STATISTICS IN MEDICINE
Statist. Med. (in press)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2899

Bayesian modelling of tuberculosis clustering from DNA fingerprint data

Allison N. Scott[1], Lawrence Joseph[1,2,*,†], Patrick Bélisle[2], Marcel A. Behr[3] and Kevin Schwartzman[1,4]

[1] *Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montreal, Que., Canada H3A 1A2*
[2] *Division of Clinical Epidemiology, Department of Medicine, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Que., Canada H3G 1A4*
[3] *Division of Infectious Diseases and Microbiology, A5.156, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Que., Canada H3G 1A4*
[4] *Respiratory Epidemiology Unit, Montreal Chest Institute, 3650 Saint Urbain Street, Montreal, Que., Canada H2X 2P4*

SUMMARY

A combination of continuous and categorical tests, none of which is a gold standard, is often available for classification of subject status in epidemiologic studies. For example, tuberculosis (TB) molecular epidemiology uses select mycobacterial DNA sequences to provide clues about which cases of active TB are likely clustered, implying recent transmission between these cases, *versus* reactivation of previously acquired infection. The proportion of recently transmitted cases is important to public health, as different control methods are implemented as transmission rates increase. Standard typing methods include IS*6110* restriction fragment length polymorphism (IS*6110* RFLP), but recently developed polymerase chain reaction based genotyping modalities, including mycobacterial interspersed repetitive unit-variable-number tandem repeat and spoligotyping provide quicker results. In addition, it has recently been suggested that results from IS*6110* RFLP can be used to create a continuous measure of genetic relatedness, called the nearest genetic distance. Whichever method is used, estimation of cluster rates is rendered difficult by the lack of a gold standard method for classifying cases as clustered or not. Since many of these methods are relatively new, their properties have not been extensively investigated. Misclassification errors subsequently lead to sub-optimal estimation of risk factors for clustering. Here we show how Bayesian latent class models can be used in such situations, for example to simultaneously analyse *Mycobacterium tuberculosis* DNA data from all three of the above methods. Using the data collected at the Public Health Unit in Montreal, we estimate the proportion of clustered cases and the operating characteristics of each method using information from all three methods combined, including both continuous and dichotomous measures from IS*6110* RFLP. A misclassification-adjusted regression model provides estimates of the

[*] Correspondence to: Lawrence Joseph, Division of Clinical Epidemiology, Department of Medicine, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Que., Canada H3G 1A4.
[†] E-mail: Lawrence.Joseph@mcgill.ca

Received 19 June 2006
Accepted 9 March 2007

effects of risk factors on the clustering probabilities. We also discuss how one must carefully interpret any inferences that arise from a combination of continuous and dichotomous tests. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: Bayesian analysis; diagnostic test; molecular epidemiology; ROC curve; sensitivity and specificity; tuberculosis

1. INTRODUCTION

The World Health Organization estimates that one-third of the world's population is infected with *Mycobacterium tuberculosis*, the causative agent of tuberculosis (TB), and that another individual is infected every second [1]. The prevalences in most developed countries are lower, and in Montreal, Canada, the setting of our study, approximately 150–200 people are diagnosed with active disease each year [2]. Nevertheless, even in the areas of low prevalence, it is important for public health officials to track the transmission pathways of infectious diseases in order to promote optimal control measures. In particular, it is useful to know the proportion of cases that are reactivations of previously acquired infection, *versus* recent disease transmission. Although no gold standard method for this exists, it is usually accomplished by comparing *M. tuberculosis* strains from new cases to strains from those previously diagnosed with active TB in the community. It is also important to elicit the demographic factors associated with recent transmission, usually accomplished by a regression model using the possibly misclassified test results as the outcome.

TB molecular epidemiology exploits selected bacterial DNA targets to serve as markers for *M. tuberculosis* strains [3]. The most common method of DNA fingerprinting used is IS*6110*-based restriction fragment-length polymorphism (IS*6110*) RFLP [4]. An important practical limitation of IS*6110* RFLP is that results are usually obtained weeks to months after the diagnosis of TB, because of the need to grow large numbers of bacteria to extract sufficient DNA. Therefore, while useful for documenting transmission events, IS*6110* RFLP often provides data once outbreaks are well established. In contrast, a number of polymerase chain reaction (PCR) based typing modalities have been recently developed, including spoligotyping [5, 6] and mycobacterial interspersed repetitive unit-variable-number tandem repeat (MIRU-VNTR, or here, simply abbreviated to MIRU), which offer the possibility of obtaining DNA fingerprints from small numbers of bacteria [7]. The major advantage of rapid typing of isolates is the capacity to immediately provide a strain designation, which in turn may facilitate prompt public health intervention. While the technical feasibility to generate genotypes with these methods has been well demonstrated, their public health utility depends on the method's properties.

As the PCR-based methods are relatively new, their sensitivities and specificities are not well documented to date [8, 9]. IS*6110* RFLP is far from a perfect indicator of recent transmission, so that exactly how to estimate the properties of new methods is not clear. Further, IS*6110* RFLP can either be used as a dichotomous method, or, as has recently been proposed, as a continuous data method *via* a measure called the nearest genetic distance (NGD) [10]. Therefore, one has the choice to compare the new results with the results from a continuous or dichotomous method. Whichever is selected, neither is a gold standard, and any model must take this into account.

In this paper, the above example is used to illustrate how Bayesian latent class models can be employed to estimate the properties of combinations of dichotomous and continuous tests in the absence of a gold standard. While we present two main models, one to analyse three dichotomous

DOI: 10.1002/sim

tests, and another including results from two dichotomous and one continuous test, the methods are easily extendable to other combinations. We use these models to estimate the properties of our DNA methods and the proportion of TB cases that are clustered. We compare results from the two different models, and discuss the various assumptions behind each. Also considering results from a model which (incorrectly) assumes IS*6110* RFLP to be a gold standard, we show that accounting for the imperfections in each method can lead to very different estimates. We also present, for the first time, estimates of the risk factors for recent transmission that are adjusted for misclassification errors.

Bayesian latent class models for dichotomous diagnostic tests arising from laboratory tests in the absence of a gold standard test have been discussed by many in the past [11–15]. Similar models for continuous data and ROC curves have been investigated [16], although they assumed availability of a gold standard test result for each patient. Choi *et al.* [17] estimated disease prevalence and predictive probabilities for individual level continuous test results in the absence of a gold standard, but only considered results from a single test, and assumed the availability of a training sample of test results to estimate the distributions of truly disease positive and truly disease negative subjects. Erkanli *et al.* [18] proposed a non-parametric analysis using mixtures of Dirichlet processes to model distributions within diseased and non-diseased subjects, leading to non-parametric ROC curve estimation, but again assumed a gold standard test. Since it is often the case in practice that results are available on more than one test, none of which can be considered as a gold standard, we extend these models to allow consideration of one or more dichotomous test results in addition to a continuous test, in the absence of a gold standard. Another unique feature of our paper is that, to our knowledge, this is the first time such a latent class model has been applied to diagnostic tests relying on DNA sequence data.

