Bayesian Meta-Analysis of the Accuracy of a Test for Tuberculous Pleuritis in the Absence of a Gold Standard Reference

Nandini Dendukuri,¹∗ Ian Schiller,² Lawrence Joseph,¹ and Madhukar Pai¹

¹Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal H3A 1A2, Canada
²Division of Clinical Epidemiology, McGill University Health Centre, Montreal H3A 1A1, Canada
∗email: nandini.dendukuri@mcgill.ca

SUMMARY. Absence of a perfect reference test is an acknowledged source of bias in diagnostic studies. In the case of tuberculous pleuritis, standard reference tests such as smear microscopy, culture and biopsy have poor sensitivity. Yet meta-analyses of new tests for this disease have always assumed the reference standard is perfect, leading to biased estimates of the new test’s accuracy. We describe a method for joint meta-analysis of sensitivity and specificity of the diagnostic test under evaluation, while considering the imperfect nature of the reference standard. We use a Bayesian hierarchical model that takes into account within- and between-study variability. We show how to obtain pooled estimates of sensitivity and specificity, and how to plot a hierarchical summary receiver operating characteristic curve. We describe extensions of the model to situations where multiple reference tests are used, and where index and reference tests are conditionally dependent. The performance of the model is evaluated using simulations and illustrated using data from a meta-analysis of nucleic acid amplification tests (NAATs) for tuberculous pleuritis. The estimate of NAAT specificity was higher and the sensitivity lower compared to a model that assumed that the reference test was perfect.

KEY WORDS: Bayesian; Bivariate model; Diagnostic test accuracy; Latent class model; Meta-analysis.

1. Introduction

Lack of a gold standard reference test is an acknowledged problem in studies of tuberculosis (TB) diagnostics (Pai et al., 2004). Reference standards for latent TB infection do not exist, and extrapulmonary TB and childhood TB have imperfect reference standards. Most studies resort to using composite reference standards based on a mix of clinical data, microbiological tests, or response to therapy. A recent review of over 40 meta-analyses published so far in the field of TB diagnostics found that none of them adjusted for the bias due to an imperfect reference standard (Pai et al., 2010).

Conveniently assuming the reference test is perfect leads to biased estimates of sensitivity and specificity of the test under evaluation. If the tests are independent conditional on the disease status, the new test’s sensitivity and specificity will be underestimated due to nondifferential misclassification (Walter and Irwig, 1988). Conversely, if the tests are conditionally dependent with positive correlation between them, the new test’s properties may be overestimated (Dendukuri and Joseph, 2001). Though approaches to correct for this bias in individual studies have been discussed, there has been limited attention to the same problem in the context of a meta-analysis. A further problem in meta-analyses is that different primary studies may use different reference standards.

Models for meta-analysis of diagnostic tests have sought to improve over naive univariate pooling of sensitivity or specificity (Macaskill et al., 2010). Moses, Shapiro, and Littenberg (1993) described meta-analysis of the diagnostic odds ratio. The hierarchical summary receiver operating characteristic (HSROC) model of Rutter and Gatsonis (2001) expressed the sensitivity and false positive probability in each study as functions of an underlying bivariate normal model, whereas another bivariate model described by Reitsma et al. (2005) assumed that the vector of (logit(sensitivity), logit(specificity)) itself follows a bivariate normal distribution. Harbord et al. (2007) showed that in the absence of covariates, the likelihood functions of both HSROC and bivariate models are algebraically equivalent, providing identical pooled sensitivity and specificity and between study variance estimates from a frequentist viewpoint. They note that the HSROC construction leads more naturally to a summary receiver operating characteristic curve (SROC), while the model of Reitsma et al. (2005) leads to pooled sensitivity and specificity. The HSROC model has been recommended in the absence of a standard cut-off to define a positive result (Macaskill et al., 2010).

At least three articles have described meta-analysis models for diagnostic tests in the absence of a perfect reference. Walter, Irwig, and Glasziou (1999) proposed a latent class model which assumes the test under evaluation and reference test both measure the same unobservable (latent) variable, the true disease status. They showed that the model is identifiable when assuming sensitivity and specificity remain identical but prevalence varies across studies. In practice, the assumption of identical sensitivity and specificity in all studies is difficult to justify given the variability in population and design aspects of individual studies. Chu, Chen, and Louis

© 2012, The International Biometric Society
(2009) described a more general model where sensitivity and specificity of both the test under evaluation and the reference test, as well as the prevalence, are treated as random effects. It allows for correlation between all four pairs of sensitivity and specificity parameters. This model can be conceptualized as an extension of the model in Reitsma et al. (2005) to the case where there is no gold standard. Sadatsafavi et al. (2010) described a model where one parameter (i.e., sensitivity or specificity of one of the tests) varies across studies. So far, none of the meta-analysis models adjusting for an imperfect reference considered the situation when different reference standards are used in the selected studies. Also, neither of the hierarchical models (Chu et al., 2009; Sadatsafavi et al., 2010), hypothesize the mechanism by which the variation across studies may arise, and accordingly do not provide an SROC. As discussed by Arends et al. (2008), the relation between sensitivity and specificity estimates across studies needs to be specified to determine an SROC.

