Biometrics. Author manuscript; available in PMC 2013 July 30.

Published in final edited form as:

Biometrics. 2012 December ; 68(4): 1285–1293. doi:10.1111/j.1541-0420.2012.01773.x.

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 Pai**[1]

[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal H3A 1A2, Canada

[2]Division of Clinical Epidemiology, McGill University Health Centre, Montreal H3A 1A1, Canada

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.

Keywords

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

* nandini.dendukuri@mcgill.ca.

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.

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 (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, T_1) and the reference test (T_2). 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 t_{1j} and t_{2j} denote the vectors of results from study j for T_1 and T_2, respectively. The sensitivity of the reference test is defined by $S_2 = P(T_2 = 1/D = 1)$ and its specificity is defined by $C_2 = P(T_2 = 0/D = 0)$. 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 T_1 is based on an underlying continuous latent variable. However, we assume that the continuous latent variable (Z_1) follows a normal distribution, and that a positive result on T_1 corresponds to a higher value on Z_1 than a negative result. Both parameterizations give similar results when T_2 is assumed to be perfect. Our model assumes that among patients with $D = 0$, $Z_1 \sim N\{-\frac{\alpha_j}{2}, \exp(-\frac{\beta}{2})\}$ and when $Z_1 \sim N\{\frac{\alpha_j}{2}, \exp(\frac{\beta}{2})\}$. 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 α_j, and the ratio of their standard deviations is $exp(\beta)$. Each study is assumed to use a different cut-off value, θ_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_1 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 θ_j and α_j and a between-study level for parameters Λ, Θ and β that are common to all studies.

Based on the above assumptions, the sensitivity of T_1 in the jth study is given by $S_{1j}=\Phi\{-\frac{\theta_j-\frac{\alpha_j}{2}}{\exp(\frac{\beta}{2})}\}$, while its specificity is given by $C_{1j}=\Phi\{\frac{\theta_j+\frac{\alpha_j}{2}}{\exp(-\frac{\beta}{2})}\}$. Thus, increasing values of θ_j induce a negative correlation between sensitivity and specificity of T_1 across studies, while increasing values of α_j induce a positive correlation between them. The overall sensitivity and specificity of the index test may be summarized as $\Phi\{-\frac{\Theta-\frac{\Lambda}{2}}{\exp(\frac{\beta}{2})}\}$ and $\Phi\{\frac{\Theta+\frac{\Lambda}{2}}{\exp(-\frac{\beta}{2})}\}$, 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 Θ 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 α, θ, and β can be obtained from the predictive distribution of these parameters leading to predicted values of sensitivity and specificity in a future study (j') as follows: $\hat{S}_{1j'}=\Phi\{-\frac{\hat{\theta}_{j'}-\frac{\hat{\alpha}_{j'}}{2}}{\exp(\frac{\hat{\beta}}{2})}\}$ and $\hat{C}_{1j'}=\Phi\{\frac{\hat{\theta}_{j'}+\frac{\hat{\alpha}_{j'}}{2}}{\exp(-\frac{\hat{\beta}}{2})}\}$, 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 jth study, ($P(D=1/\text{Study}=j)=\pi_j$), as follows:

$$\begin{aligned}
&L(\Theta,\Lambda,S_2,C_2,\sigma_\alpha^2,\sigma_\theta^2,\beta,\pi_j,\alpha_j,\theta_j,j\\
&=1,\ldots,J|t_{1j},t_{2j},j=1,\ldots,J)\\
&=\prod_{j=1}^{J}\left[\pi_j\Phi\left\{-\frac{\theta_j-\frac{\alpha_j}{2}}{\exp\left(\frac{\beta}{2}\right)}\right\}S_2+(1-\pi_j)\Phi\left\{-\frac{\theta_j+\frac{\alpha_j}{2}}{\exp\left(\frac{-\beta}{2}\right)}\right\}(1-C_2)\right]^{t_{1j}\cdot t_{2j}}\\
&\times\left[\pi_j\Phi\left\{-\frac{\theta_j-\frac{\alpha_j}{2}}{\exp\left(\frac{\beta}{2}\right)}\right\}(1-S_2)+(1-\pi_j)\Phi\left\{-\frac{\theta_j+\frac{\alpha_j}{2}}{\exp\left(\frac{-\beta}{2}\right)}\right\}C_2\right]^{t_{1j}\cdot(1-t_{2j})}\\
&\times\left[\pi_j\Phi\left\{\frac{\theta_j-\frac{\alpha_j}{2}}{\exp\left(\frac{\beta}{2}\right)}\right\}S_2+(1-\pi_j)\Phi\left\{\frac{\theta_j+\frac{\alpha_j}{2}}{\exp\left(\frac{-\beta}{2}\right)}\right\}(1-C_2)\right]^{(1-t_{1j})\cdot t_{2j}}\\
&\times\left[\pi_j\Phi\left\{\frac{\theta_j-\frac{\alpha_j}{2}}{\exp\left(\frac{\beta}{2}\right)}\right\}(1-S_2)+(1-\pi_j)\Phi\left\{\frac{\theta_j+\frac{\alpha_j}{2}}{\exp\left(\frac{-\beta}{2}\right)}\right\}C_2\right]^{(1-t_{1j})\cdot(1-t_{2j})}.
\end{aligned}\tag{1}$$

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 Λ, Θ, and β 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, β was assumed to follow a U(−0.75, 0.75) distribution. The pooled "cut-off" parameter, Θ was assumed to follow a U(−1.5, 1.5)

distribution. Parameters σ_α and σ_θ were assumed to follow U(0,2) distributions. For the π_j, 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—α_j, θ_j, β—we chose to assume β 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 β 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 β 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, HSROC, 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_2 and C_2 by S_{2j} and C_{2j}, respectively, in (1), and defining independent prior distributions for the parameters of each reference standard, e.g., $S_{2j} \sim \text{Beta}(sa_j, sb_j)$ and $C_{2j} \sim \text{Beta}(ca_j, cb_j)$. 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 j', we could assume the accuracy is the same in both studies ($S_{2j} = S_{2j'}$ and $C_{2j} = C_{2j'}$) 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 ($\text{cov}s_j$) in the jth study and between their specificity ($\text{cov}c_j$) 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

$$P(T_1=u, T_2=v|\text{Study}=j) = \pi_j\{S_{1j}^u(1-S_{1j})^{(1-u)}S_2^v(1-S_2)^{(1-v)}+(-1)^{|u-v|}\text{cov}s_j\}+(1-\pi_j)\{C_{1j}^{(1-u)}(1-C_{1j})^u C_2^{(1-v)}(1-C_2)^v+(-1)^{|u-v|}\text{cov}c_j\}$$

We defined uniform prior distributions over the covariance parameters as follows: $\text{cov}s_j \sim U(0, \min(S_{1j}, S_2) - S_{1j}S_2)$, $\text{cov}c_j \sim U(0, \min(C_{1j}, C_2) - C_{1j}C_2)$. 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_2 and C_2, 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_2 and C_2 - informative, noninformative, and degenerate at $S_2 = C_2 = 1$. 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_2 was Beta(57,38) (95% credible interval from 0.5 to 0.7) and the informative prior distribution over C_2 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 sensitivity and specificity of about 0.15. The average coverage 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 (σ_a and σ_θ), with the bias and coverage worsening with increasing number of studies.

3. There was no difference in the performance of the model when using informative or noninformative prior distributions over S_2 and C_2. This suggests that prior distributions 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 σ_a and σ_θ 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 β was least well estimated. Bias in estimation did not decrease with increasing number of studies nor higher sample size per study. Wide average length ensured high average coverage above 95% for most scenarios considered.

