ranges given above for the sensitivity and specificity of clinical diagnosis, giving $S_2 \sim \mathrm{U}(0.39, 0.98)$ and $C_2 \sim \mathrm{U}(0.33, 1)$. We also considered an alternative bivariate normal prior distribution that allowed for a correlation between sensitivity and specificity, with

$$\begin{pmatrix} \mathrm{logit}(S_2) \\ \mathrm{logit}(C_2) \end{pmatrix} \sim \mathrm{N}\left(\mu = \begin{pmatrix} 1.72 \\ 1.94 \end{pmatrix}, \Sigma = \begin{pmatrix} 1.18 & -1.14 \\ -1.14 & 1.76 \end{pmatrix}\right),$$

Table IV. Posterior medians and 95 per cent posterior credible intervals when applying the model to simulated data: Results for cases when prior credible intervals over S_2 and C_2 were mis-specified or a covariate was involved.

N	Prior on S_2, C_2	$\pi(0.4^*)$	$S_1(0.7)$	$C_1(0.7)$	$S_2(0.9)$	$C_2(0.9)$
Impact of prior mis-specification and sample size in situations where no covariate is involved						
2000	$S_2, C_2 = 1$	0.44 (0.42, 0.47)	0.66 (0.62, 0.69)	0.66 (0.63, 0.69)	—	—
2000	Mis-specified	0.42 (0.29, 0.53)	0.75 (0.67, 0.93)	0.71 (0.66, 0.80)	0.85 (0.72, 0.93)	0.85 (0.74, 0.93)
20 000	$S_2, C_2 = 1$	0.42 (0.41, 0.43)	0.65 (0.64, 0.66)	0.67 (0.665, 0.68)	—	—
20 000	Mis-specified	0.39 (0.27, 0.49)	0.76 (0.68, 0.95)	0.72 (0.69, 0.80)	0.85 (0.72, 0.93)	0.85 (0.74, 0.93)

Covariate adjusted	Prior on S_2, C_2	α_0 (0.41)	α_1 (−1.21)	$S_1(0.7)$	$C_1(0.7)$	$S_2(0.9)$	$C_2(0.9)$
Impact of ignoring a covariate ($N = 2000$)							
Yes	NI	0.50 (−0.23, 1.16)	−1.24 (−1.68, −0.90)	0.66 (0.61, 0.76)	0.66 (0.60, 0.74)	0.90 (0.78, 0.99)	0.91 (0.70, 1.00)
Yes	I	0.49 (0.12, 0.95)	−1.21 (−1.57, −0.88)	0.66 (0.61, 0.72)	0.66 (0.61, 0.72)	0.91 (0.81, 0.97)	0.91 (0.81, 0.98)
No	NI	0.09 (−1.07, 1.28)	—	0.76 (0.61, 0.98)	0.76 (0.61, 0.98)	0.76 (0.65, 0.99)	0.71 (0.56, 0.98)
No	I	0.20 (−0.24, 0.76)	—	0.65 (0.60, 0.74)	0.66 (0.61, 0.78)	0.90 (0.76, 0.97)	0.90 (0.77, 0.97)

For all models, $p(V = 1|T_1 = 0) = 0.6$. * True values of each parameter are given in brackets. NI indicates Non-informative and I indicates Informative prior density were used for S_2 and C_2 centred over the true values, with hyperparameters as discussed in Section 4.

thereby accounting for the correlation between these parameters. Marginally, the ranges given by this bivariate prior density closely match the ranges of the two independent uniform distributions over S_2 and C_2 given above. Since the problem is identifiable, we could theoretically use low-information U(0, 1) priors for all parameters. This was also done as a robustness check.

The screening test is based on a short questionnaire, and while the reference test also employs a more in-depth questionnaire, it is largely based on a clinical examination of the patient. Nevertheless, one cannot be absolutely certain of the assumption of conditional independence between tests. We therefore also fit a model that allows for conditional dependence. In further analyses we studied the relative impact of ignoring verification bias and reference standard bias, by fitting various models where combinations of these biases are ignored. We also considered situations where adjustment for the covariate was ignored.

Table V. Summary of the posterior distributions for all unknown parameters in screening for dementia.