We discuss here an extension of the HSROC model, which assumes that variation in sensitivity and specificity across studies arises due to use of different cut-off values for defining a positive test and/or differences in diagnostic accuracy, to account for an imperfect reference. The HSROC model is similar in concept to the receiver operating characteristic (ROC) model of Tosteson and Begg (1988). The hierarchical structure accounts for within- and between-study variability.

Section 2 introduces our motivating data set on in-house nucleic acid amplification tests for TB pleuritis. In Section 3, we describe our model and parameter estimation, including how to extract summary statistics and plot an SROC. The performance of the model is investigated through a series of simulations in Section 4, and in Section 5, we apply the model to the TB data. We conclude with a discussion.

2. Evaluating in-House Nucleic Acid Amplification Tests for Tuberculous Pleuritis

Recognition of the global burden of TB has led to renewed interest in combating the disease. Several bodies, including the World Health Organization, have recognized the need for improved diagnostic tests in order to successfully identify patients and prevent further cases (Pai, Ramsay and O’Brien, 2008, Pai et al., 2010). Development of new tests for TB has in turn led to a need for appropriate statistical methods to evaluate them.

Tuberculous pleuritis is an extrapulmonary form of TB that affects the pleural lining of the lungs and causes fluid collection (effusion). Standard tests for this disease are smear microscopy, culture of the pleural fluid and pleural biopsy (histopathological examination). Smear microscopy of the pleural fluid has low sensitivity and is therefore rarely positive (<10%) in pleural TB cases, and culture of the pleural fluid is known to have poor sensitivity ranging from 25% to 58% (Berger and Mejia, 1973; Bueno et al., 1990; Seibert et al., 1991; Valdes et al., 1998; Light, 2010). A composite test based on both biopsy and culture is believed to have nearly 80% sensitivity (Berger and Mejia, 1973; Bueno et al., 1990; Seibert et al., 1991; Valdes et al., 1998; Light, 2010). However, the invasive nature of biopsy and time delays with culture has led to interest in other types of tests. Nucleic acid amplification tests (NAATs), based on amplification and detection of nucleic acid sequences in clinical specimens, are one such option. In-house NAATs, developed in research laboratories are less expensive compared to commercial alternatives, but are not well standardized.

Pai et al. (2004) identified 11 studies of in-house NAATs of the IS6110 target, a commonly used gene target for Mycobacterium tuberculosis that is used as a rapid test for tuberculous pleuritis (Table 1). Due to the lack of standardization across laboratories and differences in test procedures used, it is reasonable to anticipate that each primary study used different criteria to define a positive test. Thus it would be reasonable to use a HSROC-type model to meta-analyze this data. Another source of heterogeneity was the variety of reference standards. Primary studies used different composite reference standards based on different combinations of culture, microscopy, biopsy and clinical data (including signs, symptoms and clinical response to empiric TB therapy) (Pai et al., 2004).

Table 1 lists the reference standards used in each study together with the plausible range of values for the sensitivity of each reference standard. These ranges were determined based on several clinical studies and literature reviews, including systematic reviews (Berger and Mejia, 1973; Bueno et al., 1990; Seibert et al., 1991; Valdes et al., 1998; Light, 2010). Despite the highly variable sensitivity, all reference standards are believed to have high specificity ranging from 90% to 100%. We return to analyze these data in Section 5 with the hierarchical model, described in Section 3, for pooling sensitivity and specificity estimates across studies while adjusting for the imperfect and varied nature of the reference tests in the primary studies.

3. A Model for Meta-Analysis of Diagnostic Tests in the Absence of a Perfect Reference Test

We assume that J diagnostic studies are included in the meta-analysis, and that each study provides the cross-tabulation between the test under evaluation (the index test, T1) and the reference test (T2). Both tests are assumed to be dichotomous, taking the value of 1 when positive and 0 when negative. Both tests are assumed to be imperfect measures of a common underlying dichotomous latent variable D, the true disease status. Let tij and tij denote the vectors of results from study j for T1 and T2, respectively. The sensitivity of the reference test is defined by S2 = P(T2 = 1|D = 1) and its specificity is defined by C2 = P(T2 = 0|D = 0). In the simplest version of the model, we assume that the same reference standard is used in all studies.

