Journal of Biomedical Informatics 53 (2015) 237–242

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Towards probabilistic decision support in public health practice: Predicting recent transmission of tuberculosis from patient attributes

Hiroshi Mamiya [a,e,*], Kevin Schwartzman [d,e,f], Aman Verma [a,e], Christian Jauvin [a], Marcel Behr [b,e,f], David Buckeridge [a,c,e]

[a] McGill Clinical and Health Informatics, McGill University, 1140 Avenue Pine, Montréal, Québec H3A 1A3, Canada
[b] McGill University Health Centre, 1650 Cedar Avenue, Room A5.156, Montreal, H3G 1A4, Canada
[c] Agence Sociosanitaire de Montréal, Direction de la santé publique, 1301 Rue Sherbrooke Est, Montreal, Quebec H2L 1M3, Canada
[d] Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute, McGill University Health Centre, 3650 Rue Saint-Urbain, Montreal, Quebec H2X 2P4, Canada
[e] Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, Quebec H3A 1A2, Canada
[f] McGill International TB Centre, McGill University Health Centre, 1650 Cedar Avenue, Room A5.156, Montreal, Quebec, H3G 1A4, Canada

ARTICLE INFO

Article history:
Received 29 September 2014
Accepted 11 November 2014
Available online 20 November 2014

Keywords:
Decision support model
Statistical prediction
Public health
Tuberculosis
Transmission

ABSTRACT

Objective: Investigating the contacts of a newly diagnosed tuberculosis (TB) case to prevent TB transmission is a core public health activity. In the context of limited resources, it is often necessary to prioritize investigation when multiple cases are reported. Public health personnel currently prioritize contact investigation intuitively based on past experience. Decision-support software using patient attributes to predict the probability of a TB case being involved in recent transmission could aid in this prioritization, but a prediction model is needed to drive such software.

Methods: We developed a logistic regression model using the clinical and demographic information of TB cases reported to Montreal Public Health between 1997 and 2007. The reference standard for transmission was DNA fingerprint analysis. We measured the predictive performance, in terms of sensitivity, specificity, negative predictive value, positive predictive value, the Receiver Operating Characteristic (ROC) curve and the Area Under the ROC (AUC).

Results: Among 1552 TB cases enrolled in the study, 314 (20.2%) were involved in recent transmission. The AUC of the model was 0.65 (95% confidence interval: 0.61–0.68), which is significantly better than random prediction. The maximized values of sensitivity and specificity on the ROC were 0.53 and 0.67, respectively.

Conclusions: The characteristics of a TB patient reported to public health can be used to predict whether the newly diagnosed case is associated with recent transmission as opposed to reactivation of latent infection.

1. Background and significance

Tuberculosis (TB) is a communicable disease caused by *Mycobacterium tuberculosis* (MTB). Despite organized control efforts, TB continues to occur in developed countries. In Canada, TB is concentrated in immigrants from high TB incidence countries, inner-city residents, and Aboriginal persons [1]. Upon infection, approximately 90% of individuals remain asymptomatic and non-infectious with latent tuberculosis infection (LTBI). After months or years of latency, approximately 5–10% of persons with LTBI develop active TB disease due to a complex array of biological, genetic and environmental factors [2]. Individuals with reactivated TB can transmit infection to others in the absence of timely detection and intervention.

Contact investigation is a core public health strategy to prevent and control TB. It involves identification, medical evaluation, and treatment of individuals who have had contact with a newly diagnosed case, often called an index case. Evidence of recent infection or active TB disease among contacts suggests ongoing transmission, and treatment is provided to infected individuals to prevent subsequent active TB, thereby interrupting transmission.

A more recent approach to identifying TB transmission is DNA fingerprint analysis. With this approach, DNA is isolated from the MTB organisms cultured from patient samples. Mycobacterial DNA is then characterized with respect to the presence and number of target sequences; transmission is assumed to have occurred

* Corresponding author at: 1140 Avenue Pine, Montréal, Québec H3A 1A3, Canada.
E-mail address: hiroshi.mamiya@mail.mcgill.ca (H. Mamiya).

http://dx.doi.org/10.1016/j.jbi.2014.11.006
1532-0464/

between cases with matching "fingerprints". Although this method is capable of identifying transmission involving persons with limited contact, results can take weeks [3], and the method can be applied only to persons with active disease from whom it is possible to obtain a positive MTB culture. In particular, DNA fingerprinting is not able to establish transmission to individuals with LTBI, or to those with culture-negative active TB (a frequent manifestation in children, for example). For these reasons, contact investigation remains critical to the rapid assessment and interruption of transmission.

