Statistics
in Medicine

Research Article

Received 04 June 2010 Accepted 17 May 2011 Published online 22 July 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4320

# Modeling continuous diagnostic test data using approximate Dirichlet process distributions

**Martin Ladouceur,[a] Elham Rahme,[b] Patrick Bélisle,[b] Allison N. Scott,[a] Kevin Schwartzman[a,c] and Lawrence Joseph[a,b*†]**

**There is now a large literature on the analysis of diagnostic test data. In the absence of a gold standard test, latent class analysis is most often used to estimate the prevalence of the condition of interest and the properties of the diagnostic tests. When test results are measured on a continuous scale, both parametric and nonparametric models have been proposed. Parametric methods such as the commonly used bi-normal model may not fit the data well; nonparametric methods developed to date have been relatively complex to apply in practice, and their properties have not been carefully evaluated in the diagnostic testing context. In this paper, we propose a simple yet flexible Bayesian nonparametric model which approximates a Dirichlet process for continuous data. We compare results from the nonparametric model with those from the bi-normal model via simulations, investigating both how much is lost in using a nonparametric model when the bi-normal model is correct and how much can be gained in using a nonparametric model when normality does not hold. We also carefully investigate the trade-offs that occur between flexibility and identifiability of the model as different Dirichlet process prior distributions are used. Motivated by an application to tuberculosis clustering, we extend our nonparametric model to accommodate two additional dichotomous tests and proceed to analyze these data using both the continuous test alone as well as all three tests together. Copyright © 2011 John Wiley & Sons, Ltd.**

**Keywords:** nonparametric Bayesian analysis; diagnostic test; Dirichlet process prior; latent class model; receiver operating characteristic curve; sensitivity and specificity; identifiability

## 1. Introduction

Tracking the transmission pathways of infectious diseases such as tuberculosis (TB) is important in planning resource allocation and infection control strategies. In TB, it is of interest to know the proportion of recently transmitted cases versus those that are reactivations of previously acquired infection. The nearest genetic distance (NGD) is a continuous measure of genetic closeness in which lower values are indicative of a higher probability of recent transmission [1]. As a relatively new measure, the distributions of NGD values within recent and reactivated cases are not well known, but distributions with long tails are possible, meaning the standard bi-normal model may not fit well, and a nonparametric model may be preferable. Two dichotomous tests are also available which use different DNA sequences, MIRU [2] and spoligotyping [3]. In order to make full use of all available data, a model including a nonparametric component for the NGD data and dichotomous components for the other two tests is desirable.

An additional complication is that none of the three available tests provide perfect diagnoses of recent versus reactivated cases. For the two dichotomous tests, this implies that some individuals are

[a]*Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montreal, Quebec, H3A 1A2, Canada*
[b]*Division of Clinical Epidemiology, McGill University Health Centre, Royal Victoria Hospital, 687 Pine Avenue West, V Building, Montreal, Quebec, H3A 1A1, Canada*
[c]*Respiratory Epidemiology Unit, Montreal Chest Institute, 3650 Saint Urbain Street, Montreal, Quebec, H2X 2P4*
**Correspondence to: Lawrence Joseph, Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montreal, Quebec, H3A 1A2, Canada.*
[†]*E-mail: lawrence.joseph@mcgill.ca*

misclassified, and for the continuous test, it means that the distributions of test data with recent and non-recent transmission groups overlap. This occurs because the NGD measure does not directly detect the presence or absence of a recent transmission of TB but rather measures the degree to which a marker for the recent transmission is manifest.

In general, ignoring the imperfections of diagnostic tests leads to biased estimates of the prevalence and other parameters of interest, including the sensitivity, specificity, and predictive values for dichotomous tests [4, 5] and receiver operating characteristic (ROC) curves for continuous tests. Many authors have developed statistical methods to estimate disease prevalence and properties of diagnostic tests for the case of dichotomous imperfect tests [4–10]. Estimates are derived mainly via latent class analysis, with inferences typically involving the EM algorithm or the Gibbs sampler. Recently, articles on modeling continuous diagnostic test data in the absence of a gold standard have appeared, including Scott *et al.* [11] and Branscum *et al.* [12].

When continuous tests are used, the challenge is to estimate the distributions of test results within truly positive and truly negative groups of individuals, from which the test properties will follow. Complicating the task of density estimation within groups is the classification problem when no gold standard test exists to definitively place each subject into their correct group. Bi-normal models have been proposed [11], but even approximate normality is not guaranteed, and using the wrong distributional shape can lead to biased estimates. Nonparametric models [13, 14] may therefore be preferable, but these models come with their own challenges because they are clearly non-identifiable: If arbitrarily shaped densities are possible, then any subject can be positive or negative regardless of their test value. A careful balance is therefore required between flexibility in the model and *a priori* information concerning likely location and shape of test results within positive and negative sub-populations.

Although several Bayesian nonparametric models have appeared in the literature, the performance of these models in practice has not been extensively investigated. This is especially true in the diagnostic testing problem, where most methods have assumed either that data from a gold standard test are available or that sufficient information is available from other tests to render the model identifiable [14–18]. In this paper, we propose using a simplified version of a Dirichlet processes model nested within a latent class diagnostic testing framework that accounts for the lack of a gold standard test. In particular, we place approximate nonparametric Dirichlet prior distributions directly on the continuous test data within each group. Through simulations using both truly normally and non-normally distributed data within groups, we study the advantages and disadvantages of nonparametric models in the diagnostic testing context, paying particular attention to the effects of changing various prior parameters associated with the Dirichlet process. This allows us to investigate the delicate balance required between model flexibility needed to fit non-normal data and model identifiability required for accurate diagnosis of each subject. Nonparametric latent class models also exist for ordinal diagnostic test data [19, 20].

Several authors, including Albert and Dodd [21] and Pepe and James [22], discuss limitations of latent class models for diagnostic test data. For example, model misspecification, especially concerning any dependence structure between tests, can lead to poor estimation. Further, these problems are not solved by increasing sample size because models can remain non-identifiable; nor can they typically be addressed by comparing models via methods such as Bayes Factors, again because non-identifiability often prevents data alone from distinguishing between competing models. These problems require careful attention, for example, justification as to whether tests may be conditionally dependent or independent based on clinical considerations of physical properties of the tests being used. Researchers using these methods must often admit that their analysis is conditional on their choice of model being correct and run several different models to check for robustness of the main inferences. Another problem is that the definition of a truly positive case is not made explicit by a latent class model, which, rather than classifying patients as positive or negative, provides a probability of disease for each subject. This, however, can be seen as an accurate depiction of the underlying problem when no gold standard classifier exists, and so the model is simply reflecting reality. The issue of conditional independence does not arise in most of our paper, which deals with a single continuous test, although we do briefly discuss this issue in the context of our TB data in Section 4. In that same section, we run several different models for our data and compare results across each.

The outline of the rest of the paper is as follows. In Section 2, we describe our Bayesian nonparametric methods in the context of a latent class model for continuous diagnostic test data. Section 3 describes the simulations used to investigate the properties of our nonparametric model, whereas in Section 4 we apply the models to the analysis of our three tests for recent tuberculosis transmission, comparing our results with those obtained from the standard bi-normal model [11]. As data from additional dichotomous tests

are available in this data set, we first extend our model to include the combination of continuous and dichotomous test data. We conclude with a discussion in Section 5.

## 2. An approximate Dirichlet process applied to continuous diagnostic testing data

Ferguson [23] introduced the Dirichlet process as a means to generate random density functions, which can be viewed as a limit of Dirichlet distributions as the dimension goes to infinity. We notate the process by $G \sim \mathrm{DP}(\alpha G_0)$, where $G_0$ is the baseline distribution, providing the expectation of the process, and $\alpha$ is the precision parameter, which controls the variability of the random probability measure $G$ about $G_0$. Small values of $\alpha$ will allow larger deviations from $G_0$ compared with larger values, and as $\alpha \to \infty$ the variability is minimized, and the density $G$ becomes identical to $G_0$.