Section 2 describes the study setting and the three methods which give rise to our data. Section 3 presents our models, and discusses how to estimate not only the proportion of clustered cases, but also the properties of each genetic method, as well as estimates of the effects of various demographic factors on the rates of recent transmission. The results of applying our models to our TB DNA data are given in Section 4, and we conclude with a discussion.

2. BACKGROUND

2.1. Study setting

The study setting and data collection have been described previously [8]. Briefly, the study population consisted of residents of Montreal diagnosed with active TB over a three-year period, from 1 January 1996 to 31 December 1998. All diagnosed cases of TB are reported to the local public health authorities, who enter clinical and demographic information into a notifiable disease registry. Clinical isolates are archived at the provincial public health laboratory. Over 85 per cent of all diagnosed cases have a positive culture.

One isolate per patient was selected for this study and IS*6110* RFLP, MIRU-VNTR and spoligotyping were performed according to standard methods [4, 5, 7, 19]. IS*6110* RFLP fingerprints were clustered using Molecular Fingerprint Analyzer 2.0 (Stanford University).

2.2. Molecular methods

Standard IS 6110 RFLP: In molecular epidemiology studies, genotyping is used to characterize the different *M. tuberculosis* isolates. Molecular genotyping identifies DNA sequences within the

DOI: 10.1002/sim

M. tuberculosis genome to determine the different *M. tuberculosis* isolates within a community. If two clinical isolates from two different patients have identical (or, sometimes, very similar) genotypes, it is considered likely that transmission of TB took place between them.

The standard genotyping method, IS*6110* RFLP typing [4] is based on a small DNA sequence that is capable of creating copies of itself within the genome [20, 21]. IS*6110* RFLP typing cleaves the *M. tuberculosis* genome and identifies the number of copies of IS*6110* within the genome and the size of the genomic DNA fragments they are on, creating a pattern that looks similar to a bar code. Isolates with the same number of fragments of matching molecular weights are considered clustered, and transmission of TB between those individuals is suspected. However, IS*6110* RFLP typing has significant drawbacks: it is time consuming, labour intensive, is suspected to have poor specificity in isolates with five or less copies of IS*6110*, and can only produce results in three to six weeks [22, 23]. Furthermore, it has been noted that the IS*6110* fingerprint pattern can evolve in an individual or within an epidemiologic cluster over time, resulting in band additions, deletions or shifts [24, 25]. Therefore, this method will have far from perfect sensitivity and specificity for detecting recent transmission.

Nearest genetic distance (NGD): Recently, Salamon *et al.* [10] created a measure of genetic distance based on IS*6110* RFLP. Assuming an exponential rate of transposition that depends on the number of copies of IS*6110* within the genome, they were able to create a measure of time since two isolates diverged from a common ancestor. Isolates are considered clustered if the genetic distance between them and the isolate they are genetically most related to (their nearest genetic distance (NGD)) is below a certain cut-off value. This continuous measure allows for fingerprint evolution, and depending on the cut-off chosen, can allow IS*6110* RFLP to be used in isolates with less than six copies of IS*6110*.

Spoligotyping and MIRU-VNTR: Two new PCR-based genotyping methods have recently been developed. Spoligotyping measures the presence or absence of 40 unique sequences at a particular site in the *M. tuberculosis* genome [6]. MIRU-VNTR measures the number of copies of repeated sequences at 12 different locations in the genome [7]. Isolates are considered clustered when their genotypes are identical. These methods are less expensive, less labour intensive, and faster to perform compared to IS*6110* RFLP typing, so that genotyping results can be obtained within a week. This rapidity gives public health authorities the opportunity to include the results of these methods as new cases are diagnosed. Many public health organizations have switched from IS*6110* RFLP to spoligotyping and MIRU-VNTR for these reasons [26], even though their properties are not yet well known.

For a review of these genotyping methods and their relevance to the understanding of TB transmission, see Barnes and Cave [3].

3. STATISTICAL METHODS

We have data from two dichotomous methods (MIRU and spoligotyping), and IS*6110* RFLP data, which can be used in the conventional fashion as a dichotomous method (clustered or not), or converted into a continuous measure using the NGD. We will perform separate analyses of the data from these three methods, one for each form of IS*6110* RFLP data, and also create a model that allows the probability of clustering to depend on various covariates, including age, sex, country of origin, and pulmonary disease status. All of our main models assume that none of our methods provide perfectly accurate results, but we also compare our results to a very simple model that assumes IS*6110* RFLP dichotomous data to be the gold standard, as has often been the case in the past literature. Below we divide our methods into three subsections, first describing how to estimate all unknown parameters when the data consist of three imperfect dichotomous tests. We next present a model for the situation of two dichotomous tests with one continuous test, and finally, the latter model is extended to include covariates that can modify the probability of clustering.

3.1. Three dichotomous tests

Bayesian estimation of the prevalence of a condition given results from one, two or three dichotomous diagnostic tests has been previously discussed [12, 13], and we will first apply these previously developed methods to detect clustering using data from our three DNA sequence tests. Let $C+$ and $C-$ denote the true clustering status, 'clustered' and 'not clustered', respectively. Similarly, let $T+$ and $T-$ represent test positive and test negative outcomes on a given dichotomous test, i.e. IS*6110* RFLP, MIRU, or spoligotyping. In the absence of a perfectly accurate method, estimating the clustering rate θ will depend on the test characteristics, particularly the sensitivity $\mathrm{Se} = P(T+|C+)$ and specificity $\mathrm{Sp} = P(T-|C-)$, where $P(A|B)$ denotes the conditional probability of A given B. For any single test, the probability of testing positive is the sum of the probabilities of being a true positive and the probability of being a false positive, i.e. $P(T+) = p = \theta\mathrm{Se} + (1-\theta)(1-\mathrm{Sp})$. In the single test situation when Se and Sp are exactly known, algebraically solving for θ, a prevalence estimate adjusted for sensitivity and specificity is given by $\theta = p - (1-\mathrm{Sp})/(\mathrm{Se}+\mathrm{Sp}-1)$. Note that in the case that $\mathrm{Se} = \mathrm{Sp} = 1$ (i.e. a gold standard test) then $\theta = p$, and no adjustment is needed.