Ideally, contact investigation should be conducted for all infectious cases immediately upon diagnosis. In practice, however, limited resources in public health may necessitate prioritization of contact tracing among multiple infectious individuals. In general, patient features related to infectiousness, such as pulmonary and laryngeal disease, cavitary lesions on chest radiography, positive sputum acid fast smears, and younger age are considered when assessing the urgency of investigation [4].

Previous molecular epidemiological investigations of TB transmission have identified additional clinical and demographic predictors associated with involvement in transmission chains. These features include HIV infection, drug-resistant TB, homelessness, and intravenous (IV)-drug use [5]. The degree to which these features are used to prioritize contact tracing depends on the intuition and experience of public health officials, and the potential use of these patient features to predict the probability that a case is involved in recent transmission has not been explored.

Statistical and machine learning algorithms, which estimate the probability of an event as a function of input variables, can analyze many patient variables to assist medical decision making. Such prediction models can inform decisions about diagnosis and therapy when applied to patient data contained in electronic health records (EHR) [6]. Although the models are most frequently used for clinical decision support, their application in public health is rare. In TB control, using known risk factors for transmission associated with a newly diagnosed active TB case to predict recent transmission appears feasible and should aid timely and evidence-based decision making in prioritization of contact investigation.

The rapid identification of community transmission allows timely intervention. Although it is often considered the gold-standard for detection of transmission, the impact of DNA fingerprinting is hampered by its slow turnaround time. A decision-support tool that uses readily available clinical and demographic features of an active TB case would permit more rapid, evidence-based decision making in prioritizing contact investigation. As an initial step towards creating such a decision-support tool, we developed and evaluated a statistical learning model to estimate the probability of a given case of active TB being involved in recent transmission.

2. Materials and methods

2.1. Source of data

The data used to develop and evaluate the model were obtained from 1844 active TB cases reported to the public health department between January 1, 1996 and December 31, 2007 in Montreal, Quebec, Canada. In Quebec, as in all provinces in Canada and states in USA, every diagnosis of active TB must be reported by name to the local public health department along with standardized demographic, clinical, and microbiologic information. Hence, as part of routine public health practice, clinical and epidemiological data were collected by public health nurses from patients and treating clinicians and were stored in a database. We extracted these data in non-nominal form for our study, which was approved by the McGill Faculty of Medicine Institutional Review Board.

2.2. Definition of the dependent variable

We used DNA fingerprinting based on the IS*6110* target sequence, by restriction-fragment length polymorphism (RFLP) analysis, to assess the involvement of each case in recent transmission. IS*6110* is a repetitive DNA sequence in the MTB genome, and its frequency and insertion location vary from one MTB strain to another. This sequence is highly preserved, however, as the bacteria propagate from one host to another, thus making it possible to identify the same MTB strain in cases belonging to a chain of transmission [7]. Based on standardized IS*6110* – RFLP methodology [8], cases with MTB isolates that shared identical numbers and insertion locations of the IS*6110* sequence were deemed members of the same TB "cluster." Cases with unique IS*6110* patterns were deemed unique. As the discriminative ability of the RFLP method is compromised for MTB strains with few copies of the IS*6110* element [9], we used the results of a secondary genotyping method, spoligotyping, for strains that contain less than six IS*6110* elements.

2.3. Selection of predictors

Table 1 lists the independent variables initially included in the prediction model, which were selected by review of previous epidemiological studies exploring transmission of TB. For the *countries of origin* variable, we created three categories: Canadian born, Haitian born, and born in other countries. In previous work, Haitian birth was identified as a risk factor for transmission in Montreal [10], hence we used a distinct category for these individuals. For the *Area of residence* variable, we used health administrative areas on the island of Montreal (Centres de santé et de services sociaux – CSSS) as a unit of the analysis. There are 10 CSSS areas in Montreal, and areas with a similar frequency of genotype-defined clustering were merged to create four areas of residence. Multi-drug resistant TB (MDR-TB) disease, which is resistant to at least two of the main anti-TB drugs, Isoniazid (INH) and rifampicin, has been most strongly associated with recent transmission [11–13]. However, MDR-TB disease rarely occurs in Canada [14], so we considered for inclusion in our model the more frequently observed INH

Table 1
Patient features used to predict transmission of TB.

Name of features	Format	%Missingness
Site of infection (Pulmonary involvement or not)	Binary	0.00
Sputum AFB smear positive	Binary	7.02
Previous diagnosis of TB (Active or Latent Tuberculosis)	Binary	8.67
Cavitary lesion on chest X-ray	Binary	0.39
Drug-resistant disease	Binary	3.54
HIV test result	Binary	48.52
Age (year)	Continuous	0.26
Country of origin	Categorical	0.19
Gender	Binary	0.97
Being Canadian born Aboriginal	Binary	88.98
Being homeless	Binary	85.63
Being alcoholic	Binary	24.16
Being intravenous drug user	Binary	24.03
Area of residence on the island of Montreal by health administrative region	Categorical	2.57
Living in apartment	Binary	0.00
Coughing	Binary	4.38

resistance. The *Previous diagnosis of TB* variable included previous diagnosis of active disease and LTBI.