Assume that we apply a continuous diagnostic test to a sample of $n$ individuals from a population where $\theta$ is the unknown prevalence of the disease. We assume a Beta prior distribution for the prevalence, $\theta \sim \mathrm{Beta}(\alpha_\theta, \beta_\theta)$. Let $X = \{x_1, \ldots, x_n\}$ be the observed test results across individuals for that diagnostic test. Let $Z = \{z_1, \ldots, z_n\}$ be the latent data that represents the dichotomous unknown true status of each individual, with $z_i = 1$ and $z_i = 0$ denoting individuals with and without the disease, respectively. We assume distinct Dirichlet processes to model the test results within diseased and non-diseased populations, with parameters given by $\alpha_i$ and $G_0^i$, where $i = 1, 2$ for diseased and non-diseased groups, respectively. Let $g_0^i(x_j|\psi_i)$ denote the parametric density function belonging to $G_0^i$ with possibly vector parameter $\psi_i$. We will assume that $g_0$ is a normal density in this paper, but for low values of $\alpha$, a sample from the Dirichlet process can be very far from normal.

Given the continuous data $X$ and latent data $Z$, the likelihood function of a sample from the mixture of two Dirichlet processes assuming no ties is given by Petrone and Raftery [24] as

$$g(X, Z|\theta, \alpha_1, \alpha_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \frac{\alpha_1^{n_1(z)}\alpha_2^{n_2(z)}}{\alpha_1^{[n_1(z)]}\alpha_2^{[n_2(z)]}} \prod_{i=1}^{n} (\theta\phi(x_i|\psi_1))^{z_i} ((1-\theta)\phi(x_i|\psi_2))^{(1-z_i)}$$

where $n_1(z) = \sum_{i=1}^n z_i$ and $n_2(z) = n - \sum_{i=1}^n z_i$ denote the number of diseased and non-diseased individuals, respectively, and $\alpha^{[k]} = \alpha(\alpha+1)\ldots(\alpha+k-1)$, with $\alpha^{[0]} = 1$. Note that $\psi_i = (\mu_i, \sigma_i^2)$, where $(\mu_i, \sigma_i^2)$ are the respective means and variances of the assumed normal baseline density functions, $i = 1, 2$. For ease of notation, we replace $g_0^1$ and $g_0^2$ by $\phi$, the density function of a normal distribution.

Our model is equivalent to a mixture of two Dirichlet processes. Even though our data are continuous, samples from a Dirichlet process are discrete with probability 1 [23]. This would not normally be a problem because any continuous density can be reasonably approximated by a discrete function, but Petrone and Raftery [24] discussed potential inferential problems caused by this discreteness property in assigning group membership in mixture problems of the type we have here. In particular, for very low values of $\alpha$, the inferences obtained about the prevalence of disease can potentially be biased towards 50%. Intuitively, this arises from the flexibility of nonparametric densities because if the distributions within diseased and non-diseased groups can take on almost any shape over the range of data as implied by low values of $\alpha$, then it becomes very difficult to separate out who may be diseased versus non-diseased on the basis of this continuous measure. This leads to a classification no better than chance and thus an estimated prevalence of 50%. This problem is not unique to mixtures of Dirichlet processes and will occur with any highly flexible nonparametric model. Thus, the two major objectives of the simulations in Section 3 are to evaluate the practical impact of the bias identified by Petrone and Raftery [24] and to investigate the use of different values of $\alpha$, which controls the degree of flexibility of the nonparametric distributions. Thus, we carefully investigate the balance required between flexibility to accommodate non-normality (lower values of $\alpha$ leading to higher flexibility) and identifiability of the models (higher values of $\alpha$ implying more prior knowledge of the within-group densities, leading to better identified models).

Exact inferences are difficult in practice because Dirichlet processes have an infinite number of parameters. Therefore, as suggested by Mira and Petrone [25], we approximated the exact posterior Dirichlet process by a finite Dirichlet distribution by selecting a finite partition of the real line, essentially treating the Dirichlet parameters as representing non-normalized multinomial probabilities on these intervals. Thus, our final model is a finite approximation of a Dirichlet process. If the dimension (number of bins or partitions of the real line) of the multinomial is large, the approximation will be very good in practice.

From our experience, taking 12 to 15 bins across each distribution works well. For example, we used 25 bins for each of the simulations in Section 3.

Let $B = (\beta_1, \ldots, \beta_k)$ be a partition of the real line, and define $N = (n_1, \ldots, n_k)$, where the sample size $n = \sum_{i=1}^{k} n_i$, and $n_i$ represents the number of observations from data $X$ that fall into the partition $\beta_i$, $i = 1, \ldots, k$. Define latent data $Y = (y_1, \ldots, y_k)$, where $y_i$ represents the number of truly diseased subjects out of $n_i$. Note that $y_i$ is the sum of the latent data $Z$ that fall into the interval $\beta_i$.

Within each disease state, the prior probability associated with each partition is defined to be the area under the nonparametric baseline density curve belonging to that disease state for that partition. Let $p_1 = (p_{11}, \ldots, p_{k1})$ and $p_2 = (p_{12}, \ldots, p_{k2})$ be the vectors of partition probabilities within diseased and non-diseased subjects, respectively, where the $p_{j1}$'s and $p_{j2}$'s represent the probabilities for the $j$th partition. For a given partition, the form of the likelihood is multinomial, with cell probabilities given by the vectors $p_1$ and $p_2$ for the diseased and non-diseased populations, respectively. The likelihood function of the observed and latent data can then be expressed as

$$L(X, Y|\theta, p_{11}, \ldots, p_{k1}, p_{21}, \ldots, p_{k2}) \propto \prod_{j=1}^{k} (\theta p_{j2})^{y_j} ((1-\theta) p_{j1})^{n_j - y_j} \ . \tag{1}$$

Letting the prior distributions for the vectors $p_1$ and $p_2$ be Dirichlet with parameters $(\alpha_i G_0^i(\beta_1), \ldots, \alpha_i G_0^i(\beta_k))$, $i \in \{1, 2\}$, we have

$$g(p_{11}, \ldots, p_{k1}) \propto \prod_{j=1}^{k} p_{j1}^{\alpha_1 G_0^1(\beta_j)-1} \quad \text{and} \quad g(p_{12}, \ldots, p_{k2}) \propto \prod_{j=1}^{k} p_{j2}^{\alpha_2 G_0^2(\beta_j)-1}.$$

The posterior density is formed by multiplying the prior distributions discussed above with the likelihood function. Inference then proceeds using the Gibbs sampler. An R implementation of these methods is available from the authors.

In Section 4, we return to the analysis of our TB data. First, however, we will investigate the performance of our simple nonparametric latent class model in practice through a series of simulations.

## 3. Evaluating the performance of a nonparametric latent class model

To evaluate our model, we carried out a simulation study under three main sets of conditions. The first two sets examine simulated data that arise from normal distributions with differing degrees of prior information, whereas the third set uses non-normally distributed data. The first two sets allow us to compare the performance of our model with the 'correct' bi-normal model, when nonparametric flexibility is not needed. The third set will show what is gained from using a flexible model when the bi-normal model does not fit the data well. For each set of conditions, we controlled the degree of overlap between the two distributions as well as the sample sizes within both diseased and non-diseased groups, which in turn imply a pre-specified disease prevalence when the sample is representative of the population. The results will provide clues about the properties of our model, guide the choice of prior parameters, and estimate the magnitude of the bias raised by Petrone and Raftery [24]. This understanding is crucial to the effective use of the model in practice. Because simulation set #1 is intended to investigate the effect of changing prior information on the bias, we ran only a single example for each combination of parameters. For simulation sets #2 and #3, we created and analyzed 200 data sets for each scenario and report average bias and credible interval coverage for our nonparametric model across the 200 estimates.