Like Rutter and Gatsonis (2001), we assume that the observed dichotomous result on T1 is based on an underlying continuous latent variable. However, we assume that the continuous latent variable (Z1) follows a normal distribution, and that a positive result on T1 corresponds to a higher value on Z1 than a negative result. Both parameterizations give similar results when T2 is assumed to be perfect. Our model assumes that among patients with D = 0, Z1 ∼ N(−α2, exp(−β2)) and when D = 1, Z1 ∼ N(α2, exp(p2)). This model can also be conceptualized as a binomial regression model with a probit link.

Within the jth study, the difference in the means of these two distributions is αj, and the ratio of their standard
deviations is $\exp(\beta)$. Each study is assumed to use a different cut-off value, $\theta_j$, to define a positive result. We define a hierarchical prior distribution (Spiegelhalter, Abrams, and Myles, 2004) on the mean difference (or diagnostic accuracy), $\alpha_j \sim N(\Lambda, \sigma_\alpha^2)$, allowing for variation in the distribution of $Z_j$ in each study. Similarly, a hierarchical prior $\theta_j \sim N(\Theta, \sigma_\theta^2)$ allows for variation in the the cut-off values across studies. This structure is equivalent to a hierarchical model with two levels—a within-study level for study-specific parameters $\theta_j$ and $\alpha_j$, and a between-study level for parameters $\Lambda$, $\Theta$ and $\beta$ that are common to all studies.

Based on the above assumptions, the sensitivity of $T_1$ in the $j$th study is given by $S_{1j} = \Phi\left(-\frac{\theta_j - \alpha_j}{\exp(\beta)/2}\right)$, while its specificity is given by $C_{1j} = \Phi\left(\frac{\theta_j + \alpha_j}{\exp(\beta)/2}\right)$. Thus, increasing values of $\theta_j$ induce a negative correlation between sensitivity and specificity of $T_1$ across studies, while increasing values of $\alpha_j$ induce a positive correlation between them. The overall sensitivity and specificity of the index test may be summarized as $\Phi\left(-\frac{\theta - \alpha}{\exp(\beta)/2}\right)$ and $\Phi\left(\frac{\theta + \alpha}{\exp(\beta)/2}\right)$, respectively. The utility of a single pooled estimate will depend on the degree of heterogeneity between studies (Macaskill et al., 2010). A more informative approach to summarizing the data may be via an SROC plot obtained by plotting the overall sensitivity versus the overall specificity as $\Theta$ spans its range. Predicting parameter values in a future study is another way of studying the heterogeneity in a meta-analysis (Spiegelhalter et al., 2004). If the credible intervals around the predicted values are much wider than those around the pooled estimates, it would suggest that the pooled estimates cannot be generalized to individual studies. Predicted values of $\alpha$, $\theta$, and $\beta$ can be obtained from the predictive distribution of these parameters leading to predicted values of sensitivity and specificity in a future study ($j^\ast$) as follows: $\hat{S}_{1j^\ast} = \Phi\left(-\frac{\theta_j - \alpha_j}{\exp(\beta)/2}\right)$ and $\hat{C}_{1j^\ast} = \Phi\left(\frac{\theta_j + \alpha_j}{\exp(\beta)/2}\right)$, where the hat notation denotes the predicted value of a parameter.

### 3.1 Estimation

The likelihood function of the observed data across the $J$ studies can be expressed in terms of the sensitivity and specificity of each test, and the prevalence in the $j$th study, $P(D = 1|\text{Study} = j) = \pi_j$, as follows:

$$L(\Theta, \Lambda, S_2, C_2, \sigma_\alpha^2, \sigma_\theta^2, \beta, \pi_j, \alpha_j, \theta_j, j) = 1, \ldots, J|t_1, t_2, j = 1, \ldots, J$$

$$\log\left(\frac{\pi_j \Phi\left(-\frac{\theta_j - \alpha_j}{\exp(\beta)/2}\right)}{1 - \pi_j \Phi\left(\frac{\theta_j + \alpha_j}{\exp(\beta)/2}\right)} \cdot \left(1 - S_2\right) \cdot \left(1 - C_2\right) \cdot \exp\left(t_1 t_2\right)\right)$$

$$\log\left(\frac{(1 - \pi_j) \Phi\left(-\frac{\theta_j + \alpha_j}{\exp(\beta)/2}\right)}{\pi_j \Phi\left(\frac{\theta_j - \alpha_j}{\exp(\beta)/2}\right)} \cdot \left(1 - S_2\right) \cdot \left(1 - C_2\right) \cdot \exp(t_1 t_2)\right)$$