2.4. Multiple imputation

Because excluding the cases with missing data would reduce the size of the dataset and possibly introduce bias [15], we filled in the missing values using multiple imputation, which replaces missing information with plausible values drawn from a conditional distribution formed by the observed covariates [16]. We used fully conditional specification approach [17], where multivariable imputation model is specified for each incomplete variables. As an example, a linear regression model was fit to the observed data in order to predict the missing values of the continuous age variable. Similarly, logistic regression and multinomial regression models were specified to predict the missing values of binary and categorical variables. We used all the variables in our dataset shown in Table 1 as the predictor variables for the imputations models.

Imputation was repeated multiple times to generate m different complete datasets, each of which was used to train and test the model [16]. The resulting m measured performances of prediction and their corresponding variances were pooled into a single final estimate using Rubin's rule [16]. In this study, 15 imputed datasets were created, which should be sufficient given the proportion of missingness [16]. We used the MICE (Multivariate Imputation by Chained Equation) package [18] in the R programming environment [19] to generate imputed datasets. Homelessness and Aboriginal status were not useful predictors and were not explored further, since both variables had a very large proportion of missing values. Missing values in all other predictors, including HIV infection, were imputed.

2.5. Prediction model

A logistic regression model was used to predict whether a given case was involved in recent transmission or was reactivation of LTBI. We selected logistic regression due to its wide recognition in the medical community [20], and its comparable classification performance to other methods such as artificial neural networks and support vector machines [21–24]. In addition, it provides a straightforward interpretation of the predictive power of independent variables in the form of an odds ratio [23].

Interaction of predictor variables may occur, so the plausibility of adding interaction terms was assessed by the following selection approach; the Bayesian Information Criterion (BIC) of a model containing all the predictors (i.e. main model) was compared to the BIC derived from the same model plus the interaction term. Lower BIC indicates better model fit, and BIC tends to penalize the larger model (i.e. the model with additional parameter) [25]. Therefore, the interaction term was considered a strong predictor if the model containing the term showed a lower BIC than the main model. Because BIC-based model selection is substantially more conservative than other statistical criteria such as Akaike's Information Criterion and the likelihood ratio test due to its stronger penalization mechanism, we were less likely to select interaction terms resulting from noise of the development data. All possible combinations of two variables were tested in this fashion. We performed the screening of interaction terms before the model selection process described below, since testing all possible combinations of interaction terms by the model selection method (Bayesian Model Averaging) can result in a very large model space. As well, we assessed the linearity of the continuous age variable in relation to the logit of the dependent variable. To do this analysis, we used a fractional polynomial, which tests the feasibility of linear and various non-linear functions of continuous variables [26].

2.6. Model selection

We identified the optimal model using Bayesian Model Averaging (BMA). BMA estimates the posterior probabilities of all possible models given the training data, chooses a set of candidate models according to their posterior probability distributions, and averages the coefficients of the selected models using the posterior probabilities as weights [27]. Posterior model probability for each of candidate models can be analytically approximated by BIC, as described by Hoeting et al. [27].

BMA assumes that no single model will correctly explain the observed data and creates a synthetic model as an approximation to a true unobserved model. Unlike conventional model selection approach for generalized linear model, such as stepwise selection which bases inference and prediction on a single best model, BMA offers a solution to account for model uncertainly by averaging the coefficient of multiple models [27]. Compared to stepwise model selection, BMA was shown to perform superior in selecting collect model [28] and often estimate smaller standard deviations for the parameters of interest [29].

Instead of determining the effect of a predictor by its p-value from statistical significant test, BMA reports the posterior probability of each predictor. This value is simply the sum of the posterior probabilities of the models that contain a given variable, and represents the normalized probability of a coefficient having a non-zero value given the training data, $P(\beta \neq 0|D)$. A posterior effect probability of more than 0.5 indicates evidence for the effect of a variable. For prediction, we included in the final model variables whose posterior effect probability were not zero. We implemented our model in the R programming environment using the BMA package [30].