### *3.1. The scenarios investigated*

We used two sample sizes, $n = 1000$ and $n = 5000$, and both large and small degrees of overlap, defined by areas under the ROC curve (AUC) of 0.70 and 0.90, respectively. These were intended to simulate data from diagnostic tests with moderate to very good test properties, as measured by the degree of separation between test results from positive and negative subjects. Taking the four possible combinations of two sample sizes and two AUC values and keeping the prevalence at $\theta = 0.25$ created four scenarios. We kept the prior information about the prevalence fixed at a uniform density on [0,1] and assumed that our two precision parameters $\alpha_1$ and $\alpha_2$ were equal, meaning in practice that there were equal amounts

of prior information about the positive and negative test result distributions. These choices kept the total number of simulations to a reasonable level while still covering a range of realistic scenarios.

The sample sizes may seem larger than those found in many studies using diagnostic tests, and we initially investigated the model properties with a sample size of 400. However, even under ideal conditions using very strong prior distributions all centered on the correct values, it became clear that a size of $n = 400$ is insufficient for accurate results. Therefore, results for $n = 400$ are not reported, and we emphasize that it will not usually be possible to accurately estimate the prevalence of a disease or diagnostic test properties in a latent class model with such small sample sizes. This conclusion holds not only for our nonparametric model but also for the standard bi-normal model in the absence of a gold standard test. To provide an idea as to why this is true, consider a binomial distribution with probability of success of $\theta = 0.5$, corresponding to a prevalence of 50%, and 400 trials. If the exact number of successes were known, then the binomial confidence interval will have a width of approximately 10%. However, in our model each binomial outcome is not directly observed but is a latent variable, so that this 10% value will be the lower bound for the interval width. Further, our simulations showed that the actual widths will be much larger than 10% in most practical situations, even when very strong prior information is used. One way around this is to use additional tests, as we discuss in Section 4.

*3.1.1. Data generation and prior distributions for simulation set #1.* In this first set of simulations, we generated continuous test results from two sets of overlapping normal distributions, first $N(0, 1)$ and $N(2, 1)$(AUC $= 0.9$), and then $N(0, 1)$ and $N(0.75, 1)$(AUC $= 0.7$) for the diseased and non-diseased groups, respectively. We assume that the population distributions are known and thus use 'perfect' prior information that match the true parameter values. That is, we used point prior values matching the true means and variances when using the bi-normal model and used baseline densities $g$ that match the true densities for the nonparametric models. Assuming that the normal means and variances are exactly known is of course unrealistic but is helpful as a preliminary step towards understanding the effect of the choice of nonparametric precision parameter, $\alpha$, and to investigate the possible bias issue raised by Petrone and Raftery [24] without distraction from other modeling issues. We considered $\alpha$ values of 5, 10, 50, 100, 250, 500, 750, 1000, 5000, 10,000, and 100,000.

*3.1.2. Data generation and prior distributions for simulation set #2.* Data generation for our second set of simulations is identical to that from simulation set #1, but we now begin to assume less than perfect prior information, both in the bi-normal model and for the nonparametric baseline densities. We do so in two stages. First, we will center the prior densities about their true values but now with some uncertainty about these values. In the second stage, we will retain this uncertainty about the means and variances but also move the centers of the prior densities away from the true values to estimate the extent to which our model is able to use the data to recover the true values despite 'off-centered' prior densities. In particular, when AUC $= 0.9$ for the parametric models, we start by assuming that the means of the diseased, $\mu_1$, and non-diseased, $\mu_2$, have $N(0, 1)$ and $N(2, 1)$ prior distributions, respectively, and become $N(-1, 1)$ and $N(3, 1)$, respectively, in the second set of simulations. When AUC $= 0.7$, the prior distributions for the mean of the $D+$ and $D-$ distributions in the first and second set of simulations are $N(0, 0.25)$, $N(0.75, 0.25)$, $N(-1, 0.25)$, and $N(1.75, 0.25)$, respectively. Across all cases, the prior distributions for the standard deviations remain fixed at a uniform density on the range $[0.8,\ 1.2]$.

Similar to the choices in the preceding paragraphs, for the nonparametric models, when AUC $= 0.9$, the distributions $g_0^1$ and $g_0^2$ in the first and second set of simulations are $N(0, 1)$, $N(2, 1)$, $N(-1, 1)$, and $N(3, 1)$, respectively. When AUC $= 0.7$, the corresponding distributions are changed to $N(0, 1)$, $N(0.75, 1)$, $N(-1, 1)$, and $N(1.75, 1)$. We varied the precision parameters to be $\alpha = 10$, 100, or 1000, corresponding to small, moderate, and large amounts of prior information about the shapes of the density of the test results.

*3.1.3. Data generation and prior distributions for simulation set #3.* One important objective of our simulations was to investigate the difference between our nonparametric model and a standard bi-normal model when the normality assumption is violated. Here we generated test result data from a $\chi^2_3$ density for the $D+$ population and from an equal mixture of two normal densities for the $D-$ population. The means and standard deviations of the two normal densities were chosen according to the desired AUC. To obtain an area under the ROC curve of 0.9, we used a mixture of a $N(5, 1.5)$ and a $N(12, 1.5)$. For an AUC of 0.7, we used a mixture of a $N(3, 1.5)$ and a $N(5.5, 1.5)$.

We will continue to use normal baseline distributions in our nonparametric models, and so do not assume that we know the density shape *a priori*. We therefore investigate whether the flexibility gained by the nonparametric model results in reasonable estimation of the prevalence and test characteristics and provides at the same time a good estimate of the shapes of the distributions of the non-normal test results.

Table I summarizes the parametric and nonparametric prior parameter values that will be investigated in simulation set #3. The first group of prior parameters closely matched the means and variances (but not the shapes) found in the data, whereas the second group were moved off-center by one standard deviation, similar to simulation set #2. As in the preceding paragraphs, we used $\alpha$ values of 10, 100, or 1000.

### 3.2. *Results from the simulations*

We focus on reporting results for the prevalence because the model performance for this parameter should be similar to that for the other parameters. Indeed, in order to estimate the prevalence accurately, one needs to accurately ascertain which subjects are true positives or true negatives, in turn implying reasonable estimation for the sensitivity and specificity leading to ROC curves. Although one can conceive that the prevalence is well estimated because of equally poor estimation of both sensitivity and specificity whose misclassified subjects balance out to provide an accurate estimate of the prevalence, this does not generally occur [26, 27].

*3.2.1. Results from simulation set #1.* Median prevalence estimates from one run of each nonparametric model as a function of the precision parameter $\alpha$, sample size, and AUC are displayed in Figure 1. Estimates for $\alpha > 1000$ are omitted because they were similar to those from a model with $\alpha = 1000$. The results for $\alpha = 1000$ are also very similar to those from the exact parametric bi-normal model because the likelihood function (1) converges to the bi-normal model as $\alpha \longrightarrow \infty$.