	Prior on S_2 and C_2			
	$S_2 = C_2 = 1$	$S_2, C_2 \sim$ Beta(1,1)	$S_2 \sim$U(0.39, 0.98), $C_2 \sim$U(0.33, 0.99)	Bivariate normal prior*
S_1	0.91 (0.75, 0.99)	0.91 (0.69, 1.00)	0.91 (0.70, 1.00)	0.91 (0.70, 1.00)
C_1	0.85 (0.83,0.86)	0.88 (0.86, 0.91)	0.88 (0.86, 0.91)	0.88 (0.86, 0.91)
S_2	—	0.50 (0.37, 0.66)	0.50 (0.40, 0.66)	0.51 (0.38, 0.66)
C_2	—	0.99 (0.93, 1.00)	0.98 (0.94, 0.99)	0.98 (0.95, 0.99)
α_1	2.32 (1.70, 3.02)	2.38 (1.68, 6.41)	2.45 (1.68, 4.01)	2.40 (1.65, 3.98)
π when age <75	0.01(0.007, 0.02)	0.02 (0.0004,0.06)	0.02 (0.004, 0.05)	0.02 (0.004, 0.06)
π when age $\geqslant 75$	0.12 (0.09,0.15)	0.21 (0.15, 0.30)	0.21 (0.15, 0.29)	0.21 (0.15, 0.29)

*See Section 5.

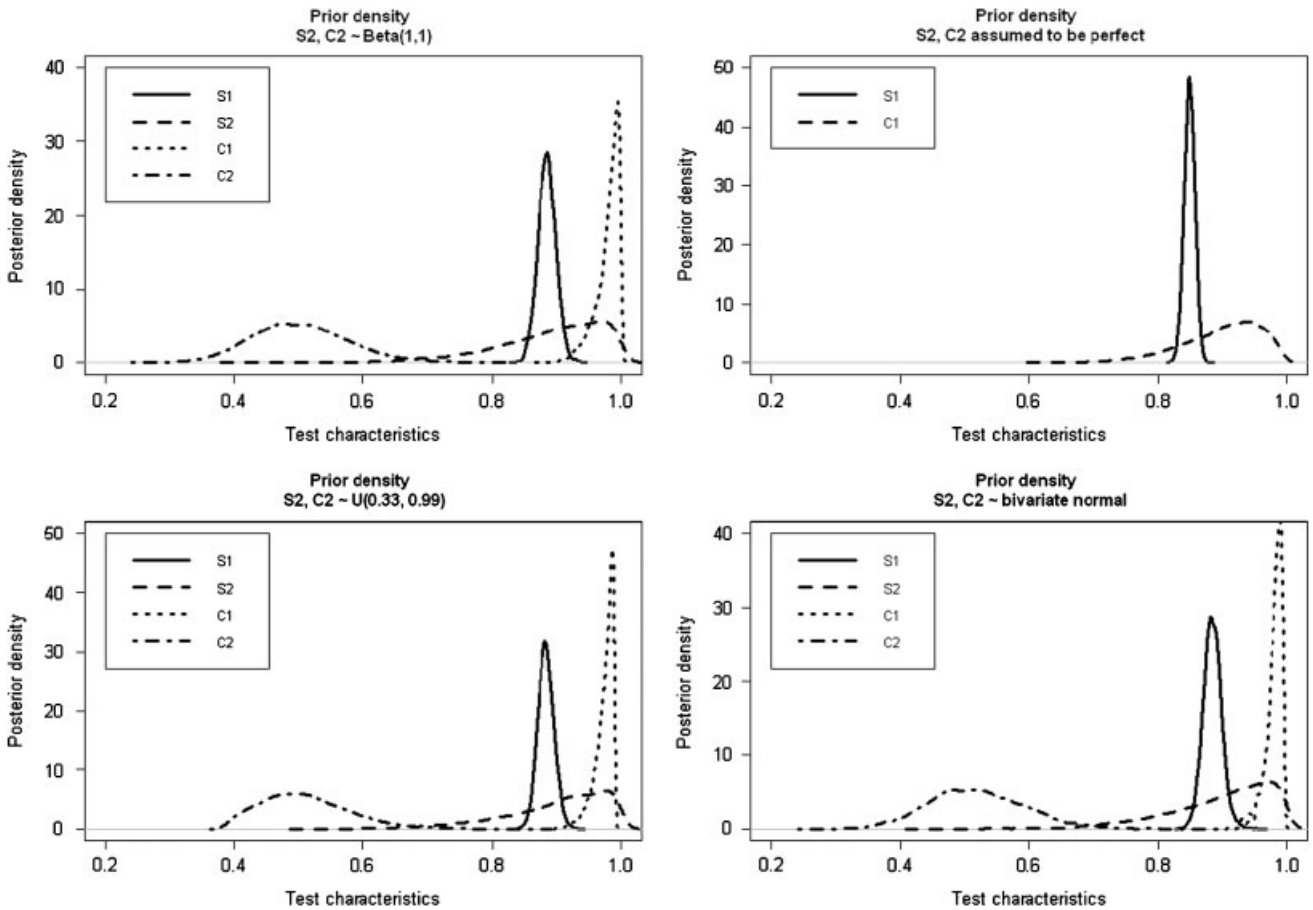

Figure 1. Posterior density functions obtained with different prior distributions over sensitivity and specificity of reference test for dementia.