Test properties Se and Sp from each test are not usually exactly known, and so have to be estimated along with θ. This presents a non-identifiable estimation problem if data from only one or two tests are available [12, 15, 27]. Formally, a model expressed *via* the density function $f(y|\theta)$ is identifiable if $f(x|\theta_1) = f(x|\theta_2)$ for all x implies $\theta_1 = \theta_2$. In non-identifiable problems, it is clear that the data alone cannot provide consistent estimation of θ. Estimates can nevertheless be derived *via* Bayesian methods, where prior information can separate out the likelihood of θ_1 from θ_2 when the data cannot distinguish between these values. When using three conditionally independent tests, however, the problem becomes identifiable, as there are seven degrees of freedom (from the eight possible outcomes, see Table I) from which to estimate the seven unknown parameters (Se and Sp from each of three tests, and θ). In practice, this means that estimates of all parameters can be found by maximum likelihood or Bayesian methods with non-informative prior distributions, avoiding the reliance on finding good prior estimates of test properties. Unlike the one test situation, however, this method relies on the unverifiable assumption of conditional independence between the three tests. Tests are conditionally independent if their results are independent of each other, conditional on the true status of the individual being tested. In our data set, however, three rather different genetic methodologies are being used, so that conditional independence seems reasonable, as the target DNA sequences used for each test of clustering are for the most part unrelated to one another (but see Section 5 for further discussion).

To review how estimates from the three test model are constructed, consider Table I. When three test results are available for each subject, each result could be either positive or negative, leading to eight possible combinations for the observed data. It is convenient to also consider the (latent) true status of each individual, leading to 16 possible combinations of observed and latent data. Let $Y_1, \ldots, Y_8$ be latent data that represents the number of true positive subjects out of $a, b, \ldots, h$, subjects in each possible category for the observed test results, respectively. Given

Table I. Likelihood contributions of all possible combinations of observed and latent data for the case of three diagnostic tests.

Truth	Test 1 result	Test 2 result	Test 3 result	Likelihood contribution per subject	Number of subjects
+	+	+	+	$\theta\text{Se}_1\text{Se}_2\text{Se}_3$	Y_1
+	+	+	−	$\theta\text{Se}_1\text{Se}_2(1-\text{Se}_3)$	Y_2
+	+	−	+	$\theta\text{Se}_1(1-\text{Se}_2)\text{Se}_3$	Y_3
+	+	−	−	$\theta\text{Se}_1(1-\text{Se}_2)(1-\text{Se}_3)$	Y_4
+	−	+	+	$\theta(1-\text{Se}_1)\text{Se}_2\text{Se}_3$	Y_5
+	−	+	−	$\theta(1-\text{Se}_1)\text{Se}_2(1-\text{Se}_3)$	Y_6
+	−	−	+	$\theta(1-\text{Se}_1)(1-\text{Se}_2)\text{Se}_3$	Y_7
+	−	−	−	$\theta(1-\text{Se}_1)(1-\text{Se}_2)(1-\text{Se}_3)$	Y_8
−	+	+	+	$(1-\theta)(1-\text{Sp}_1)(1-\text{Sp}_2)(1-\text{Sp}_3)$	$a-Y_1$
−	+	+	−	$(1-\theta)(1-\text{Sp}_1)(1-\text{Sp}_2)\text{Sp}_3$	$b-Y_2$
−	+	−	+	$(1-\theta)(1-\text{Sp}_1)\text{Sp}_2(1-\text{Sp}_3)$	$c-Y_3$
−	+	−	−	$(1-\theta)(1-\text{Sp}_1)\text{Sp}_2\text{Sp}_3$	$d-Y_4$
−	−	+	+	$(1-\theta)\text{Sp}_1(1-\text{Sp}_2)(1-\text{Sp}_3)$	$e-Y_5$
−	−	+	−	$(1-\theta)\text{Sp}_1(1-\text{Sp}_2)\text{Sp}_3$	$f-Y_6$
−	−	−	+	$(1-\theta)\text{Sp}_1\text{Sp}_2(1-\text{Sp}_3)$	$g-Y_7$
−	−	−	−	$(1-\theta)\text{Sp}_1\text{Sp}_2\text{Sp}_3$	$h-Y_8$

Note: The parameter θ represents the prevalence, and Se_i and Sp_i represent the sensitivity and specificity, respectively, of the ith test. The vector of observed numbers of subjects with each possible combination of test results is given by $(a, b, \ldots, h)$, and within each of these cells, we have the unobserved number of truly positive subjects, represented by the vector of latent data, $(Y_1, Y_2, \ldots, Y_8)$.

results from three dichotomous tests, one can calculate the probability of being in a cluster, given by the positive predictive values. For example, if a subject is positive on all three tests, then

$$P(\text{subject is a true positive}) = \frac{\theta\text{Se}_1\text{Se}_2\text{Se}_3}{\theta\text{Se}_1\text{Se}_2\text{Se}_3 + (1-\theta)(1-\text{Sp}_1)(1-\text{Sp}_2)(1-\text{Sp}_3)}$$

The numerator of this expression results from the probability of being positive, i.e. the prevalence θ, multiplied by the conditional probability of testing positively on each test, i.e. the three sensitivities. Similar reasoning leads to the second expression in the denominator, and to the rest of the quantities derived in Table I. The full likelihood function of the observed and latent data is proportional to the product of each entry in the likelihood contribution column of Table 1 raised to the power of the corresponding entry in the number of subjects column of the table.

Estimates are derived either by maximizing the likelihood function [27], or by Bayesian methods *via* the addition of a prior distribution over the prevalence θ and all test parameters, Se_i and Sp_i, $i = 1, 2, 3$. We used beta prior distributions, and if uniform or other low information priors are used, Bayesian methods will provide similar numerical estimates to maximum likelihood, but have the added advantage of increased precision if reliable prior information is available, especially in small data sets [12]. In either case, analytic solutions are not feasible, so the EM algorithm or the Gibbs sampler are typically used in practice to maximize the likelihood or approximate Bayesian posterior distributions, respectively. Once the above seven parameters are estimated, marginal posterior distributions can easily be estimated for any function of these parameters using

DOI: 10.1002/sim

the Gibbs sampler output [28]. Therefore, posterior distributions for the likelihood ratios and the probabilities of being truly positive given any combination of test results are available, as are the positive and negative predictive values for each test.

For comparison purposes, we will also present results of the prevalence of recent TB transmission and the sensitivities and specificities of MIRU and spoligotyping when considering IS*6110* RFLP to be a perfect reference standard. We will thus expose the potential bias in the evaluation of these recently developed tests if misclassification is ignored.