The above results suggest that the model we are proposing performs well on average, including when using noninformative 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 standards based on a review of the literature. We transformed the prior information on the plausible ranges of the sensitivity and specificity of the reference tests for TB pleuritis into Beta(α, β) prior distributions. This was done by equating the mid-point of the range to the mean of the Beta distribution $\frac{\alpha}{\alpha+\beta}$, and matching one quarter of the range to its standard deviation $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. 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 reference 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 reference 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 between the tests in each study. We considered models with informative prior distributions over S_2 and C_2, as well as noninformative Beta(1,1) distributions. For comparison, we also considered the model that assumed the reference test was perfect 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 adjusted for conditional dependence (DIC = 145.7). The worst fitting model was the one that assumed conditional independence between index and reference test (DIC = 171.1 with informative prior, DIC = 155.1 with noninformative prior). 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 ($\Phi\{-\frac{T-\frac{\Lambda}{2}}{\exp(\frac{\beta}{2})}\}$) versus Specificity ($\Phi\{\frac{T+\frac{\Lambda}{2}}{\exp(-\frac{\beta}{2})}\}$) at the posterior mean values of Λ and β as the value of T varies over the 95% credible interval of Θ.

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 σ_α^2 and σ_θ^2 followed a U(0,5) distribution, while the second alternative assumed that $\frac{1}{\sigma_\alpha^2}$ and $\frac{1}{\sigma_\theta^2}$ 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 σ_α and σ_θ 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 interval 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 for the IS6110 target do not improve over the best-known reference of culture and biopsy in terms of sensitivity. The higher DIC of the models with informative prior distributions can be attributed to a disagreement between the observed data and the prior information. In particular, the posterior credible intervals for the sensitivity of culture and culture+clinical data cover wider ranges than we had provided, including higher values.

6. Discussion

We have presented a Bayesian hierarchical model adjusting for the imperfect nature of the reference standard in a 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 $G!^P$ 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_2, C_{1j}, C_2) and not ($1 - \pi_j$, $1 - C_{1j}$, $1 - C_2$, $1 - S_{1j}$, $1 - S_2$) 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 when: (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 θ_j and α_j 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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

Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. Medical Decision Making. 2008; 28:621–638. [PubMed: 18591542]

Berger HW, Mejia E. Tuberculous pleurisy. Chest. 1973; 63:88–92. [PubMed: 4630686]

Bernatsky S, Joseph L, Bélisle P, Boivin JF, Rajan R, Moore A, Clarke A. Bayesian modeling of imperfect ascertainment methods in cancer. Statistics in Medicine. 2005; 24:2365–2379. [PubMed: 15977290]

Bueno EC, Clemente GM, Castro CB, Martin ML, Ramos RS, Glez-Rio GAPMJ. Cytologic and bacteriologic analysis of fluid and pleural biopsy specimens with Cope's needle. Study of 414 patients. Archives of Internal Medicine. 1990; 150:1190–1194. [PubMed: 2353852]

Chu H, Chen S, Louis TA. Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard. Journal of the American Statistical Association. 2009; 104:512–523. [PubMed: 19562044]

Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics. 2001; 57:158–167. [PubMed: 11252592]

Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: A multiple latent variable model. Statistics in Medicine. 2009; 28:441–461. [PubMed: 19067379]

Gelman A, Rubin DB. Inferences from iterative simulation using multiple sequences (with discussion). Statistical Science. 1992; 7:457–511.

Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007; 8:239–251. [PubMed: 16698768]

Light RW. Update on tuberculous pleural effusion. Respirology. 2010; 15:451–458. [PubMed: 20345583]

Macaskill, P., Gatsonis, C., Deeks, J., Harbord, R., Takwoingi, Y. Deeks, JJ.Bossuyt, PM., Gatsonis, C., editors. The Cochrane Collaboration. Chapter 10: Analysing and presenting results. Handbook for Diagnsotic Test Accuracy Reviews Version 1.0. 2010. http://srdta.cochrane.org/

McLachlan, G., Peel, D. Finite Mixture Modeling. Hoboken, NJ: John Wiley and Sons Limited; 2000.

Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: Data analytic approaches and some additional considerations. Statistics in Medicine. 1993; 12:1293–1316. [PubMed: 8210827]

Pai M, Flores LL, Hubbard A, Riley LW, Colford JM. Nucleic acid amplification tests in the diagnosis of tuberculos pleuritis: A systematic review and meta-analysis. Biomed Central (BMC) Infectious Diseases. 2004; 4:1–14.

Pai M, Ramsay A, O'Brien R. Evidence-based tuberculosis diagnosis. PLoS Medicine. 2008; 5:e156. [PubMed: 18651788]

Pai M, Minion J, Steingart K, Ramsay A. New and improved tuberculosis diagnostics: Evidence, policy, practice and impact. Current Opinion in Pulmonary Medicine. 2010; 16:271–284. [PubMed: 20224410]

Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology. 2005; 58:982–990. [PubMed: 16168343]

Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic accuracy evaluations. Statistics in Medicine. 2001; 20:2865–2884. [PubMed: 11568945]

Sadatsafavi M, Shahidi N, Marra F, FitzGerald MJ, Elwood KR, Guo N, Marra CA. A statistical method was used for the meta-analysis of tests for latent TB in the absence of a gold standard, combining random effect and latent-class methods to estimate test accuracy. Journal of Clinical Epidemiology. 2010; 63:257–269. [PubMed: 19692208]

SAS Institute Inc. SAS/STAT® *User's Guide*. 2. SAS Institute Inc; Cary NC: 2009.

Schiller, I., Dendukuri, N. [Accessed March 2011.] HSROC: Joint meta-analysis of diagnostic test sensitivity and specificity with or without a gold standard reference test version 1.0.0. 2011. Available at http://cran.r-project.org/web/packages/HSROC/index.html

Seibert AF, Haynes J, Middleton R, Bass JB. Tuberculous pleural effusion: Twenty-year experience. Chest. 1991; 99:883–886. [PubMed: 1901261]

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B. 2002; 64:583–639.

Spiegelhalter, DJ., Thomas, A., Best, NG. WinBUGS version 1.4.3 User Manual. Medical Research Council Biostatistics Unit; United Kingdom: 2007.

Spiegelhalter, DJ., Abrams, KR., Myles, JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. New York: John Wiley and Sons Limited; 2004.

Tosteson AN, Begg CB. A general regression methodology for ROC curve estimation. Medical Decision Making. 1988; 8:204–215. [PubMed: 3294553]

Valdes L, Alvarez D, Jose SE, Penela P, Valle JM, Garcia-Pazos JM, Suarez J, Pose A. Tuberculous pleurisy: A study of 254 patients. Archives of Internal Medicine. 1998; 158:2017–2021. [PubMed: 9778201]

Walter SD, Irwig LM. Estimation of error rates, disease prevalence, and relative risk misclassified data: A review. Journal of Clinical Epidemiology. 1988; 41:923–937. [PubMed: 3054000]

Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. Journal of Clinical Epidemiology. 1999; 52:943–951. [PubMed: 10513757]