$$(1)$$

### Table 1

<table>
<thead>
<tr>
<th>Study</th>
<th>Author (Year)</th>
<th>$T_1 = 1$</th>
<th>$T_1 = 0$</th>
<th>$T_2 = 1$</th>
<th>$T_2 = 0$</th>
<th>Reference test</th>
<th>Sensitivity of reference test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Chan (1996)</td>
<td>11</td>
<td>1</td>
<td>14</td>
<td>75</td>
<td>Culture</td>
<td>20–60%</td>
</tr>
<tr>
<td>2</td>
<td>Gunisha (2001)</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>25</td>
<td>Culture</td>
<td>20–60%</td>
</tr>
<tr>
<td>3</td>
<td>Almeda (2000)</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>16</td>
<td>Culture/Clinical data</td>
<td>20–70%</td>
</tr>
<tr>
<td>4</td>
<td>Tan (1997)</td>
<td>16</td>
<td>6</td>
<td>0</td>
<td>43</td>
<td>Culture/Clinical data</td>
<td>20–70%</td>
</tr>
<tr>
<td>5</td>
<td>Portillo-Gomez(2000)</td>
<td>16</td>
<td>0</td>
<td>1</td>
<td>56</td>
<td>Culture/Biopsy</td>
<td>70–90%</td>
</tr>
<tr>
<td>6</td>
<td>De Lassence (1992)</td>
<td>9</td>
<td>0</td>
<td>6</td>
<td>10</td>
<td>Culture/Biopsy</td>
<td>70–90%</td>
</tr>
<tr>
<td>7</td>
<td>Mangiapan (1996)</td>
<td>13</td>
<td>0</td>
<td>4</td>
<td>25</td>
<td>Culture/Biopsy</td>
<td>70–90%</td>
</tr>
<tr>
<td>8</td>
<td>Querol (1995)</td>
<td>17</td>
<td>2</td>
<td>4</td>
<td>84</td>
<td>Culture/Biopsy</td>
<td>70–90%</td>
</tr>
<tr>
<td>9</td>
<td>Tan, Jama (1995)</td>
<td>7</td>
<td>0</td>
<td>3</td>
<td>13</td>
<td>Culture/Biopsy</td>
<td>70–90%</td>
</tr>
<tr>
<td>10</td>
<td>Villena (1998)</td>
<td>14</td>
<td>1</td>
<td>19</td>
<td>97</td>
<td>Culture/Biopsy</td>
<td>70–90%</td>
</tr>
<tr>
<td>11</td>
<td>Villegas (2000)</td>
<td>31</td>
<td>7</td>
<td>11</td>
<td>63</td>
<td>Culture/Biopsy</td>
<td>70–90%</td>
</tr>
</tbody>
</table>
To carry out Bayesian estimation, we need to specify prior distributions over the set of unknown parameters. Our overall strategy was to use noninformative (objective) prior distributions for most parameters, but using prior parameter values that cover a reasonable range. The priors for $\lambda$, $\Theta$, and $\beta$ were selected so that the resulting marginal distributions on the pooled sensitivity or specificity were approximately uniform over (0,1). The pooled "difference in means" parameter was assumed to have prior density $\Lambda \sim U(-3,3)$. The log of the ratio between the two standard deviations, $\beta$ was assumed to follow a $U(-0.75, 0.75)$ distribution. The pooled "cut-off" parameter, $\Theta$ was assumed to follow a $U(-1.5, 1.5)$ distribution. Parameters $\sigma_1$ and $\sigma_0$ were assumed to follow $U(0,2)$ distributions. For the $\pi$, $S_2$, and $C_2$ parameters we used Beta prior distributions. As we will illustrate in our example, some of these prior distributions may be informative. When an objective prior distribution was desired we used the Beta(1,1) distribution.

The total number of degrees of freedom available is $3J$, with each study contributing 3 degrees of freedom. The total number of parameters to be estimated is at least $J+7$. Therefore, a minimum of 4 studies would be required to reasonably estimate this model without any informative prior distributions. Since the two parameters $S_{1j}$ and $C_{1j}$ are defined as a function of three parameters $-a_j$, $\theta_j$, $\beta$ we chose to assume $\beta$ was the same across all studies to avoid problems of non-identifiability. A similar assumption was made in Rutter and Gatsonis (2001). If there is a need to allow $\beta$ to vary across studies, informative prior distributions would be needed over these additional parameters.

There being no analytical solution to the marginal posterior distributions, we used a Gibbs sampler algorithm to obtain a sample from the marginal posterior distributions of the parameters of interest. Most full-conditional distributions were of known forms, except that of $\beta$ for which we used a Metropolis–Hastings step (see Web Appendix A for a listing of the full-conditional distributions). We have developed an R package, HSRROC, to implement this algorithm (Schiller and Dendukuri, 2011). To assess convergence of the models in Section 5 we ran five different chains of 50,000 iterations starting at disparate initial values and calculated the Gelman–Rubin statistic for comparing variability within and between chains (Gelman and Rubin, 1992). Convergence was achieved fairly rapidly for all the models we considered. We dropped the first 10,000 iterations in each of the five chains and reported summary statistics based on the remaining 200,000 iterations. We have also written programs in WinBUGS (Spiegelhalter, Thomas, and Best, 2007) and using PROC MCMC in SAS (SAS Institute Inc., 2009) to implement the model and verified that all programs provide identical results up to Monte Carlo error.