3.2. Two dichotomous tests and one continuous test

Alternatively, we can treat the IS*6110* RFLP data as continuous, using the NGD measure. We first consider the situation where data from a single continuous test are available, and later show how the likelihood function changes when there are also data available from one or more dichotomous tests. We assume that the NGD values follow normal distributions, within each of the clustered and non-clustered sub-populations, so that $\mathrm{N}(\mu_{\mathrm{C}}, \sigma^2_{\mathrm{C}})$ is the density function for the results of the NGD values conditional on being in a cluster, and $\mathrm{N}(\mu_{\mathrm{NC}}, \sigma^2_{\mathrm{NC}})$ is the density function conditional on not being in a cluster. Similar to the methods of Section 3.1, we assume that there are latent data z_i, $i = 1, 2, \ldots, n$, representing the (unknown) true status of each subject, with $z_i = 1$ if the subject is clustered, and $z_i = 0$ otherwise. The likelihood function for the observed and latent data is

$$f(\underset{\sim}{X}, \underset{\sim}{Z} | \theta, \mu_{\mathrm{C}}, \sigma^2_{\mathrm{C}}, \mu_{\mathrm{NC}}, \sigma^2_{\mathrm{NC}}) = \prod_{i=1}^{n} \left(\theta \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{C}}} \exp\left\{ -\frac{1}{2}(x_i - \mu_{\mathrm{C}})^2/\sigma^2_{\mathrm{C}} \right\} \right)^{z_i} \times \left((1-\theta) \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{NC}}} \exp\left\{ -\frac{1}{2}(x_i - \mu_{\mathrm{NC}})^2/\sigma^2_{\mathrm{NC}} \right\} \right)^{1-z_i} \quad (1)$$

where $\underset{\sim}{X} = (x_1, x_2, \ldots, x_n)$ are the observed NGD data, and $\underset{\sim}{Z} = (z_1, z_2, \ldots, z_n)$ are the latent data. Thus, the likelihood contribution from each subject i is a normal distribution, with the contribution going to either the first or second normal distribution in the above expression, depending on whether the subject is considered to be clustered or not, respectively.

The problem reduces to one of fitting a normal mixture model, which can be solved *via* maximum likelihood or Bayesian methods [29]. We use the latter, and thus require prior distributions over the parameter space $(\theta, \mu_{\mathrm{C}}, \sigma^2_{\mathrm{C}}, \mu_{\mathrm{NC}}, \sigma^2_{\mathrm{NC}})$. We will use a beta prior distribution for θ, normal distributions for μ_{C} and μ_{NC}, and uniform distributions over a finite range for σ_{C} and σ_{NC}.

Of course, once the parameters of the normal curves are estimated, one can estimate any function of these parameters, including ROC curves [30]. In addition, it is of clinical interest to know the probability of being in a cluster given any NGD value, x. Once $(\theta, \mu_{\mathrm{C}}, \sigma^2_{\mathrm{C}}, \mu_{\mathrm{NC}}, \sigma^2_{\mathrm{NC}})$ have been estimated, the posterior probabilities $P(\text{clustered}|x, \mu_{\mathrm{C}}, \sigma^2_{\mathrm{C}}, \mu_{\mathrm{NC}}, \sigma^2_{\mathrm{NC}})$ are obtainable *via* the formula

$$P(\text{clustered}|x, \mu_{\mathrm{C}}, \sigma^2_{\mathrm{C}}, \mu_{\mathrm{NC}}, \sigma^2_{\mathrm{NC}}) = \frac{\theta \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{C}}} \exp\left\{ -\frac{1}{2}(x_i - \mu_{\mathrm{C}})^2/\sigma^2_{\mathrm{C}} \right\}}{\theta \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{C}}} \exp\left\{ -\frac{1}{2}(x_i - \mu_{\mathrm{C}})^2/\sigma^2_{\mathrm{C}} \right\} + (1-\theta) \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{NC}}} \exp\left\{ -\frac{1}{2}(x_i - \mu_{\mathrm{NC}})^2/\sigma^2_{\mathrm{NC}} \right\}}$$

DOI: 10.1002/sim

In addition to the continuous test, we also have our two dichotomous methods, MIRU (T_1) and spoligotyping (T_2). The full likelihood function for data from all three methods, is a combination of the above likelihood (1) with that from two dichotomous methods, a reduction of the formula implied by Table I. Therefore, in this case, the full likelihood function is

$$
\begin{aligned}
& f(\underset{\sim}{X}, \underset{\sim}{Z}, \underset{\sim}{T_1}, \underset{\sim}{T_2} | \theta, \mathrm{Se}_1, \mathrm{Se}_2, \mathrm{Sp}_1, \mathrm{Sp}_2, \mu_{\mathrm{C}}, \sigma^2_{\mathrm{C}}, \mu_{\mathrm{NC}}, \sigma^2_{\mathrm{NC}}) \\
&= \prod_{i=1}^{n} \left(\theta \mathrm{Se}_1^{T_{1i}} (1-\mathrm{Se}_1)^{(1-T_{1i})} \mathrm{Se}_2^{T_{2i}} (1-\mathrm{Se}_2)^{(1-T_{2i})} \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{C}}} \exp\left\{ -\frac{1}{2}(x_i - \mu_{\mathrm{C}})^2/\sigma^2_{\mathrm{C}} \right\} \right)^{z_i} \\
&\quad \times \Bigg((1-\theta)(1-\mathrm{Sp}_1)^{T_{1i}} \mathrm{Sp}_1^{(1-T_{1i})} \mathrm{Sp}_2^{(1-T_{2i})} (1-\mathrm{Sp}_2)^{(1-T_{2i})} \\
&\quad \times \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{NC}}} \exp\left\{ -\frac{1}{2}(x_i - \mu_{\mathrm{NC}})^2/\sigma^2_{\mathrm{NC}} \right\} \Bigg)^{1-z_i}
\end{aligned} \tag{2}
$$

where $\underset{\sim}{T_1} = (T_{11}, T_{12}, \ldots, T_{1n})$ and $\underset{\sim}{T_2} = (T_{21}, T_{22}, \ldots, T_{2n})$ are the results for the two diagnostic tests across all n subjects, respectively.

We use beta prior distributions on θ, Se_1, Se_2, Sp_1, and Sp_2, and normal/uniform priors for the normal parameters, as described above. We again use the Gibbs sampler for inferences, which again means that the posterior distributions from any function of our unknown parameters, such as the probability of being clustered given the results from one continuous and two dichotomous tests are available, as are all likelihood ratios from all combinations of tests.

Non-statisticians would not be usually be sufficiently expert to fit the above models by programming the methods themselves. Therefore, user-friendly Windows-based software programs implementing all of the above models, including generation of ROC curves and graphs of the probabilities of clustering conditional on the posterior distributions of all parameters, are freely available from www.medicine.mcgill.ca/epidemiology/Joseph/.

3.3. Inclusion of covariates

Rather than using fixed prior distributions, employing hierarchical modelling allows investigation of the effects of any covariates on the prevalence θ or any of the test properties, including any of the sensitivities, specificities, or parameters from the normal distributions of the continuous test. While one can directly model the α and β parameters *via* hierarchical distributions, for example, using the gamma distribution, we follow a more common method using the logistic function.

Here, it is of greatest interest to public health officials to investigate whether any demographic factors may influence the rate of recent transmission. We thus focus on a model for the prevalence of clustering, although similar models can easily be constructed for the other parameters. To replace the beta prior distribution for θ, let θ_i represent the probability of clustering for subject i, and model as follows:

$$
\begin{aligned}
\gamma_i &= \mathrm{logit}(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) \\
\gamma_i &\sim \mathrm{N}(\mu_{\gamma_i}, \sigma^2_{\gamma}) \\
\mu_{\gamma_i} &= \alpha + \beta_1 \mathrm{age}_i + \beta_2 \mathrm{sex}_i + \beta_3 \mathrm{origin}_i
\end{aligned}
$$

where, for example, age, sex, and country of origin are assumed to be the covariates of interest. The posterior distributions of $\exp(\beta_i)$, $i=1,2,3$, then estimate the odds ratios for the corresponding covariates. Normal prior distributions can be used for the regression parameters α and β_i, $i=1,2,3$, and a uniform prior distribution can be used for σ_γ.