Under the ideal conditions implied by this set of simulations, the prevalence of $\theta = 0.25$ is well estimated for values of $\alpha$ larger than 100, except for the case of small sample size ($n = 1000$) and low AUC (0.7), where values of $\alpha \approx 400$ are required to render the size of the bias negligible. Thus, 'low information' priors can indeed have the undesirable side effect of biasing prevalence estimates towards

**Table I.** Parametric and nonparametric prior distributions used for simulation set #3. In all nonparametric scenarios, we fit models with three choices for the precision parameter, $\alpha \in \{10, 100, 1000\}$. The data for the $D+$ group were generated from a $\chi^2_3$ density, whereas the $D-$ data were generated from a normal mixture, as described in Section 3.1.

| AUC = 0.9 | AUC = 0.7 |
|---|---|
| Parametric prior 1 | |
| $\mu_1 \sim N(2.366, 0.25)$ | $\mu_1 \sim N(2.366, 0.25)$ |
| $\mu_2 \sim N(8.5, 0.25)$ | $\mu_2 \sim N(4.25, 0.25)$ |
| $\sigma_1 \sim \text{unif}(2.25, 2.65)$ | $\sigma_1 \sim \text{unif}(2.25, 2.65)$ |
| $\sigma_2 \sim \text{unif}(3.6, 4)$ | $\sigma_2 \sim \text{unif}(3.6, 4)$ |
| Parametric prior 2 | |
| $\mu_1 \sim N(1.366, 0.25)$ | $\mu_1 \sim N(1.366, 0.25)$ |
| $\mu_2 \sim N(9.5, 0.25)$ | $\mu_2 \sim N(5.25, 0.25)$ |
| $\sigma_1 \sim \text{unif}(2.25, 2.65)$ | $\sigma_1 \sim \text{unif}(2.25, 2.65)$ |
| $\sigma_2 \sim \text{unif}(3.6, 4)$ | $\sigma_2 \sim \text{unif}(3.6, 4)$ |
| Nonparametric prior 1 | |
| $g_0^1 = N(2.366, \sqrt{6})$ | $g_0^1 = N(2.366, \sqrt{6})$ |
| $g_0^2 = N(8.5, 3.8)$ | $g_0^2 = N(4.25, 1.95)$ |
| Nonparametric prior 2 | |
| $g_0^1 = N(1.366, \sqrt{6})$ | $g_0^1 = N(1.366, \sqrt{6})$ |
| $g_0^2 = N(9.5, 3.8)$ | $g_0^2 = N(5.25, 1.95)$ |

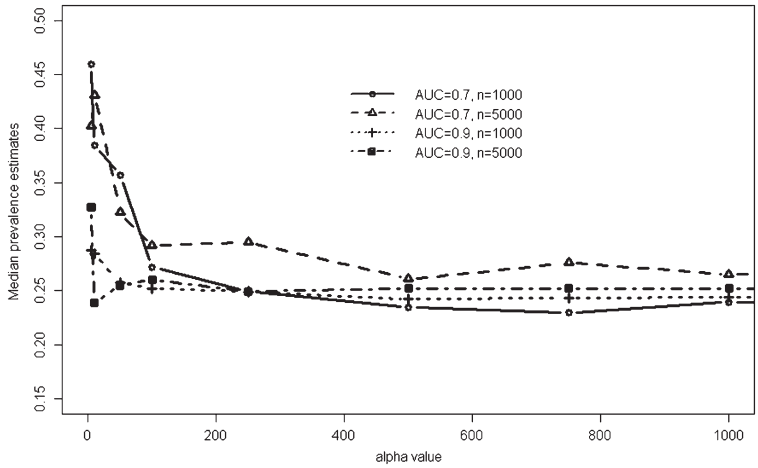


**Figure 1.** Median prevalence estimates as a function of the precision parameter, the sample size, and the degree of overlap between the two distributions.

50% as originally discussed by Petrone and Raftery [24], but this effect becomes negligible as the value of $\alpha$ increases, the rate depending on the sample size and degree of overlap between $D+$ and $D-$ test results.

These results suggest that when the distributions of test results in the two populations are normally distributed and when a high degree of information concerning their means and variances are known *a priori*, it is possible to obtain accurate prevalence estimates from both the parametric and nonparametric models, as long as $\alpha$ is not set too low. However, researchers will not generally know the exact means and variances of the two distributions of test results nor know that they are exactly normally distributed. We next compare the two models when less prior information about these parameters is assumed.

*3.2.2. Results from simulation set #2.* Table II provides all prevalence estimates from simulation set #2. When the AUC $= 0.9$, the credible intervals from all four parametric models capture the true prevalence estimate regardless of the prior distributions used. However, the choice of prior density is important to nonparametric model performance. Under the first set of prior densities, where the prior distributions are centered on the true mean values, all models performed extremely well, with very low bias and high coverage probabilities. However, when prior distributions were not centered on the true values, the results depended on the value of $\alpha$ chosen and on the sample size. Using a high value for $\alpha$ combined with poor prior information creates a 'false certainty' that leads to poor performance because it is difficult for the nonparametric model to move back towards the true densities when high prior values of $\alpha$ are used. Thus, coverages tend to be low. With $\alpha = 10$, however, the prevalence is much better estimated, as the weight of the 'false' prior information is decreased.

When the AUC $= 0.7$, all four parametric models again capture the true prevalence estimate but with extremely wide credible intervals. Obviously, when the two test result distributions overlap considerably, latent class models will have difficulty distinguishing diseased from non-diseased subjects. We therefore also tried to estimate the prevalence under more precise prior distributions, specifically $\mu_1 \sim N(0, 0.10)$, $\mu_2 \sim N(0.75, 0.10)$, and $\sigma_1, \sigma 2 \sim \text{unif}(0.99, 1.01)$, but the credible intervals for the prevalence remained very wide. Therefore, the parametric models do not perform well even with quite strong prior information when the distributions of test results are not widely separated.

Conversely, under centered prior distributions, the nonparametric models performed very well when AUC $= 0.7$, except when $\alpha = 10$, where the bias towards 50% remained large. When the baseline distributions are off centered (prior set #2), the model still performed well when $\alpha = 100$ or 1000, except when $\alpha = 1000$ and the sample size was 1000, the relatively small sample size not being informative enough to overcome the strong prior information, unlike the case when a larger sample size of 5000 is used. Once again, bias towards 50% is evident for models using $\alpha = 10$, and coverages are low.

Figure 2 illustrates the true ROC curve along with estimates from the various models when the AUC $= 0.9$. Among the 16 models, 10 (63%) had well-estimated ROC curves that were similar in shape and AUC to the true curve. These included all of the parametric models. The nonparametric models in

**Table II.** Posterior medians and 95% credible intervals for the prevalence (true value $\theta = 0.25$) for a first run across all models related to simulation set #2. In addition, bias and coverage probabilities based on 200 repetitions of each scenario are given for the nonparametric models. Data were simulated from two normal densities. The two sets of prior distributions are detailed in Section 3.1.

| | AUC = 0.9 | | AUC = 0.7 | |
|---|---|---|---|---|
| | $n = 1000$ | $n = 5000$ | $n = 1000$ | $n = 5000$ |
| | Parametric—first (centered) set of prior distributions | | | |
| | 0.226 | 0.216 | 0.239 | 0.317 |
| | [0.122; 0.456] | [0.151; 0.333] | [0.018; 0.824] | [0.033; 0.807] |
| | Parametric—second (not centered) set of prior distributions | | | |
| | 0.215 | 0.215 | 0.163 | 0.195 |
| | [0.116;0.428] | [0.151; 0.329] | [0.003; 0.540] | [0.006; 0.513] |
| | Nonparametric—first (centered) set of prior distributions | | | |
| $\alpha = 1000$ | 0.234 | 0.244 | 0.230 | 0.265 |
| 95% CrI | [0.192; 0.276] | [0.220; 0.268] | [0.119; 0.332] | [0.198; 0.335] |
| (Bias, coverage) | (0.002, 0.995) | (0.001, 1.000) | (0.002, 0.980) | (0.013, 1.000) |
| $\alpha = 100$ | 0.276 | 0.274 | 0.302 | 0.360 |
| 95% CrI | [0.213; 0.341] | [0.226; 0.325] | [0.107;0.489] | [0.185; 0.478] |
| (Bias, coverage) | (0.016, 1.000) | (0.017, 1.000) | (0.096, 0.975) | (0.101, 0.940) |
| $\alpha = 10$ | 0.324 | 0.290 | 0.458 | 0.465 |
| 95% CrI | [0.225; 0.430] | [0.216; 0.391] | [0.274;0.613] | [0.323;0.614] |
| (Bias, coverage) | (0.077, 1.000) | (0.079, 0.885) | (0.202, 0.130) | (0.202, 0.090) |
| | Nonparametric—second (not centered) set of prior distributions | | | |
| $\alpha = 1000$ | 0.194 | 0.153 | 0.332 | 0.254 |
| 95% CrI | [0.157; 0.232] | [0.134; 0.176] | [0.282; 0.385] | [0.213; 0.299] |
| (Bias, coverage) | (−0.020, 0.885) | (−0.102, 0.000) | (0.091, 0.035) | (−0.005, 0.945) |
| $\alpha = 100$ | 0.197 | 0.147 | 0.333 | 0.231 |
| 95% CrI | [0.136; 0.259] | [0.112; 0.197] | [0.216; 0.458] | [0.127; 0.383] |
| (Bias, coverage) | (−0.069, 0.410) | (−0.102, 0.000) | (0.047, 0.945) | (0.006, 1.000) |
| $\alpha = 10$ | 0.314 | 0.339 | 0.385 | 0.414 |
| 95% CrI | [0.220; 0.395] | [0.201; 0.402] | [0.260; 0.536] | [0.291; 0.544] |
| (Bias, coverage) | (0.061, 0.985) | (0.056, 0.800) | (0.168, 0.060) | (0.161, 0.120) |