3.2 Multiple Reference Standards

It is possible that different studies in a meta-analysis use different reference standards. This can be accommodated by replacing parameters $S_j$ and $C_j$ by $S_{2j}$ and $C_{2j}$, respectively, in (1), and defining independent prior distributions for the parameters of each reference standard, e.g., $S_{2j} \sim Beta(a_{2j}, b_{2j})$ and $C_{2j} \sim Beta(c_{2j}, d_{2j})$. As in the case of a single reference standard, these prior distributions may be informative or objective. When the same reference standard is used in two different studies $j$ and $jt$, we could assume the accuracy is the same in both studies ($S_{2j} = S_{2jt}$ and $C_{2j} = C_{2jt}$) or different and allow for hierarchical prior distributions on $logit(S_{2j})$ and $logit(C_{2j})$ as in Bernatsky et al. (2005).

3.3 Conditional Dependence Between Index and Reference Tests

In the absence of a gold standard reference it is important to adjust for conditional dependence between multiple tests carried out on the same subjects (Dendukuri and Joseph, 2001) in order to adjust for unexplained correlation between the tests within each latent class. The model in (1) can be extended to adjust for conditional dependence in a number of ways (Dendukuri, Wang, and Hadgu, 2009). For the application in Section 5 of this article we consider modeling conditional dependence by the addition of covariance terms between the sensitivity of index and reference tests ($cov_{S_j}$) in the jth study and between their specificity ($cov_{C_j}$) in the jth study as in Dendukuri and Joseph (2001) and Chu et al. (2009). The joint probability of the two tests in the jth study is

$$P(T_1 = u, T_2 = v | \text{Study} = j) = \pi_j \left\{ S_{1j}^u (1 - S_{1j})^{1-u} S_{2j} (1 - S_{2j})^{1-v} + (-1)^{u-v} cov_{S_j} ight\}$$
$$+ (1 - \pi_j) \left\{ C_{1j}^{u-v} (1 - C_{1j})^{-u} C_{2j}^{v-u} (1 - C_{2j})^v + (-1)^{u-v} cov_{C_j} \right\}$$

We defined uniform prior distributions over the covariance parameters as follows: $cov_{S_j} \sim U(0, \min(S_{1j}, S_{2j}) - S_{1j}, S_{2j})$, $cov_{C_j} \sim U(0, \min(C_{1j}, C_{2j}) - C_{1j}, C_{2j})$. Models adjusting for conditional dependence were fit using WinBUGS.

4. Simulation Study

We fit the model in (1) to simulated datasets generated under eight scenarios in order to examine the impact of: (i) the number of studies ($J = 5, 10, 20$, or $35$), (ii) the range of sample sizes of the primary studies (n = 50–200 or n = 200–500), and (iii) sensitivity to the prior distributions for $S_j$ and $C_j$, on the performance of our model. Sensitivity to the prior distributions was examined by fitting each data set with three different sets of prior distributions for $S_j$ and $C_j$ - informative, noninformative, and degenerate at $S_j = C_j = 1$. We estimated the frequentist properties of the model in terms of bias (average absolute difference between true value and posterior median), as well as average coverage and average length of the 95% credible intervals of the key parameters in the model across 500 datasets generated under each of the eight scenarios.

Results for scenarios where sample sizes of individual studies ranged from 50 to 200 are summarized in Table 2 and in Web Table 1. Results for scenarios with larger sample sizes are in Web Table 2. In all eight scenarios the true values of the pooled index test sensitivity and specificity was 0.9, while the reference test was assumed to have low sensitivity of 0.6 and higher specificity of 0.95 like many tests for TB. The informative prior distribution over $S_j$ was Beta(57,38) (95% credible interval from 0.5 to 0.7) and the informative prior distribution over $C_j$ was Beta(95,5) (95% credible interval from 0.8997 to 0.9834). The prevalence in individual studies ranged from 0.15 to 0.4. We used a Beta(2.75,8.25) prior
distribution over the prevalence to increase the chances that
the Gibbs sampler converged to the more meaningful mode
(with prevalence <0.5, and sensitivity and specificity >0.5)
in all 100 datasets.

The following general observations can be made from these
results:

1 When allowing the reference standard to be imperfect,
bias in overall sensitivity and specificity was less than
0.05 across all scenarios, while the average coverage was
very high, exceeding 95% for a number of scenarios. The
average length of the 95% credible interval decreased with
increasing number of studies in the meta-analysis, though the sample sizes of individual studies did not
appear to have a substantial impact in the scenarios
considered.

2 When incorrectly assuming the reference standard was
perfect, there was a bias in estimation of overall sen-
sitivity and specificity of about 0.15. The average cov-
verage for these parameters was very poor (less than
70% in all scenarios), decreasing to 0 as the number of
studies increased. There was also considerable bias in
the estimation of the heterogeneity parameters ($\sigma_\alpha$ and
$\sigma_\beta$), with the bias and coverage worsening with in-
creasing number of studies.