4. RESULTS

Data were available on 393 subjects with recently diagnosed active TB in Montreal. The average age was 45.1 (SD=21.1), 44.7 per cent of cases were female, 59.2 per cent were smear positive cases of TB, and 81.9 per cent of cases were foreign born. The mean NGD value was 98.9 (SD=55.4), with range from 11.3 to 270.6 (Figure 1). For each genotyping modality, we determined whether isolates had a unique pattern in the data set (referred to as negative in our models) or whether that isolate had a pattern matching that of another isolate (called positive). From this, the rate of subjects positive on MIRU was 45.8 per cent, while 65.4 per cent were positive on spoligotyping, and 19.1 per cent were positive on standard IS6110 RFLP typing (Table II). All three methods agreed on the clustering status of only 117 of the 393 cases, less than 30 per cent agreement. Of note is that spoligotyping found 197 cases positive that were categorized as negative by IS*6110* RFLP, and 104 of these 197 cases were also labelled as positive by MIRU. The large discrepancies in results across methods means that *at least* two of the three methods perform very poorly, if not all three, casting doubt on the clinical utility of these methods. This certainly is true when results of each method are used alone, and while the situation improves if information from all methods is combined, as is the case here, considerable uncertainty remains.

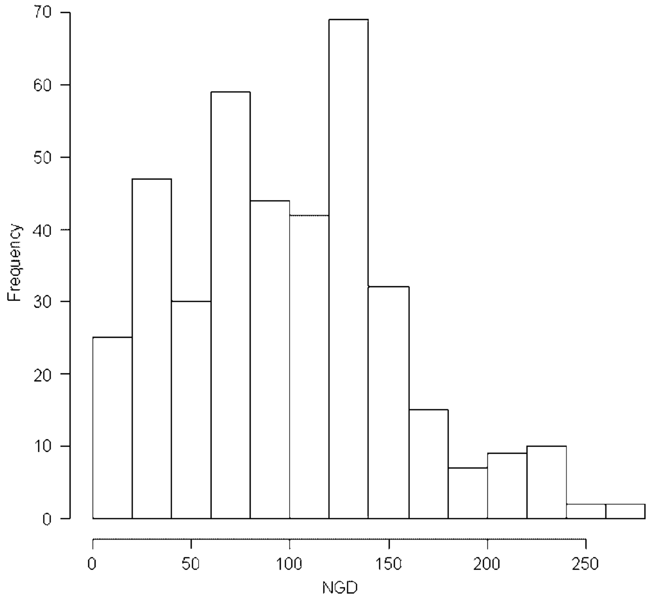

Figure 1. Histogram of the nearest genetic distance (NGD) values from 393 subjects.

DOI: 10.1002/sim

Table II. Dichotomous data on 393 subjects, all tested with standard IS*6110* RFLP, MIRU, and spoligotyping.

IS*6110* RFLP	MIRU	Spoligotyping	Number of cases
+	+	+	35
+	+	−	2
+	−	+	25
+	−	−	13
−	+	+	104
−	+	−	39
−	−	+	93
−	−	−	82

Table III. Posterior medians and 95 per cent credible intervals for the prevalence of clustering, and the sensitivities and specificities of the three tests for clustering from the analysis of the data in Table II.

Variable		Posterior median	95 per cent credible interval
Prevalence of clustering		0.55	(0.26, 0.77)
IS*6110* RFLP	Sensitivity	0.26	(0.19, 0.39)
	Specificity	0.89	(0.81, 0.95)
	Positive predictive value	0.74	(0.40, 0.93)
	Negative predictive value	0.50	(0.26, 0.79)
MIRU	Sensitivity	0.61	(0.51, 0.91)
	Specificity	0.72	(0.62, 0.90)
	Positive predictive value	0.73	(0.46, 0.93)
	Negative predictive value	0.61	(0.35, 0.95)
Spoligotyping	Sensitivity	0.94	(0.78, 0.99)
	Specificity	0.67	(0.44, 0.98)
	Positive predictive value	0.77	(0.38, 0.99)
	Negative predictive value	0.91	(0.57, 0.99)

4.1. Results from three methods, all treated as dichotomous

As discussed above, when three conditionally independent dichotomous tests are used, the problem is identifiable, and no substantive prior input is required for reasonable Bayesian estimation of all parameters. We therefore used uniform (Beta(1,1)) prior distributions over the feasible range (from 0 to 1) for all seven parameters, including the prevalence of clustering θ, and the sensitivities and specificities of the three diagnostic tests. The results, based on the data in Table II, are summarized in Table III.

Treating IS*6110* RFLP as a gold standard would result in a clustering prevalence of 19.1 per cent with 95 per cent confidence interval extending from 15.3 per cent to 23.3 per cent, a relatively small interval. This would imply that approximately 80 per cent of cases are reactivations of previously acquired disease. Considering information from all three methods combined, however, the clustering prevalence rises to 55 per cent, with 95 per cent credible interval of 26–77 per cent, a much wider interval. This is largely driven by not considering any of the methods to be a perfect gold standard, and by the large discrepancies between results from the different methods. The sensitivity of IS*6110* RFLP is estimated to be only 26 per cent, but again with a wide credible

Table IV. Positive likelihood ratios (LR+, see the text for definition) and 95 per cent credible intervals for clustering for each possible test combination.

IS*6110* RFLP	MIRU	Spoligotyping	LR+ (95 per cent credible interval)
+	+	+	17.8 (6.78, 282.10)
+	+	−	0.52 (0.02, 5.77)
+	−	+	3.89 (0.30, 69.56)
+	−	−	0.10 (0.0041, 0.69)
−	+	+	5.8 (2.58, 92.30)
−	+	−	0.17 (0.0070, 2.00)
−	−	+	1.23 (0.17, 22.12)
−	−	−	0.037 (0.0015, 0.13)

interval. Since the sensitivity of 'exact match' IS*6110* RFLP may be low, the criteria for an IS*6110* RFLP match are sometimes relaxed, so that two fingerprints with one band difference are considered as clustered. Allowing for this definition here raises the clustering rate from 19 to 43 per cent, much closer to the rate found by our latent class model, and well within the credible interval from this analysis.

The properties of MIRU and spoligotyping were also generally poor, with the exception of the sensitivity of spoligotyping. It is interesting to compare these results with the properties which would be estimated if IS*6110* RFLP is considered as the gold standard. The sensitivity and specificity of MIRU would have been 49 per cent (95 per cent CI 38–61 per cent) and 55 per cent (95 per cent CI 49–61 per cent), respectively. The same properties for spoligotyping would have been estimated as 80 per cent (95 per cent CI 69–88 per cent) and 38 per cent (95 per cent CI 33–44 per cent), respectively. All of the point estimates are lower and all intervals narrower than the corresponding estimates from the Bayesian latent class model, as shown in Table III.