Table VI. Impact of ignoring various biases on sensitivity and specificity estimates.

Adjustment for							
Verification bias	Imperfect reference	Conditional dependence	Covariate	S_1	C_1	S_2	C_2
No	No	—	Yes	0.98 (0.92, 1.00)	0.34 (0.29, 0.40)	—	—
No	Yes	No	Yes	0.98 (0.92, 1.00)	0.35 (0.29, 0.42)	0.87 (0.59, 0.98)	0.98 (0.93, 0.99)
No	Yes	Yes	Yes	0.61 (0.45, 0.87)	0.23 (0.15, 0.32)	0.47 (0.39, 0.78)	0.92 (0.86, 0.98)
Yes	No	—	Yes	0.91 (0.75, 0.99)	0.85 (0.83,0.86)	—	—
Yes	Yes	No	Yes	0.92 (0.70, 1.00)	0.88 (0.86, 0.91)	0.50 (0.40, 0.66)	0.98 (0.94, 0.99)
Yes	Yes	Yes	Yes	0.88 (0.65, 0.99)	0.87 (0.85, 0.89)	0.47 (0.39, 0.63)	0.98 (0.95, 0.99)
Yes	No	—	No	0.81 (0.56, 0.97)	0.85 (0.83, 0.86)	—	—
Yes	Yes	Yes	No	0.57 (0.02, 0.98)	0.81 (0.79, 0.86)	0.66 (0.40, 0.96)	0.94 (0.91, 0.99)

The impact of adjusting for reference standard bias can be seen in Table V. When treating clinical diagnosis as a perfect reference standard, the sensitivity of the screening test was estimated to be 0.91 (95 per cent credible interval (CrI) 0.75, 0.99) and its specificity was 0.85 (95 per cent CrI 0.83, 0.86). When using an informative prior density the specificity of the screening test somewhat increased, while the credible interval for the sensitivity widened. But the greatest impact was seen in the estimated prevalence, which roughly doubled in both age groups when the reference test was considered as imperfect. This largely occurs because the sensitivity of clinical diagnosis dropped to 50 per cent, whereas its specificity stayed close to 100 per cent. Similar estimates were obtained when using the bivariate normal prior distribution for (S_2, C_2), and the results were also similar whether using non-informative or informative prior distributions. This suggests that the observed data are in agreement with the prior information on sensitivity and specificity of clinical diagnosis reported in earlier studies, and that the sample size was large enough to outweigh the prior information in this identifiable problem. Plots of the posterior density functions of the sensitivity and specificity parameters of the two tests, obtained with the different prior distributions are given in Figure 1.

Table VI summarizes the results of our sensitivity analysis of the impact of ignoring various biases on our estimates. As others have also found, ignoring verification bias causes the estimate of the sensitivity of the screening test to increase, while the specificity decreases. When adjusting for verification bias, further adjustment for conditional dependence appears to only slightly affect the estimates of test properties in this data set, lowering the two sensitivity estimates and increasing the width of their posterior credible intervals, while the two specificity estimates remained virtually unchanged. The posterior median and 95 per cent credible interval of the conditional covariance parameters were: covp = 0.02 (0, 0.10) and covn = 0.01 (0, 0.02). Ignoring the covariate decreased the sensitivity and specificity of the screening test while raising the sensitivity of the reference test.

6. Discussion

We have presented methods that simultaneously adjust for reference standard bias and verification bias when evaluating the properties of a new diagnostic test. By applying them to simulated data sets we have shown our methods to work well, although the nature of the problem is such that credible intervals will tend to be wide with smaller sample sizes [18]. These analyses also show that incorrectly assuming the reference test to be a perfect gold standard can result in biased estimates of the sensitivity and specificity of the test under evaluation, as well as inaccurate prevalence estimates, even when the properties of the reference test are very good. It is therefore important to adjust for even small imperfections in the reference test.

Gustafson [20] discusses how to express non-identified models in two parts, one component with parameters for which the data can be informative, the other component having parameters upon which data have little effect, regardless of the sample size. For the latter set of parameters, no amount of data will lessen the impact of prior information on final inferences. One must then derive a range of plausible prior densities for these parameters, and check robustness of posterior inferences across the different prior choices. Our results emphasize this point by displaying the effect of different choices of prior distributions while retaining the same model and data. We also considered joint prior modeling of sensitivity and specificity along with the usual independent prior choices. Alternatively, one can address identifiability by increasing the amount of data, as occurs in our example when the covariate age is included. Here, the covariate sufficiently increased the amount of data to render our model identifiable, which will happen in practice if the covariate has a large enough effect.