3 There was no difference in the performance of the model
when using informative or noninformative prior distribu-
tions over $S_i$ and $C_j$. This suggests that prior dis-
tributions we considered were dominated by the data
even when the number of studies was as low as $J = 5$.

4 When allowing the reference standard to be imperfect,
the bias in estimating parameters $\sigma_\alpha$ and $\sigma_\beta$ was around
0.15 when $J = 5$ studies were included in the meta-
analysis. It decreased with increasing number of studies
in the meta-analysis. The average length of the 95%
credible interval for these parameters decreased with
increasing number of studies in the meta-analysis as well
with higher sample size per study.

5 The parameter $\beta$ was least well estimated. Bias in esti-
mation did not decrease with increasing number of
studies nor higher sample size per study. Wide aver-
age length ensured high average coverage above 95%
for most scenarios considered.

The above results suggest that the model we are propos-
ing performs well on average, including when using nonin-
formative prior distributions. We also considered a simulated
scenario that resembled our tuberculous pleuritis data, i.e.,
$J = 11$ studies using three different reference standards across
studies. We considered both conditionally independent and
dependent models. In both cases we observed similar results
to those described above (data not shown), suggesting the
model can be applied to the data at hand.

5. Meta-Analysis of Nucleic Acid Amplification Tests

As explained in Section 2, we determined plausible ranges
for the sensitivity and specificity of the three reference stan-
dards based on a review of the literature. We transformed
the prior information on the plausible ranges of the sensi-
tivity and specificity of the reference tests for TB pleuritis
into Beta($\alpha$, $\beta$) prior distributions. This was done by equat-
ing the mid-point of the range to the mean of the Beta dis-
tribution $\frac{\alpha + \beta}{\alpha + \beta + 1}$. Ranges for the sensitivity
of each reference test are given in Table 1. The prior range
for specificity across all references tests is 0.9–1. The prior
distributions over the sensitivities of the three different refer-
ce tests were: (i) culture: $S_{21} \sim$ Beta(9.2, 13.8), (ii) culture and
clinical data: $S_{22} \sim$ Beta(6.678, 8.162), (iii) culture and
Biopsy: $S_{23} \sim$ Beta(50.4, 12.6). The specificities of all refer-
ce tests were assumed to have the same prior distribution,
$C_{2j} \sim$ Beta(71.25, 3.75), $j = 1–2, 3–4, 5–11$. The prevalence
in each study was assumed to follow a Beta(1,1) distribution.
We did not consider hierarchical priors as two of the reference
standards were used in only two studies each.

We considered the model in (1) that assumes conditional
independence between the index and reference tests, together
with a model that adjusted for conditional dependence be-
tween the tests in each study. We considered models with
informative prior distributions over $S_i$ and $C_j$, as well as non-
informative Beta(1,1) distributions. For comparison, we also
considered the model that assumed the reference test was per-
fect in all studies. Model fit was compared using the Deviance
Information Criterion (DIC) (Spiegelhalter et al., 2002).

We found that the DIC was lowest (indicating the best fit)
for the model with noninformative prior distributions that ad-
justed for conditional dependence (DIC = 145.7). The worst
fitting model was the one that assumed conditional indepen-
dence between index and reference test (DIC = 171.1 with
informative prior, DIC = 155.1 with noninformative prior).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Statistic</th>
<th>I</th>
<th>NI</th>
<th>GS</th>
<th>I</th>
<th>NI</th>
<th>GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_i$</td>
<td>AB</td>
<td>0</td>
<td>0</td>
<td>0.15</td>
<td>0.03</td>
<td>0.03</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>AL</td>
<td>0.24</td>
<td>0.23</td>
<td>0.12</td>
<td>0.13</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AC</td>
<td>1</td>
<td>1</td>
<td>0.09</td>
<td>0.98</td>
<td>0.95</td>
<td>0</td>
</tr>
<tr>
<td>$C_j$</td>
<td>AB</td>
<td>0.03</td>
<td>0.04</td>
<td>0.13</td>
<td>0.02</td>
<td>0.03</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>AL</td>
<td>0.23</td>
<td>0.25</td>
<td>0.2</td>
<td>0.12</td>
<td>0.13</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>AC</td>
<td>0.96</td>
<td>0.95</td>
<td>0.09</td>
<td>0.95</td>
<td>0.93</td>
<td>0</td>
</tr>
<tr>
<td>$\sigma_\alpha$</td>
<td>AB</td>
<td>0.07</td>
<td>0.04</td>
<td>0.42</td>
<td>0</td>
<td>0.01</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>AL</td>
<td>1.63</td>
<td>1.64</td>
<td>0.71</td>
<td>1.12</td>
<td>1.13</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>AC</td>
<td>0.99</td>
<td>0.98</td>
<td>0.56</td>
<td>0.96</td>
<td>0.95</td>
<td>0</td>
</tr>
<tr>
<td>$\sigma_\beta$</td>
<td>AB</td>
<td>0.02</td>
<td>0.00</td>
<td>0.12</td>
<td>0.02</td>
<td>0.01</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>AL</td>
<td>0.98</td>
<td>0.96</td>
<td>0.55</td>
<td>0.47</td>
<td>0.47</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>AC</td>
<td>0.97</td>
<td>0.98</td>
<td>0.90</td>
<td>0.97</td>
<td>0.96</td>
<td>0.44</td>
</tr>
<tr>
<td>$\beta$</td>
<td>AB</td>
<td>0.23</td>
<td>0.28</td>
<td>0.14</td>
<td>0.2</td>
<td>0.24</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>AL</td>
<td>1.35</td>
<td>1.36</td>
<td>1.18</td>
<td>1.19</td>
<td>1.21</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>AC</td>
<td>1</td>
<td>1</td>
<td>0.98</td>
<td>0.96</td>
<td>0.96</td>
<td>0.94</td>
</tr>
</tbody>
</table>
The estimates of the sensitivity and specificity of the index test within each study as well as overall from the best-fitting model are given in Table 3.