Table IV presents the positive likelihood ratios for all possible combinations. These likelihood ratios represent the factor which the pre-test odds of being clustered need to be multiplied by to determine the post-test odds of being clustered. Note that the term 'likelihood ratio' is used in the sense of the diagnostic testing literature, and is not equivalent to the LR test from standard hypothesis testing. When all three tests classify the subject as positive (LR+ = 17.8) or negative (LR+ = 0.037), a large degree of confidence can be placed in the test results. Although the credible intervals are wide, a subject with three positive results has *at least* six times the odds of being clustered post-test compared to pre-test, using the lower limit of the LR+ of 6.78. Similarly, a subject with three negative tests has *at least* seven times less the odds of being clustered post-test compared to pre-test, using the reciprocal of the upper credible interval limit for LR+ of 0.13. When at least one of the methods is discordant with the other two, results are much less certain, and only when MIRU and spoligotyping both agree do the LR+ credible intervals not overlap with the null value of 1.

4.2. Results when IS6110 RFLP is treated as continuous NGD data

When analysing the IS*6110* RFLP data converted into the continuous NGD measure, we used N(0, 10,000) prior distributions for the unknown means, and uniform[0, 200] distributions for the unknown standard deviations. Uniform[0,1], or Beta(1,1), prior distributions were used for the properties of the two dichotomous methods and for the prevalence of clustering.

Table V. Posterior medians and 95 per cent credible intervals for the prevalence of clustering, the sensitivities and specificities of the two dichotomous tests for clustering, and the posterior means and medians for the NGD data.

Variable		Posterior median	95 per cent credible interval
Prevalence of clustering		0.25	(0.16, 0.37)
NGD	Mean for unclustered subjects	118.6	(110.5, 127.4)
	SD for unclustered subjects	48.6	(44.4, 53.4)
	Mean for clustered subjects	38.4	(26.1, 53.4)
	SD for clustered subjects	21.3	(10.4, 32.8)
MIRU	Sensitivity	0.78	(0.67, 0.89)
	Specificity	0.65	(0.58, 0.72)
	Positive predictive value	0.43	(0.27, 0.59)
	Negative predictive value	0.90	(0.80, 0.96)
Spoligotyping	Sensitivity	0.97	(0.90, 0.99)
	Specificity	0.45	(0.38, 0.54)
	Positive predictive value	0.37	(0.24, 0.54)
	Negative predictive value	0.98	(0.92, 0.99)

As shown in Table V, the estimated mean NGD value for clustered cases was close to 40, while the mean for unclustered cases was approximately 120, about three times as large. Comparing results from Table III to those in Table V, the most striking difference occurs in the estimated clustering rate, where the posterior median is cut from 55 to 25 per cent. This is presumably in large part owing to the increased weight of the IS*6110* RFLP results, as continuous tests carry more information compared to dichotomous test values. When results from three dichotomous tests are used, all contribute roughly equally to the final results, the difference in information content from one test to another largely depending on their sensitivities and specificities. However, when a continuous test is used, the information content for that test is largely determined by the ratio of heights of the two normal curves for any given observed NDG value, and these ratios can be very large for some points. In practice, this means that the data from the continuous test can sometimes overwhelm that from the dichotomous tests. The sensitivities and specificities of MIRU and spoligotyping also vary between the two analyses, but with considerable overlaps between the two sets of credible intervals. As is the case with the dichotomous analysis reported in the previous section, none of the parameters are accurately estimated, even though close to 400 subjects were available for analysis.

Figure 2 presents a ROC curve for NGD, which shows that regardless of the cut-off value chosen, the test provides far from definitive classification into clustered and non-clustered cases. Nevertheless, a range of NGD cut-off values provide sensitivities and specificities both above 80 per cent. For example, a cut-off value of 70 on the NGD scale gives a sensitivity of 97 per cent and a specificity of 81 per cent. Figure 3 presents the probability of a case being clustered, depending on the four possible groupings of the two dichotomous methods, and as a function of the continuous NGD value. Note in particular the weight of the NGD data for larger values, where, regardless of the MIRU and spoligotyping outcomes, the probability of being in a cluster approaches zero.

Comparing results from Table IV with those in Figure 3 highlights the main difference when switching the form of IS*6110* RFLP data used from dichotomous to continuous. When all three tests are dichotomous, the data seems to indicate very poor sensitivity of IS*6110* RFLP, as both MIRU and spoligotyping are positive much more often compared to IS*6110* RFLP. Since all three

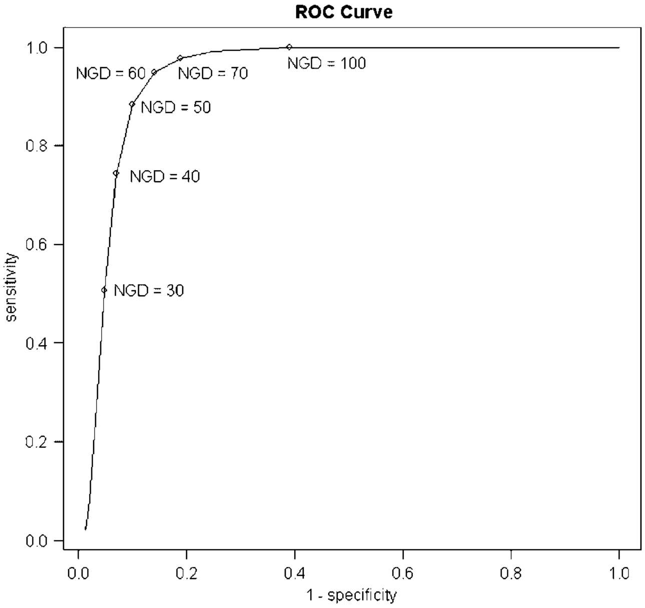

Figure 2. ROC curve for the nearest genetic distance (NGD) measure.

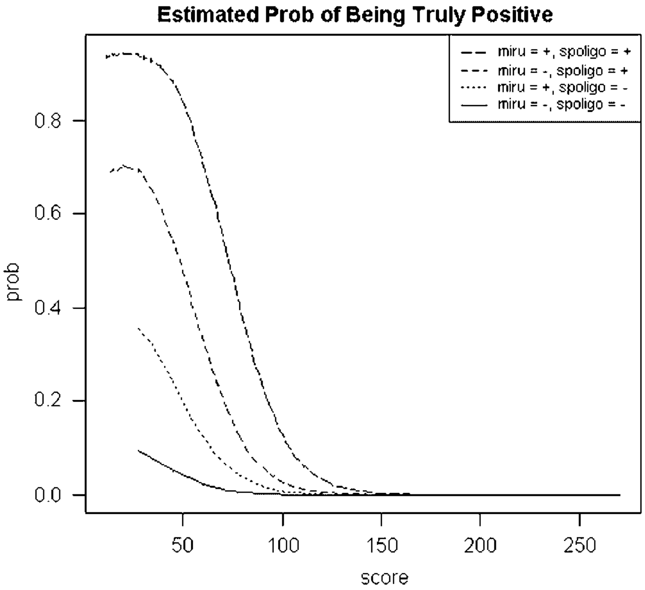

Figure 3. Probability of being clustered as a function of the nearest genetic distance (NGD) grouped by the four possible outcomes of the MIRU and spoligotyping dichotomous tests.