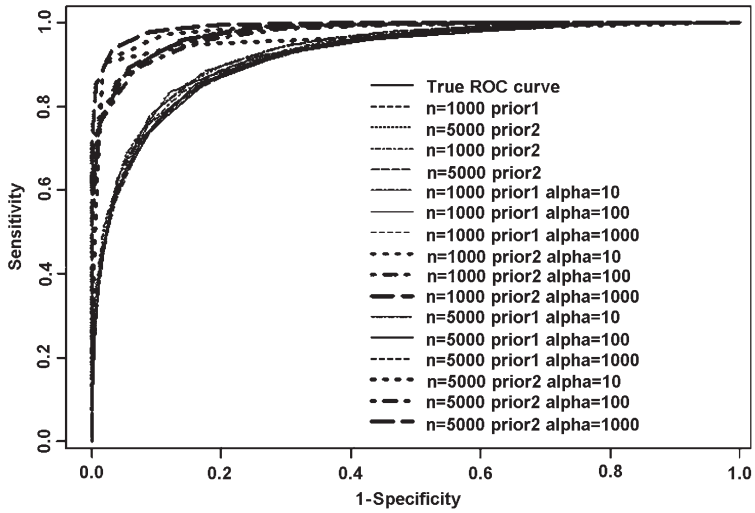


**Figure 2.** True ROC curve and ROC curves estimated from all 16 models when AUC = 0.9. The top six ROC curves are obtained from nonparametric models where the baseline distributions were off-centered, indicated as prior2, whereas prior1 indicates centered prior distributions.

which the baseline densities matched the true density also gave reasonable ROC curves, but when the baseline densities were off-centered, the ROC curves were too high because the densities were estimated to be further apart than they really were.

**Table III.** Posterior medians and 95% credible intervals for the prevalence (true value $\theta = 0.25$) for one run across all models related to simulation set #3. In addition, bias and coverage probabilities based on 200 repetitions of each scenario are given for the nonparametric models. Data were simulated from non-normal densities as described in Section 3.1, which also details two sets of prior distributions.

| | AUC = 0.9 | | AUC = 0.7 | |
|---|---|---|---|---|
| | $n = 1000$ | $n = 5000$ | $n = 1000$ | $n = 5000$ |
| | Parametric—first (centered) set of prior distributions | | | |
| | 0.403 | 0.482 | 0.025 | 0.253 |
| | [0.347; 0.454] | [0.461; 0.501] | [0.001; 0.014] | [0.002; 0.465] |
| | Parametric—second (not centered) set of prior distributions | | | |
| | 0.426 | 0.479 | 0.009 | 0.001 |
| | [0.376; 0.473] | [0.457; 0.500] | [0.003; 0.053] | [0.000; 0.007] |
| | Nonparametric—first (centered) set of prior distributions | | | |
| $\alpha = 1000$ | 0.272 | 0.224 | 0.172 | 0.112 |
| | [0.220; 0.326] | [0.183; 0.266] | [0.090; 0.263] | [0.059; 0.162] |
| (Bias, coverage) | (0.011, 0.985) | (−0.029, 0.885) | (−0.106, 0.190) | (−0.137, 0.000) |
| $\alpha = 100$ | 0.258 | 0.239 | 0.207 | 0.172 |
| | [0.156; 0.346] | [0.140; 0.330] | [0.074; 0.383] | [0.061; 0.350] |
| (Bias, coverage) | (0.011, 1.000) | (−0.001, 1.000) | (−0.040, 1.000) | (−0.049, 1.000) |
| $\alpha = 10$ | 0.355 | 0.358 | 0.405 | 0.399 |
| | [0.243; 0.476] | [0.248; 0.471] | [0.222; 0.573] | [0.270; 0.532] |
| (Bias, coverage) | (0.112, 0.740) | (0.111, 0.505) | (0.172; 0.510) | (0.171; 0.380) |
| | Nonparametric—second (not centered) set of prior distributions | | | |
| $\alpha = 1000$ | 0.243 | 0.141 | 0.264 | 0.125 |
| | [0.193; 0.294] | [0.113; 0.173] | [0.202; 0.331] | [0.100; 0.155] |
| (Bias, coverage) | (−0.009, 0.995) | (−0.106, 0.000) | (0.034, 0.865) | (−0.103, 0.000) |
| $\alpha = 100$ | 0.175 | 0.120 | 0.231 | 0.150 |
| | [0.106; 0.260] | [0.072; 0.224] | [0.116; 0.393] | [0.091; 0.306] |
| (Bias, coverage) | (−0.062, 0.935) | (−0.114, 0.215) | (−0.023, 1.000) | (−0.085, 0.870) |
| $\alpha = 10$ | 0.339 | 0.366 | 0.410 | 0.418 |
| | [0.226; 0.466] | [0.256; 0.473] | [0.224; 0.541] | [0.298; 0.542] |
| (Bias, coverage) | (0.092, 0.935) | (0.091, 0.690) | (0.160, 0.285) | (0.158, 0.265) |

*3.2.3. Results from simulation set #3.* Recall that in this set of simulations the underlying data were non-normally distributed, so that a nonparametric model was needed. As shown in Table III, even when the AUC = 0.9, all four parametric estimates are far from the true value, and their credible intervals are too narrow, leading to very poor estimation. The latent class models have great difficulty distinguishing the two distributions when forced to falsely assume normality.

For the nonparametric models with centered baseline distributions and precision parameters of 100 or 1000, the prevalence is well estimated, even though the Dirichlet model's baseline shape is incorrect, being normally distributed. Biases tend to be small and coverages very high. When $\alpha = 10$, however, the prevalence estimate is biased towards 50%, as seen previously, and coverages drop correspondingly. When the baseline distributions are off center, coverages depend largely on the sample size, which dictates the size of the interval. With sample sizes of 5000, the intervals are too narrow to include the true prevalence value much of the time. Nevertheless, even when the data are not normally distributed, when $\alpha = 100$ or 1000 and the sample size is 1000, there is sufficient flexibility to let the data alter the shape of these distributions, as shown in Figure 3. Thus, the nonparametric models perform considerably better than the parametric models under non-normality.

When the AUC = 0.7, the prevalence estimates across the parametric models are heavily biased, and most have very large credible intervals. The nonparametric model tends to work well regardless of prior choice for $\alpha = 100$. Despite the difficult nature of the problem owing to the large overlap and non-normality of the test results, the nonparametric models with $\alpha = 100$ are flexible enough to mold to the distributional shapes. However, for $\alpha = 10$ the bias towards 50% remained large, and for $\alpha = 1000$ the strong prior information did not allow sufficient flexibility to stray from baseline normality, leading once again to poor results.