For comparison, this table also includes estimates from a model that assumed all studies used a perfect reference test and a model using the informative prior distributions. In general, the best-fitting model gave estimates of sensitivity, specificity and prevalence that were intermediate between the other two models in Table 3. The wide credible intervals around these estimates imply similar inferences, showing poor sensitivity for the NAAT with heterogeneity across studies, and consistently high specificity. This is also reflected in the SROC curves in Figure 1, which plot Sensitivity (\( \Phi \)) versus Specificity (\( 1 - \Phi \)) at the posterior mean values of \( \Lambda \) and \( \beta \) as the value of \( T \) varies over the 95% credible interval of \( \Theta \).

Our hierarchical model makes a number of unverifiable assumptions that could have an important effect on the estimation of the parameters of interest (Spiegelhalter et al., 2004). Therefore we carried out a series of sensitivity analyses to check its robustness. We changed the distribution of the latent variable \( Z_1 \) to logistic, allowed the random effects to follow a \( t(4) \) distribution rather than a normal distribution, and considered two alternative prior distributions over the between-study variability. Under the first alternative both \( \sigma^2_\alpha \) and \( \sigma^2_\beta \) followed a \( U(0, 5) \) distribution, while the second alternative assumed that \( \frac{1}{\sigma^2_\alpha} \) and \( \frac{1}{\sigma^2_\beta} \) both followed a Gamma(shape = 0.5, rate = 0.5) distribution. We found that the individual and pooled sensitivity and specificity estimates and credible intervals from the different models were fairly similar, and the DIC did not differ greatly between them (Web Table 3). Therefore we concluded that our selected model was robust for this particular application.

Finally, for comparison, we extended the conditional dependence model of Chu et al. (2009) to allow for multiple reference standards. We fit the model to our data assuming that the sensitivity and specificity of each reference standard were independent of each other and the other parameters in the model. We found that the results were very similar to those obtained from our best fitting model (Pooled sensitivity 0.69 (0.39, 0.91); Pooled specificity 0.97 (0.86, 0.99); DIC 149.4).

Based on our best fitting model, there was high variability in both the diagnostic accuracy and threshold parameters across studies (see \( \sigma_\alpha \) and \( \sigma_\beta \) in first line of Web Table 3), though the variability in accuracy was greater. The high variation in the threshold parameter across studies supports our concern that in-house NAAT tests may not be well standardized. The median and 95% credible interval for the predicted sensitivity and specificity in a future study were 0.65 (0.19, 0.96) and 0.96 (0.73, 1.00), respectively, indicating considerable heterogeneity between the observed studies. This heterogeneity is also depicted in the SROC curve in Figure 1. This suggests that the pooled sensitivity and specificity are not widely generalizable and more research is needed to study the reasons for heterogeneity in the accuracy between studies.

The posterior median and 95% equal-tailed credible intervals for the sensitivity and specificity of the four reference tests was as follows: (i) Culture: Sensitivity 0.65 (0.25, 0.96), Specificity 0.87 (0.76, 0.98) (ii) Culture and clinical data: Sensitivity 0.65 (0.33, 0.95), Specificity 0.97 (0.80, 0.99), (iii) Culture and biopsy: Sensitivity 0.85 (0.59, 0.99), Specificity 0.95 (0.84, 1.00). Thus, based on the selected model, in-house NAATs may not be well standardized, the median and 95% credible interval for the predicted sensitivity and specificity in a future study were 0.65 (0.19, 0.96) and 0.96 (0.73, 1.00), respectively, indicating considerable heterogeneity between the observed studies. This heterogeneity is also depicted in the SROC curve in Figure 1. This suggests that the pooled sensitivity and specificity are not widely generalizable and more research is needed to study the reasons for heterogeneity in the accuracy between studies.