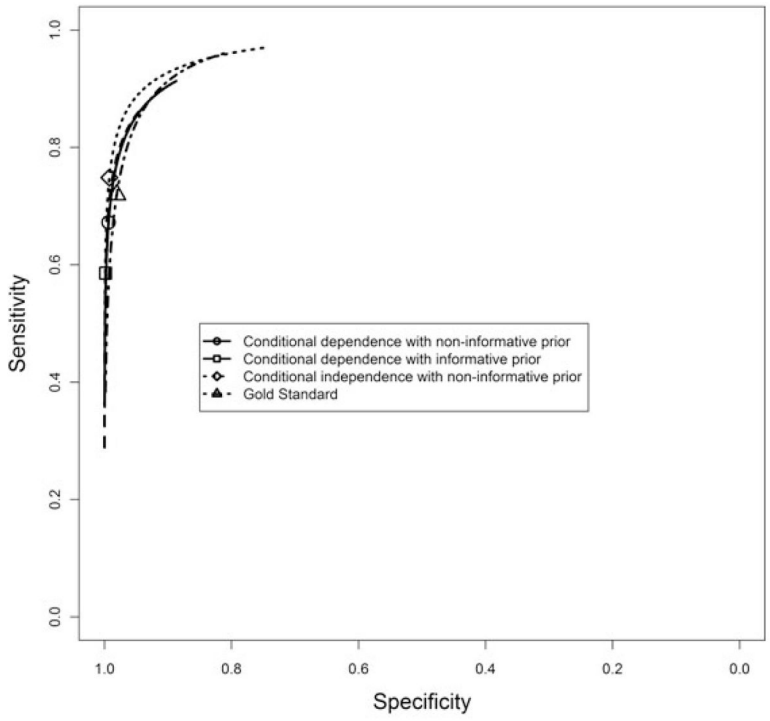

Figure 1.

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).

Table 1

Studies included in meta-analysis of in-house nucleic acid amplification tests for tuberculous pleuritis (Source: Pai et al. 2004).

		Index (T_1) and reference (T_2) test results					
Study	Author (Year)	$T_1 = 1$ $T_2 = 1$	$T_1 = 1$ $T_2 = 0$	$T_1 = 0$ $T_2 = 1$	$T_1 = 0$ $T_2 = 0$	Reference test	Sensitivity of reference test
1	Chan (1996)	11	1	14	75	Culture	20–60%
2	Gunisha (2001)	1	1	3	25	Culture	20–60%
3	Almeda (2000)	8	0	1	16	Culture/Clinical data	20–70%
4	Tan (1997)	16	6	0	43	Culture/Clinical data	20–70%
5	Portillo-Gomez (2000)	16	0	1	56	Culture/Biopsy	70–90%
6	De Lassence (1992)	9	0	6	10	Culture/Biopsy	70–90%
7	Mangiapan (1996)	13	0	4	25	Culture/Biopsy	70–90%
8	Querol (1995)	17	2	4	84	Culture/Biopsy	70–90%
9	Tan, Jama (1995)	7	0	3	13	Culture/Biopsy	70–90%
10	Villena (1998)	14	1	19	97	Culture/Biopsy	70–90%
11	Villegas (2000)	31	7	11	63	Culture/Biopsy	70–90%

Table 2

Average bias (AB), average length (AL) and average coverage (AC) of 95% credible intervals across 500 simulated datasets with individual study sample size ranging from n = 50–200.

		***J* = 10 studies**			***J* = 35 studies**		
		Prior distribution					
Parameter	**Statistic**	**I**	**NI**	**GS**	**I**	**NI**	**GS**
S_1	AB	0	0.01	0.15	0.03	0.03	0.15
	AL	0.23	0.24	0.23	0.12	0.13	0.10
	AC	1	1	0.09	0.98	0.95	0
C_1	AB	0.03	0.04	0.13	0.02	0.03	0.13
	AL	0.23	0.25	0.2	0.12	0.13	0.10
	AC	0.96	0.95	0.09	0.95	0.93	0
σ_α	AB	0.07	0.04	0.42	0	0.01	0.45
	AL	1.63	1.64	0.74	1.12	1.13	0.37
	AC	0.99	0.98	0.56	0.96	0.95	0
σ_θ	AB	0.02	0.00	0.12	0.02	0.01	0.14
	AL	0.98	0.96	0.55	0.47	0.47	0.23
	AC	0.97	0.98	0.90	0.97	0.96	0.44
β	AB	0.23	0.28	0.14	0.2	0.24	0.05
	AL	1.35	1.36	1.18	1.19	1.21	0.80
	AC	1	1	0.98	0.96	0.96	0.94

I: Informative prior distribution, NI: Noninformative prior distribution, GS: Reference test assumed to be gold standard. True values of parameters: S_1 = 0.9, C_1 = 0.9, σ_α = 0.75, σ_θ = 0.5, β = 0.25.