Figure 3 displays the density estimates of the $D+$ and $D-$ distributions. These curves were estimated by a density smoothing of the points sampled within each iteration of the Gibbs sampler output.

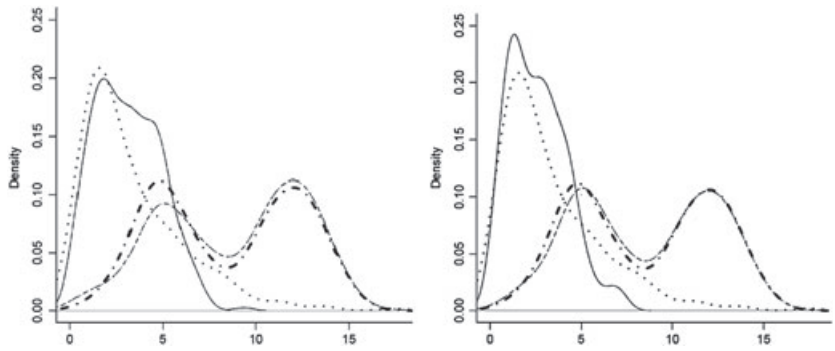


**Figure 3.** Posterior density estimates of the $D+$ and $D-$ test result distributions from a Dirichlet process prior with sample size $n = 1000$, AUC $= 0.9$, and precision parameter $\alpha = 10$. Dotted and slashed curves correspond to the true $D+$ and $D-$ test result distributions, respectively. Left and right figures are obtained when the baseline distribution $G_0$ is centered (prior set #1) or not centered (prior set #2) around the true mean test result values, respectively.

As expected, the prior distributions become more flexible as the precision parameter $\alpha$ decreases, so that even when $\alpha = 10$, the $D-$ distribution remains well estimated, at least in terms of identifying the bi-modality. The shape of the $D+$ distribution is slightly shifted to the right compared with its true location. Given the long tail of the chi-square distribution, along with the overlap in the region between 3 and 5, the tails of the $D+$ distributions are estimated to be somewhat shorter than the true density. Despite these imperfections, the nonparametric models were generally able to track the $D-$ and $D+$ density shapes despite the baseline densities being far from the true shapes.

## 4. Adding dichotomous test results to the model with application to detecting recent TB transmission

We now turn to the analysis of the data set described in Section 1. Aside from the continuous NGD values from each subject, two dichotomous tests are available which use different DNA sequences, MIRU [2] and spoligotyping [3]. As the methods of action and sequences used in these tests are all different, we assume conditional independence given the true (latent) transmission route, as did Scott *et al.* [11] in analyzing these data using a bi-normal model. If necessary, dependence between the dichotomous tests can be handled via methods described by Dendukuri and Joseph [28] or Black and Craig [8]. Accounting for possible correlations when using nonparametric models is an important topic for future research.

We will first provide inferences using the continuous test alone and then combine this test with the two additional dichotomous tests. The existence of three tests rather than a single test provides some hope that reasonable inferences (that is, credible interval lengths close to the lower bound discussed in Section 3.1) can be obtained despite our small sample size of $n = 393$ [29]. We first extend our nonparametric continuous test model to accommodate two additional dichotomous tests and then apply our model to the data from all three tests.

### *4.1. Extending the nonparametric model to include dichotomous tests*

For each subject, each dichotomous test could either be positive or negative (in our context, indicating recent transmission or reactivation, respectively). Given the true status of an individual and a partition $B_i$, $i = 1, 2, \ldots, k$, of the real line, the continuous test results could fall into any one of the $k$ bins defined by the partition. Let these occur with probabilities $P_{i1}$ for subjects who are truly positive and $P_{i0}$ for those truly negative. For each continuous data point that falls within any bin, there are eight possible combinations: two choices for each dichotomous test, as well as a third binary variable representing the latent true status for each individual. Let $N_{ijlm}$, $i = 1, 2, \ldots, k$, $j = 1, 2$, $l = 1, 2$, $m = 1, 2$ represent the number of subjects whose continuous data fall into bin $B_i$, whose true (latent) status $j$ is either positive ($j = 1$) or negative ($j = 0$), and whose dichotomous test results are either positive or negative on each test, with possible combinations $lm = (1, 1)$, $(1, 0)$, $(0, 1)$, or $(0, 0)$. Let the two

dichotomous test sensitivities and specificities be denoted by $(S_1, S_2)$ and $(C_1, C_2)$, respectively. This leads to the likelihood function for the observed and latent data

$$\prod_{i=1}^{k}\prod_{j=0}^{1}\prod_{l=0}^{1}\prod_{m=0}^{1}\left[\left\{\pi S_1^l(1-S_1)^{(1-l)}S_2^m(1-S_2)^{(1-m)}P_{i1}\right\}^{j\times N_{ijlm}}\right. \tag{2}$$
$$\left.\times\quad\left\{(1-\pi)C_1^{(1-l)}(1-C_1)^l C_2^{(m-1)}(1-C_2)^m P_{i0}\right\}^{(1-j)\times N_{ijlm}}\right].$$

Independent Beta prior distributions will be used for the prevalence and sensitivities and specificities of each dichotomous test, so that $\theta \sim \text{Beta}(\alpha_\pi, \beta_\pi)$, $S_m \sim \text{Beta}(\alpha_{Sm}, \beta_{Sm})$, and $C_m \sim \text{Beta}(\alpha_{Cm}, \beta_{Cm})$, $m = 1, 2$. As in our one test model, Dirichlet prior distributions are used for the continuous test results within positive and negative subgroups, and multiplying these prior densities by the likelihood function in equation (2) provides the posterior distribution, with inferences again provided by the Gibbs sampler.

### *4.2. Eliciting a range of prior distributions*

In order to derive prior distributions, we independently consulted with three experts familiar with tuberculosis transmission and pathways of infectious diseases. We directly elicited prior estimates of the proportions of both truly positive and truly negative subjects who would test positive or negative on each of our two dichotomous tests and asked them for means and ranges for the NGD scores within each of the positive and negative subgroups. For example, to derive the sensitivity of the dichotomic test, we asked 'Out of all patients infected with TB in Montreal who truly are newly infected cases, what proportion do you think would have a positive diagnostic recorded from the test?' These estimates were presented in the form of means and ranges, which we converted to beta distributions by matching the beta prior mean with the elicited means and the beta prior standard deviation with one quarter the elicited range, see Joseph et al. [5] or Ladouceur et al. [27] for details. For the continuous test, we asked 'On average, what would be the mean and 95% of the range for the NGD score for people who are recently transmitted', with a similar question for non-recently transmitted cases. We then converted this information to normal distributions by matching means and ranges. Throughout, we used a uniform prior distribution for the prevalence of recent transmission, our parameter of main interest. These distributions are given in Table IV, providing a range of prior distributions across which robustness to the choice of prior can be evaluated.