6. Discussion

We have presented a Bayesian hierarchical model adjusting for the imperfect nature of the reference standard in a
Bayesian Meta-Analysis in the Absence of a Gold Standard Reference

Figure 1. Summary receiver operating characteristic curves for competing meta-analysis models of sensitivity and specificity of an in-house nucleic acid amplification test for tuberculous pleuritis (IS6110).

Bivariate meta-analysis of diagnostic test sensitivity and specificity. The model allows for different reference standards to be used in individual studies, a feature commonly encountered when a gold standard reference is either nonexistent or prohibitive. Our results show that ignoring the imperfect nature of the reference may result in biased estimates of pooled sensitivity and specificity of the test under evaluation. In the case of TB pleuritis, our results show that in-house NAAT tests for the IS6110 target may have worse sensitivity but better specificity than previously estimated. An earlier meta-analysis of in-house NAATs for varied targets for TB pleuritis had estimated their pooled sensitivity and specificity as 0.71 (0.63, 0.78) and 0.93 (0.88, 0.96), respectively (Pai et al., 2004).

Our simulations show that model performance is enhanced by both the number of studies as well as the sample size per study, particularly for estimation of between-study heterogeneity. For the scenarios we considered, we did not encounter major problems with the convergence of the Gibbs sampler. However, we noticed that when using noninformative prior distributions for all parameters, the problem of permutation nonidentifiability can result in the estimates of individual study parameters converging to the mode that is not meaningful (i.e., prevalence >0.5, sensitivity and specificity <0.5 for the scenarios we considered). Consequently, the pooled estimates are not meaningful. The problem of permutation identifiability is well recognized in the literature on latent class analysis (McLachlan and Peel, 2000). The likelihood function of a model with G latent classes has G! modes. In the meta-analysis model at hand, the problem is exacerbated due to the combinations of modes across studies resulting in G!! modes. For example, for the TB dataset we considered, the likelihood function has 2^{11} modes. To distinguish between these modes we can use our substantive knowledge of the test accuracy parameters or the prevalence. We know that the specificity of all three reference tests for TB pleuritis is very high, exceeding 90%, while their sensitivity is poor. Thus before reporting the pooled sensitivity and specificity we need to ensure that the Gibbs sampler has converged to the mode of (π_j, S_{1j}, S_{2j}, C_{1j}, C_{2j}) and not (1 − π_j, 1 − C_{1j}, 1 − C_{2j}, 1 − S_{1j}, 1 − S_{2j}) in each individual study. Carefully selected initial values and weakly informative prior distributions helps to avoid this problem.

The advantage of the HSROC model (Rutter and Gatsonis, 2001), that we chose to extend, is that it models the variation in diagnostic accuracy and cut-off values, both well-recognized sources of heterogeneity across diagnostic studies (Macaskill et al., 2010). The model we have described can be considered...
a special case of the model described by Chu et al. (2009) in the situation: (i) the same reference standard is used in all studies, (ii) only the sensitivity and specificity of the index test are considered correlated, and (iii) there are no covariates affecting sensitivity and specificity of either index or reference tests. However, as noted by Harbord et al. (2007), despite the likelihood functions of the two models being equivalent, the prior distributions do not share a one-to-one relationship and hence Bayesian inference of the two models may not yield identical inferences. Further research is needed to establish the links between these two models. For the particular case of the tuberculous pleuritis data we found that both models gave similar results. The best approach for fitting a summary receiver operating characteristic function remains a topic of debate (Arends et al., 2008).

The model we described can be extended in numerous ways to accommodate some well-known practical problems. To extend the model to the situation when the index test is ordinal, additional cut-off parameters will have to be added to the model (Tosteson and Begg, 1988). As described by Rutter and Gatsonis (2001), we can express the \( \theta_i \) and \( \alpha_i \) parameters as functions of covariates. Whether adjustment for conditional dependence in a meta-analytic setting is appropriate when heterogeneity between studies is caused by known covariates is another open question.

7. Supplementary Materials
Web Appendix A (including full-conditional distributions referenced in Section 3.1, Web Tables 1 and 2 referenced in Section 4 and Web Table 3 referenced in Section 5), are available with this article at the Biometrics website on Wiley Online Library.

Acknowledgements
This work was supported by the Canadian Institutes of Health Research (CIHR) (Grant 89857). Nandini Dendukuri holds a Chercheur Boursier award from the Fonds de la Recherche en Santé du Québec. Madhukar Pai holds a New Investigator award from CIHR.

References


Received January 2011. Revised March 2012. Accepted March 2012.