Table 3

Posterior $_{2.5\%}50.0\%_{97.5\%}$ quantiles of sensitivity and specificity of in-house nucleic acid amplification tests from three meta-analysis models for TB pleuritis data.

	Model* with informative prior (DIC = 150.7)		Model* with noninformative prior (DIC = 145.7)		Model with perfect reference (DIC = 151.2)	
Study	S_1	C_1	S_1	C_1	S_1	C_1
1	$_{0.13}0.30_{0.64}$	$_{0.92}1.00_{1.00}$	$_{0.15}0.54_{0.90}$	$_{0.90}0.99_{1.00}$	$_{0.30}0.49_{0.66}$	$_{0.95}0.99_{1.00}$
2	$_{0.13}0.50_{0.88}$	$_{0.93}1.00_{1.00}$	$_{0.17}0.64_{0.94}$	$_{0.89}0.99_{1.00}$	$_{0.15}0.52_{0.81}$	$_{0.89}0.98_{1.00}$
3	$_{0.28}0.50_{0.78}$	$_{0.93}1.00_{1.00}$	$_{0.31}0.64_{0.93}$	$_{0.87}0.99_{1.00}$	$_{0.56}0.82_{0.96}$	$_{0.90}0.98_{1.00}$
4	$_{0.41}0.68_{0.94}$	$_{0.87}1.00_{1.00}$	$_{0.38}0.81_{0.99}$	$_{0.73}0.96_{1.00}$	$_{0.77}0.94_{1.00}$	$_{0.79}0.89_{0.96}$
5	$_{0.60}0.77_{0.92}$	$_{0.94}1.00_{1.00}$	$_{0.54}0.80_{0.97}$	$_{0.85}0.98_{1.00}$	$_{0.68}0.87_{0.98}$	$_{0.94}0.99_{1.00}$
6	$_{0.33}0.53_{0.72}$	$_{0.94}1.00_{1.00}$	$_{0.33}0.57_{0.81}$	$_{0.89}0.99_{1.00}$	$_{0.39}0.62_{0.82}$	$_{0.90}0.99_{1.00}$
7	$_{0.46}0.65_{0.84}$	$_{0.95}1.00_{1.00}$	$_{0.44}0.69_{0.92}$	$_{0.90}0.99_{1.00}$	$_{0.54}0.75_{0.90}$	$_{0.93}0.99_{1.00}$
8	$_{0.55}0.75_{0.94}$	$_{0.94}1.00_{1.00}$	$_{0.51}0.78_{0.97}$	$_{0.86}0.98_{1.00}$	$_{0.61}0.79_{0.92}$	$_{0.94}0.98_{1.00}$
9	$_{0.36}0.60_{0.83}$	$_{0.94}1.00_{1.00}$	$_{0.38}0.65_{0.91}$	$_{0.89}0.99_{1.00}$	$_{0.44}0.70_{0.89}$	$_{0.91}0.99_{1.00}$
10	$_{0.27}0.43_{0.62}$	$_{0.96}1.00_{1.00}$	$_{0.28}0.49_{0.84}$	$_{0.94}0.99_{1.00}$	$_{0.30}0.46_{0.62}$	$_{0.96}0.99_{1.00}$
11	$_{0.56}0.71_{0.86}$	$_{0.90}1.00_{1.00}$	$_{0.51}0.74_{0.93}$	$_{0.84}0.97_{1.00}$	$_{0.61}0.75_{0.86}$	$_{0.84}0.92_{0.97}$
Pooled	$_{0.42}0.58_{0.73}$	$_{0.97}1.00_{1.00}$	$_{0.43}0.67_{0.87}$	$_{0.92}0.99_{1.00}$	$_{0.53}0.71_{0.84}$	$_{0.94}0.98_{1.00}$

*Model adjusted for conditional dependence in addition to imperfect reference; DIC: Deviance Information Criterion.