### *4.3. A range of models for the continuous test alone and all three tests*

We investigated 12 models, including the bi-normal model of Scott et al. [11] and 11 variations of our nonparametric model across a range of prior distributions. This included three nonparametric models using the continuous NGD data alone, the other nine models including data from all three tests. The baseline distributions $G_0^1$ and $G_0^2$ where chosen to be normal, and the hyper prior for the means and

**Table IV.** 95% ranges and parameter values for the prior distributions across the three tests used in the tuberculosis analysis. Coefficients of the Beta distributions correspond to approximate equal tailed 95% credible interval ranges. $G_0^1$ ($D+$ group) and $G_0^2$ ($D-$ group) correspond to baseline densities for the Dirichlet process priors used for the nearest genetic distance (NGD) measure. The prior distribution for the prevalence of clustering was $\beta(1, 1)$ or uniform.

| MIRU | | Spoligotyping | | NGD | |
|---|---|---|---|---|---|
| Sensitivity | Specificity | Sensitivity | Specificity | $G_0^1$ | $G_0^2$ |
| | | Expert 1 | | | |
| 85% to 99% | 40% to 85% | 80% to 97% | 35% to 70% | 10 to 50 | 20 to 500 |
| $\beta(54.36, 4.73)$ | $\beta(10.95, 6.57)$ | $\beta(48.98, 6.36)$ | $\beta(9.82, 8.80)$ | $N(30, 20)$ | $N(150, 70)$ |
| | | Expert 2 | | | |
| 70% to 90% | 70% to 90% | 95% to 100% | 40% to 60% | 0 to 60 | 31 to 500 |
| $\beta(50.40, 12.60)$ | $\beta(50.40, 12.60)$ | $\beta(151.13, 3.88)$ | $\beta(49.5, 49.5)$ | $N(15, 15)$ | $N(132, 60)$ |
| | | Expert 3 | | | |
| 60% to 85% | 60% to 85% | 90% to 99% | 50% to 70% | 0 to 80 | 30 to 350 |
| $\beta(35.25, 11.75)$ | $\beta(35.25, 11.75)$ | $\beta(71.25, 3.75)$ | $\beta(57.00, 38.00)$ | $N(40, 15)$ | $N(120, 60)$ |

standard deviations were created by averaging across the information provided by all experts, leaving $g_0^1$ to be $N(28, 17)$ and $g_0^2$ to be $N(134, 63)$. For the precision parameter, we took values in the set $\alpha \in \{10, 100, 1000\}$. The next two models were obtained by combining all three tests together, with the same nonparametric baseline distributions as above, and using uniform prior distributions on [0,1] for the sensitivities and specificities of the two dichotomous tests. We limited $\alpha$ to be 10 or 100 in order to give more flexibility to the model, as strong prior information is not necessarily needed when data from other tests are available. Finally, the last six models were constructed using the prior information on the three tests provided by each expert separately, again with $\alpha \in \{10, 100\}$. This is of interest because there were some disparities in the opinion of the experts for some parameters, as displayed in Table IV.

### *4.4. Results*

When using continuous NGD data alone, $\alpha$ values of 10, 100, and 1000 gave prevalence estimates (posterior median (95% CrI)) of 23.8 (17.4, 30.9), 18.2 (10.6, 24.4), and 18.5 (13.5, 23.9), respectively. Although there is some variation, all provided a prevalence of approximately 20%, with credible interval widths ranging from 10% to 15%. Note that the estimate closest to 50% is given by the model using the smallest value of $\alpha$, as predicted by our simulation results and in accord with the bias discussed by Petrone and Raftery [24]. The value of $\alpha$ to use in practice depends on the strength of belief of the experts in their prior means, ranges, and density shape (closeness to normality) which can be difficult to elicit. Therefore, one might prefer to base the choice of $\alpha$ on practical knowledge about the tradeoff between flexibility (from lower values of $\alpha$) and concerns about bias (decreased when using higher values of $\alpha$). For example, if a moderately high value such as $\alpha = 100$ leaves sufficient flexibility, then it can be chosen on that basis.

Table V presents the results from the combination of the continuous NGD data and the two dichotomous tests, MIRU and spoligotyping. The prevalence is once again estimated to be approximately 20%, with generally narrower credible intervals compared with the case when the NGD data are used alone. Note that the lengths of the credible intervals for the prevalence are approximately 11%, which is not too far from the theoretically smallest possible length of 8% had the true transmission status been exactly known for all subjects, and the true prevalence was 20%. The sensitivities of MIRU and spoligotyping are both high, at over 95% for many of the models run for spoligotyping but closer to 85% for MIRU. Specificities are much lower, with values generally less than 50% found for spoligotyping and somewhat higher values close to 65% for MIRU. The results remained relatively robust across our range of prior distributions.

The prevalences estimated by parametric and nonparametric approaches are similar, but the advantages of using a nonparametric model can be seen when we look at the estimated shapes of the posterior distributions of each group, as seen in Figure 4. The means (standard deviations) for the clustered and non-clustered groups are, respectively, 38.4 (21.3) and 118.6 (48.6) for the parametric model. Using a noninformative nonparametric Bayesian model, these estimates are 24.1 (22.3) and 119.4 (52.6), respectively, which are not that different, but the shapes of the two densities are markedly different, with the

**Table V.** Posterior median and 95% CrI for the prevalence and test properties of the dichotomous diagnostic tests when using a model that combines information across all three tests.

| | | MIRU | | Spoligotyping | |
|---|---|---|---|---|---|
| | Prevalence | Sensitivity | Specificity | Sensitivity | Specificity |
| Results according to prior distributions from expert #1 | | | | | |
| $\alpha = 100$ | 22.4 (17.1, 28.0) | 86.4 (78.5, 93.2) | 64.7 (59.0, 70.1) | 93.7 (88.3, 97.3) | 44.2 (38.5, 49.8) |
| $\alpha = 10$ | 24.7 (19.2, 30.6) | 82.3 (74.8, 89.0) | 64.1 (58.2, 69.4) | 93.2 (87.7, 96.9) | 44.9 (39.1, 51.0) |
| Results according to prior distributions from expert #2 | | | | | |
| $\alpha = 100$ | 17.2 (12.7, 21.9) | 80.0 (72.0, 87.1) | 64.2 (59.4, 69.1) | 98.1 (95.6, 99.4) | 43.5 (38.8, 48.5) |
| $\alpha = 10$ | 20.0 (15.3, 25.0) | 78.2 (70.2, 84.8) | 64.9 (59.7, 69.6) | 97.7 (94.8, 99.3) | 44.4 (39.6, 49.2) |
| Results according to prior distributions from expert #3 | | | | | |
| $\alpha = 100$ | 23.9 (18.5, 29.6) | 77.3 (68.9, 84.9) | 65.8 (60.5, 71.0) | 96.7 (92.6, 99.0) | 48.3 (43.2, 53.5) |
| $\alpha = 10$ | 26.5 (21.0, 32.0) | 74.5 (66.6, 82.0) | 65.8 (60.6, 70.9) | 96.0 (91.8, 98.6) | 49.2 (44.0, 54.6) |
| Results according to prior distributions averaged across experts | | | | | |
| $\alpha = 100$ | 21.0 (16.0 , 26.4) | 78.0 (66.2 , 89.2) | 62.9 (57.0 , 68.3) | 97.5 (90.8 , 99.9) | 43.3 (37.8 , 49.1) |
| $\alpha = 10$ | 23.6 (18.2 , 29.5) | 74.0 (63.5 , 83.3) | 62.9 (57.3 , 68.5) | 95.7 (88.4 , 99.5) | 44.1 (38.4 , 50.0) |