DOI: 10.1002/sim

Table VI. Posterior medians and 95 per cent credible intervals for the odds ratios for the effects of covariates, using data from the two dichotomous tests (MIRU and spoligotyping) and one continuous test (NGD) for clustering.

Variable	Posterior median	95 per cent credible interval
Age	1.01	(0.99, 1.03)
Sex (1 = female, 0 = male)	1.22	(0.57, 2.67)
Canadian born	3.89	(1.41, 13.14)
Haitian born	14.84	(5.59, 53.04)
Smear-positive pulmonary disease	5.81	(1.69, 40.04)
Smear-negative pulmonary disease	4.29	(1.44, 24.46)

Note: For both smear-positive and smear-negative pulmonary disease, the reference category is extrapulmonary disease.

methods carry roughly the same amount of information (since all are dichotomous), IS*6110* RFLP fares poorly in this analysis, and the prevalence rises due to low estimated sensitivity of IS*6110* RFLP. When IS*6110* RFLP data are used in a continuous fashion, it carries more weight than the dichotomous methods because of the increased information content of continuous values from normal curves as discussed above.

4.3. Results on covariates

The posterior distributions of the effects of the covariates are summarized in Table VI. While there was no evidence for an effect from age or sex, Haitian-born cases were more likely to be clustered, as were Canadian-born cases compared to those non-Canadian not born in Haiti. A TB diagnosis can involve smear-positive pulmonary disease (highly contagious), smear-negative pulmonary disease (much less contagious), or extrapulmonary disease (not usually transmissible). Consistent with this, using extrapulmonary disease as a comparator, both positive and negative smear pulmonary disease were associated with clustering, as shown in Table VI.

In all cases, the credible intervals were wide, precluding definitive conclusions about any of the variables investigated. In large part, this is because of the uncertainty induced by the measurement error in the outcomes in the logistic regression model, demonstrating the increased accounting of all inherent uncertainty when performing analyses without a gold standard. Once again, this casts doubt on any analyses of covariates performed without taking into account this considerable extra uncertainty [15, 31].

5. DISCUSSION

We have illustrated how the estimation of TB clustering rates using data from three *M. tuberculosis* genotyping techniques can be accomplished using Bayesian latent class models. To our knowledge, this is the first application of these models to TB clustering data. Our methods incorporated the information from all three methods into a single model, while not considering any method to be a gold standard. Both dichotomous and continuous IS*6110* RFLP data were modelled, likelihood ratios and the individual probabilities of being clustered given any combination of results were produced, covariates were incorporated into a hierarchical version of the model, and in the case

DOI: 10.1002/sim

of continuous data, ROC curves were produced. In particular, we have improved on the previous work [8, 10, 12–15, 32–34] by considering continuous tests, and extending the model to include covariates.

Considering that no method is a gold standard, the wide credible intervals indicate considerable uncertainty about the true rate of clustering in Montreal, and about the properties of the three methods, even though data from almost 400 cases were available. It is clear that country of origin and both smear-positive and smear-negative pulmonary disease are associated with increased risk of recent transmission, but wide credible intervals accompany these results as well.

The wide credible intervals resulting from our analyses are not surprising, given previous work on the large sample sizes typically required in the analysis of data in the absence of a gold standard [32, 33]. As these papers showed, if misclassification errors are to be fully accounted for, data from tens of thousands of cases are often required for accurate prevalence and test property estimation, rather than the few hundred often used in practice, and indicated by standard binomial sample size calculations. Continuous tests potentially provide more information per subject, but further work on this sample size problem is required to specify the magnitude of this improvement.

Our results indicate that poor sensitivity of standard IS*6110* RFLP leads to estimates of clustering that are likely too low. Estimation of the properties of PCR-based tests indicates that these methods are also insufficiently accurate to be used alone. Improved knowledge of the properties of these methods can lead to inclusion of substantive prior distribution into the modelling effort, providing more accurate inferences (assuming that the prior information is accurate). While our findings suggest the need to interpret *M. tuberculosis* genotyping results with caution, IS*6110* typing remains the best method, at least in a low-incidence setting where the population of *M. tuberculosis* isolates shows a high degree of genetic diversity. This is in large part because IS*6110* typing has the slowest evolution rate.

There are a number of limitations to our data which indicate the need for further work in this area. First, the interval defining 'recently transmitted' disease is often considered to last up to two years after infection with *M. tuberculosis*. Therefore, our three-year data set may be too short to obtain the most reliable estimates; two-year lead-in and lag phases might lead to higher accuracy. Second, we have assumed that all *M. tuberculosis* strains evolved at similar rates, an assumption not yet verified in the literature. Third, as discussed by Murray *et al.* [35], band assignment and matching errors, as well as under-sampling on cluster analysis may add further error, although these problems were minimized here with a sampling rate of over 90 per cent [8], and the utilization of software which minimized band assignment and matching errors [10]. Fourth, the assumption of conditional independence between tests may not exactly hold, because the evolution of the DR region can be mediated by IS*6110* activity. We believe that any such activity will be quite small and have little effect on the results, because the most common means of spoligotype evolution involved deletion of spacers, with IS*6110*-mediated changes accounting for a minority of variability [36]. If necessary, methods are available that can account for conditional dependence between tests [34, 37], and these can be extended to continuous data. For example, the normal mean can be made a function of the results of one or both of the dichotomous tests. When three conditionally dependent tests are used the model can become non-identifiable, so adding additional tests, if available, could be useful. Past studies, however, have sometimes concluded that a simpler model that ignores conditional dependence may outperform a model that includes this component [15]. Finally, further studies could check whether the predictions made by the model are confirmed by contact tracing, but contact tracing itself is yet another imperfect test for clustering. It would similarly be interesting to check how the methods work in higher incidence settings.

DOI: 10.1002/sim

Importantly, we assumed that our continuous test results were normally distributed across the population, conditional on cluster status. This assumption is routinely made in the analysis of continuous diagnostic test data, and various authors have shown considerable robustness of the ROC curve to deviations from normality [38]. Clearly, other distributions can be used, and equations (1) and (2) modified as needed, with estimation proceeding *via* the Gibbs sampler. Further work on non-parametric models for continuous diagnostic test data along the lines of Erkanli *et al.* [18] would also be useful.