While our study design focuses on the case where $P(T_1)$ and $P(T_2|T_1)$ are estimated from the same study, our methods also apply when these probabilities are unbiasedly estimated from separate studies [21, 22]. While we assumed that the

sensitivities and specificities of the diagnostic tests are not functions of one or more covariates, this restriction can be removed by creating a further hierarchy on these parameters, as we did for the prevalence parameter here. For example, we could model the sensitivity or specificity of T_1 as a logistic function of one or more covariates. In a context where the verification bias is not ignorable, the problem can be handled by expressing the probabilities p_{V1} and p_{V0} in terms of the prevalence, sensitivity and specificity of Test 1. For example, $p_{V1} = P(V=1|T_1=1, D=1)ppv + (P(V=1|T_1=1, D=0)(1-ppv)$, where ppv is the positive predictive value of T_1 as given by equation (4). Of course, this would require more informative prior distributions as we are now adding two more unknown parameters to the model.

References

1. Hendrie HC, Osuntokun B, Hall KS, Ogunniyi A, Hui SL, Unverzagt FW, Gureje O, Rodenberg CA, Baiyewu O, Musick BS, Adeyinka A, Farlow MR, Oluwole SO, Class CA, Komolafe O, Brashear A, Burdine V. Prevalence of alzheimers disease and dementia in two communities: Nigerian Africans and African Americans. *American Journal of Psychiatry* 1995; **152**:1485–1492.
2. Rodenberg C, Zhou XH. Roc curve estimation when covariates affect the verification process. *Biometrics* 2000; **56**:1256–1262.
3. Hall KS, Hendnie HC, Rodgers DD, Prince CS, Pillay N, Blue AW, Bnittain HM, Norton JA, Kaufert JN, Nath A, Shelton P, Osuntokun BO, Postl BD. The development of a dementia screening interview in two distinct languages. *International Journal of Methods in Psychiatric Research* 1993; **3**:1–28.
4. Plassman BL, Khachaturiane AS, Townsend JJ, Ballg MJ, Stephens DC, Leslie CE, Tschanzh JT, Norton MC, Burke JR, Welsh-Bohmer KA, Hulette CM, Nixon RR, Tyrey M, Breitner JC. Comparison of clinical and neuropathologic diagnoses of alzheimers disease in 3 epidemiologic samples. *Alzheimer's and Dementia* 2006; **2**:2–11.
5. Wiederkehr S, Simard M, Fortin C, van Reekum R. Validity of the clinical diagnostic criteria for vascular dementia: a critical review. *Journal of Neuropsychiatry and Clinical Neuroscience* 2008; **20**:162–177.
6. Joseph L, Gyorkos T, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* 1995; **141**:263–272.
7. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983; **39**:207–215.
8. Begg CB. Biases in the assessment of diagnostic tests. *Statistics in Medicine* 1987; **6**:411–432.
9. Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy. *Statistical Methods in Medical Research* 1998; **7**:337–353.
10. Zhou XH. Effects of verification and imperfect reference standard biases on the estimated prevalence rate. *ASA Proceedings of the Joint Statistical Meetings*, Chicago, 1996; 456–461.
11. Martinez ED, Achcar JA, Louzada-Neto F. Estimator's of sensitivity and specificity in the presence of verification bias: a Bayesian approach. *Computational Statistics and Data Analysis* 2006; **51**:601–611.
12. Buzoianu M, Kadane JB. Adjusting for verification bias in diagnostic test evaluation, A Bayesian approach. *Statistics in Medicine* 2008; **27**:2453–2473.
13. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 2001; **57**:158–167.
14. Georgiadis MP, Johnson WO, Singh R, Gardner IA. Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. *Applied Statistics* 2003; **52**:63–76.
15. Gustafson P. On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science* 2005; **20**:111–140.
16. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. *Biometrics* 1990; **36**:167–171.
17. Johnson WO, Gastwirth JL, Pearson LM. Screening without a 'gold standard': the Hui–Walter paradigm revisited. *American Journal of Epidemiology* 2001; **153**:921–924.
18. Dendukuri N, Rahme E, Bélisle P, Joseph L. Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test. *Biometrics* 2004; **60**:388–397.
19. Wollman DE, Prohovnik I. Sensitivity and specificity of neuroimaging for the diagnosis of alzheimers disease. *Dialogues in Clinical Neuroscience* 2003; **5**:89–99.
20. Gustafson P. What are the limits of posterior distributions arising from nonidentified models, and why should we care? *Journal of the American Statistical Association* 2009; **104**:1682–1695.
21. Dendukuri N, McCusker J, Belzile E. The identification of seniors at risk screening tool: further evidence of concurrent and predictive validity. *Journal of the American Geriatrics Society* 2004; **52**:290–296.
22. Dendukuri N, Khetani K, McIsaac M, Brophy J. Testing for HER2-positive breast cancer: a systematic review and cost-effectiveness analysis. *Canadian Medical Association Journal* 2007; **176**:1429–1434.