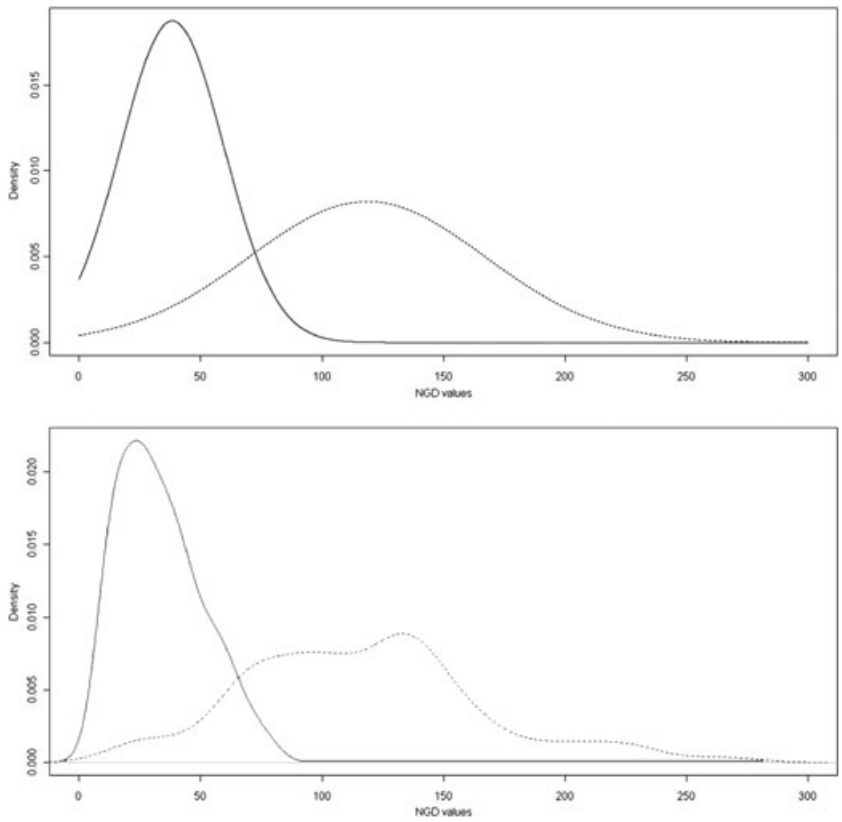


**Figure 4.** Top graph illustrates the posterior distributions from the clustered and non-clustered groups from the parametric model [11], whereas the bottom graph displays these same posterior distributions obtained from a nonparametric model that combines the three tests by using non-informative priors across all parameters and the average of the expert opinions across three experts for the baseline distributions.

parametric models missing much of the non-normality made obvious by the nonparametric models. This would be important for clinicians to know, for example, because calculating the probability of recent transmission for a subject given their particular NGD value depends on the relative heights of the $D+$ and $D-$ curves at that value [11].

## 5. Discussion

We developed and investigated the properties of a simple Bayesian nonparametric model based on an approximate Dirichlet process, designed to estimate the prevalence of a disease and test properties in the absence of a gold standard test. We first presented a method where a single continuous test is used, our goal being to replace the parametric (almost always normal) assumption about the population distributions with a more flexible family. Our choice of simulations to investigate was motivated by prototypic scenarios in diagnostic testing problems. Obviously, there are many other scenarios which could have been studied.

We used relatively large sample sizes ($n = 1000$ and $n = 5000$) because for smaller sizes the inferences from both parametric and nonparametric models were not precise, even under ideal conditions and using perfect prior choices. Thus, we conclude that for single continuous tests, small sample sizes generally require very strong prior information, and even if such prior information is available, inferences will not likely be precise. As discussed in Section 4, additional data from other tests can greatly sharpen inferences.

As Petrone and Raftery [24] pointed out, when using very low values of the precision parameter $\alpha$, the inferences obtained about the prevalence can potentially be biased. In practice, this bias corresponds to a prevalence estimate that gravitate towards 50%. Our simulations have shown that this potential bias should not discourage researchers from using nonparametric models but that some precautions are

needed. Intuitively, it makes sense that if little information about any of the test result distributions is available, any estimate of the prevalence will tend towards 50%, as that model cannot distinguish one group from another, by construction. We have observed similar phenomena for Polya tree models, showing that the bias extends to nonparametric models other than those based on Dirichlet processes.

Our simulations only investigated the situation where $\alpha_1 = \alpha_2$, but sometimes more information about test results is available for one group compared with the other. While we did not investigate that aspect in order to keep the number of simulations to a reasonable level, using stronger prior information on one population compared with another might be a good compromise in creating a flexible model that remains identifiable. Similarly, for simplicity we assumed $G_0$ and $G_1$ to be *a priori* independent, with their order being determined by prior information on the normal means. However, another prior choice could be to model these jointly, building in structures such as $G_0 \leq G_1$.

## Acknowledgements

We would like to thank Dr. Marcel Behr for providing the NGD data.

## References

1. Kulaga S, Behr M, Nguyen D, Brinkman J, Westley J, Menzies D, Brassard P, Tannenbaum T, Thibert L, Boivin JF, Joseph L, Schwartzman K. Diversity of *Mycobacterium tuberculosis* isolates in an immigrant population: evidence against a founder effect. *American Journal of Epidemiology* 2004; **159**:507–513.
2. Mazars E, Lesjean S, Banuls AL, Gilbert M, Vincent V, Gicquel B, Tibayrenc M, Locht C, Supply P. High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. *Proceedings of the National Academy of Sciences* 2001; **98**:1901–1906.
3. Hayward AC, Watson JM. Typing of mycobacteria using spoligotyping. *Thorax* 1998; **53**:329–330.
4. Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology* 1988; **41**:923–937.
5. Joseph L, Gyorkos TW, 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.
6. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine* 1999; **18**:2987–3003.
7. Johnson WO, Gastwirth JL, Pearson LM. Screening without a "gold standard": the Hui–Walter paradigm revisited. *American Journal of Epidemiology* 2001; **153**:921–924.
8. Black MA, Craig BA. Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine* 2002; **21**:2653–2669.
9. McInturff P, Johnson WO, Cowling D, Gardner IA. Modelling risk when binary outcomes are subject to error. *Statistics in Medicine* 2004; **23**:1095–1109.
10. Gustafson P. On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statistical Science* 2005; **20**:111–140.
11. Scott AN, Joseph L, Bélisle P, Behr MA, Schwartzman K. Bayesian modelling of tuberculosis clustering from DNA fingerprint data. *Statistics in Medicine* 2008; **27**:140–156.
12. Branscum AJ, Johnson WO, Hanson TE, Gardner IA. Bayesian semiparametric ROC curve estimation and disease diagnosis. *Statistics in Medicine* 2008; **27**:2474–2496.
13. Branscum AJ, Hanson TE. Bayesian nonparametric meta-analysis using Polya tree mixture models. *Biometrics* 2008; **64**:825–833.
14. Branscum AJ, Hanson TE, Gardner IA. Bayesian non-parametric models for regional prevalence estimation. *Journal of Applied Statistics* 2008; **35**:567–582.
15. Erkanli A, Sung M, Costello EJ, Angold A. Bayesian semi-parametric ROC analysis. *Statistics in Medicine* 2006; **25**:3905–3928.
16. Wang C, Turnbull BW, Grohn YT, Nielsen S. Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown. *Journal of Agricultural Biological and Environmental Statistics* 2007; **12**:128–146.
17. Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. *Statistics in Medicine* 2009; **28**:441–461.
18. Hanson TE, Branscum AJ, Gardner IA. Multivariate mixtures of Polya trees for modelling ROC data. *Statistical Modelling* 2008; **8**:81–96.
19. Zhou X, Castelluccio P, Zhou C. Nonparametric estimation of ROC curves in the absence of a gold standard. *Biometrics* 2005; **61**:600–609.
20. Albert PS. Imputation approaches for estimating diagnostic accuracy for multiple tests from partially verified designs. *Biometrics* 2007; **63**:947–957.
21. Albert P, Dodd L. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 2004; **60**:427–435.
22. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. *Biostatistics* 2007; **8**:474–484.
23. Ferguson TS. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1973; **1**:209–230.

24. Petrone S, Raftery AE. A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *Statistics & Probability Letters* 1997; **36**:69–83.
25. Mira A, Petrone S. Bayesian hierarchical nonparametric inference for change-point problems. *Bayesian Statistics* 1996; **5**:693–703.
26. Ladouceur M. Modeling continuous diagnostic test results using Dirichlet process prior distributions. *PhD thesis*, McGill University, Montreal, Canada, 2009.
27. Ladouceur M, Rahme E, Pineau CA, Joseph L. Robustness of prevalence estimates derived from misclassified data from administrative databases. *Biometrics* 2007; **63**:272–279.
28. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 2001; **57**:158–167.
29. 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.