REFERENCES

1. World Health Organization (WHO). *Eighth Annual Report on Global TB Control*, WHO, 2004.
2. Rivest P, Tannenbaum T, Bedard L. Epidemiology of tuberculosis in Montreal. *Canadian Medical Association Journal* 1998; **158**:605–609.
3. Barnes PF, Cave MD. Current concepts: molecular epidemiology of tuberculosis. *New England Journal of Medicine* 2003; **349**:1149–1156.
4. van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM, Small P. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *Journal of Clinical Microbiology* 1993; **31**:406–409.
5. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *Journal of Clinical Microbiology* 1997; **35**:907–914.
6. Hayward AC, Watson MJ. Typing of mycobacteria using spoligotyping. *Thorax* 1998; **53**:329–330.
7. 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 of the United States of America* 2001; **98**:1901–1906.
8. Scott AN, Menzies D, Tannenbaum TN, Thibert L, Kozak R, Joseph L, Schwartzman K, Behr MA. Sensitivities and specificities of spoligotyping and mycobacterial interspersed repetitive unit-variable-number tandem repeat typing methods for studying molecular epidemiology of tuberculosis. *Journal of Clinical Microbiology* 2005; **43**:89–94.
9. Gori A, Bandera A, Marchetti G, Degli Esposti A, Catozzi L, Nardi GP, Gazzola L, Ferrario G, van Embden JD, van Soolingen D, Moroni M, Franzetti F. Spoligotyping and *Mycobacterium tuberculosis*. *Emerging Infectious Diseases* 2005 **11**:1242–1248.
10. Salamon H, Behr MA, Rhee JT, Small PM. Genetic distances for the study of infectious disease epidemiology. *American Journal of Epidemiology* 2000; **151**:324–334.
11. Gastwirth JL, Johnson WO, Reneau DM. Bayesian analysis of screening data: application to AIDS in blood donors. *Canadian Journal of Statistics* 1991; **19**:135–150.
12. 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.
13. Demissie K, White N, Joseph L, Ernst P. Bayesian estimation of asthma prevalence, and comparison of exercise and questionnaire diagnostics in the absence of a gold standard. *Annals of Epidemiology* 1998; **8**:201–208.
14. Johnson WO, Gastwirth JL, Pearson LM. Screening without a 'gold standard': the Hui–Walter paradigm revisited. *American Journal of Epidemiology* 2001; **153**:921–924.
15. Gustafson P. On model expansion, model contraction, identifiability, and prior information: two illustrative scenarios involving mismeasured variables (with Discussion). *Statistical Science* 2005; **20**:111–129.
16. Zhou K, O'Malley J. A Bayesian hierarchical non-linear regression model in receiver operating characteristic analysis of clustered continuous diagnostic data. *Biometrical Journal* 2005; **47**:417–427.
17. Choi YK, Johnson WO, Thurmond MC. Diagnosis using predictive probabilities without cut-offs. *Statistics in Medicine* 2006; **25**:699–717.
18. Erkanli A, Sung M, Costello EJ, Angold A. Bayesian semi-parametric ROC analysis. *Statistics in Medicine* 2006; **25**(22):3905–3928.
19. van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD. Occurrence and stability of insertion sequences in *Mycobacterium tuberculosis* complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *Journal of Clinical Microbiology* 1991; **29**:2578–2586.

DOI: 10.1002/sim

20. Hermans PW, van Soolingen D, Dale JW, Schuitema AR, McAdam RA, Catty D, van Embden JD. Insertion element IS986 from *Mycobacterium tuberculosis*: a useful tool for diagnosis and epidemiology of tuberculosis. *Journal of Clinical Microbiology* 1990; **28**:2051–2058.
21. Segal MR, Salamon H, Small PM. Comparing DNA fingerprints of infectious organisms. *Statistical Science* 2000; **15**:27–45.
22. Rhee JT, Tanaka MM, Behr MA, Agasino CB, Paz EA, Hopewell PC, Small PM. Use of multiple markers in population-based molecular epidemiologic studies of tuberculosis. *International Journal of Tuberculosis and Lung Disease* 2000; **4**:1111–1119.
23. Warren RM, van der Spuy GD, Richardson M, Beyers N, Booysen C, Behr MA, van Helden PD. Evolution of the IS*6110*-based restriction fragment length polymorphism pattern during the transmission of *Mycobacterium tuberculosis*. *Journal of Clinical Microbiology* 2002; **40**:1277–1282.
24. Daley CL, Small PM, Schecter GF, Schoolnik GK, McAdam RA, Jacobs Jr WR, Hopewell PC. An outbreak of tuberculosis with accelerated progression among persons infected with the human immunodeficiency virus. An analysis using restriction-fragment-length polymorphisms. *New England Journal of Medicine* 1992; **326**:231–235.
25. de Boer AS, Borgdorff MW, de Haas PE, Nagelkerke NJ, van Embden JD, van Soolingen D. Analysis of rate of change of IS*6110* RFLP patterns of Mycobacterium tuberculosis based on serial patient isolates. *Journal of Infectious Diseases* 1999; **180**:1238–1244.
26. Cowan LS, Diem L, Monson T, Wand P, Temporado D, Oemig TV, Crawford JT. Evaluation of a two-step approach for large-scale, prospective genotyping of *Mycobacterium tuberculosis* isolates in the United States. *Journal of Clinical Microbiology* 2005; **43**:688–695.
27. 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.
28. Joseph L, Gyorkos T. Inferences for likelihood ratios in the absence of a gold standard. *Medical Decision Making* 1996; **16**:412–417.
29. Titterington D, Smith AFM, Makov UM. *Statistical Analysis of Finite Mixture Distributions*. Wiley: New York, 1985.
30. Hanley JA. The use of the 'binormal' model for parametric ROC analysis of quantitative diagnostic tests. *Statistics in Medicine* 1996; **15**:1575–1585.
31. Greenland S. Multiple-bias modeling for analysis of observational data. *Journal of the Royal Statistical Society, Series A* 2005; **168**:267–306.
32. Rahme E, Joseph L, Gyorkos T. Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Applied Statistics* 2000; **49**:119–228.
33. Dendukuri N, Rahme E, Bélisle P, Joseph L. Bayesian sample size determination for prevalence and diagnostic studies in the absence of a gold standard test. *Biometrics* 2004; **60**:388–397.
34. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 2001; **57**:208–217.
35. Murray M. Sampling bias in the molecular epidemiology of tuberculosis. *Emerging Infectious Disease* 2002; **8**:363–369.
36. Warren R. Microevolution of the direct repeat region of Mycobacterium tuberculosis: implications for interpretation of spoligotyping data. *Journal of Clinical Microbiology* 2002; **40**:4457–4465.
37. Black MA, Craig BA. Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine* 2002; **21**:2653–2669.
38. Hajian-Tilaki KO, Hanley JA, Joseph L, Collet JP. A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. *Medical Decision Making* 1997; **17**:94–102.

DOI: 10.1002/sim