2.7. Assessment of predictive performance

The discriminative performance of the model was assessed by plotting the Receiver Operating Characteristic (ROC) curve and calculating the Area Under the ROC Curve (AUC). In addition, we calculated negative predictive value (NPV) and positive predictive value (PPV) of the model over a range of classification thresholds. All estimates were obtained through 10 repeats of 10-fold cross-validation. Since we generated 15 imputed datasets, we applied the ten repeats of 10-fold cross-validation to each of these imputed data sets, producing 15 sets of estimates (i.e. AUC, ROC, PPV, NPV, sensitivity, and specificity), which we pooled using Rubin's rule [16].

3. Results

Of the 1829 active TB cases reported to the public health department during the study period, we excluded cases with negative or missing culture results (199 cases, 10.9%), and culture positive cases with missing DNA fingerprinting results (78 cases, 4.3%), leaving 1552 cases (84.9%) to train and test the model. Of these 1552 cases, 314 (20.2% of the enrolled cases) were clustered according to IS6110 RFLP and spoligotyping. 107 distinct matching patterns were observed, representing 107 TB clusters. The vast majority of TB clusters (67%, or 72 clusters) consisted of 2 cases; in other words, most clusters contained only one putative secondary case. The small proportion of clustered cases suggests a low level of ongoing transmission in Montreal.

Table 2 shows the distributions of potential predictors among clustered and unique cases. An interaction term for *Living in apartment* and *Cavitary lesion on chest X-ray* was included in the prediction model, since the model containing the product of the two variables had a lower BIC than the model without it. No

Table 2
Distribution of patient characteristics associated with recent transmission of TB.

Predictor	Unique cases (N = 1238, 79.8%)				Clustered cases (N = 314, 20.2%)			
			Missing data				Missing data	
	N[e]	%[d]	N	%	N	%[d]	N	%
Median age	39.7	28.9–61.9	4		40.1	27.3–54.5	0	
Living in apartment	606	49.0	0	0.0	146	46.5	0	0.0
Female	550	44.4	10	0.8	143	45.5	5	1.6
Previous diagnosis of TB	134	10.8	115	9.3	25	8.0	20	6.4
Cavitary lesion on chest X-ray[b]	213	24.6	4	0.5	70	29.4	2	0.8
HIV positive	82	6.6	630	50.9	46	14.6	123	39.2
Intravenous drug use	12	1.0	300	24.2	8	2.5	73	23.2
Alcoholic	163	13.2	302	24.4	47	15	73	23.2
Coughing	667	53.9	52	4.2	173	55.1	16	5.1
Infection with INH resistant strain[c]	100	8.1	48	3.9	25	8.0	7	2.2
Residential location on the island of Montreal (CSSS Area)								
Area A	434	35.1	2	0.2	65	20.7	1	0.3
Area B	489	39.5			164	52.2		
Area C	243	19.6			53	16.9		
Area D	70	5.7			31	9.9		
Country of origin								
Canada	185	14.9	30	2.4	77	24.5	10	3.2
Haiti	151	12.2			75	23.9		
Other countries	872	70.4			152	48.4		
Pulmonary TB	865	69.9	0	0.0	238	75.8	0	0.0
Sputum AFB smear positive[a]	422	48.8	86	9.9	120	50.4	23	9.7
Being aboriginal	1	0.1	1118	90.3	6	1.9	263	83.8
Being homeless	4	0.3	1046	84.5	5	1.6	283	90.1

[a] Among pulmonary TB cases.
[b] Among pulmonary TB cases.
[c] Among cases with resistant TB strain.
[d] Interquartile range was used instead of proportion for age.
[e] Median age was used instead of count for age variable.

Table 3
Regression coefficients and posterior effect probabilities of the predictor variables.

Predictors	OR	95% CI[a]	$P(\beta \neq 0\|D)$
Age (10 years)	0.93	(0.84–1.04)	0.71
Living in apartment	0.98	(0.81–1.17)	0.09
Female	1.00		0.00
Previous diagnosis of TB	0.99	(0.88–1.12)	0.02
Cavitary lesion on chest X-ray[e]	1.00		0.00
HIV positive	2.00	(1.22–3.27)	0.96
Intravenous drug Use	1.00		0.00
Alcoholic	1.00	(0.95–1.05)	0.01
Coughing	1.00		0.00
Infection with INH Resistance strain	1.00	(0.95–1.05)	0.01
Apartment X Cavitary[b]	2.22	(1.44–3.42)	0.99
CSSS Area[c]			
CSSS Area A	0.79	(0.47–1.31)	0.51
CSSS Area B	1.03	(0.84–1.26)	0.01
CSSS Area C	1.47	(0.66–3.26)	0.53
Country of origin[d]			
Canada	2.27	(1.58–3.26)	1.00
Haiti	2.34	(1.60–3.44)	1.00
Pulmonary TB and smear results[e]			
Pulmonary & smear negative	1.01	(0.90–1.13)	0.04
Pulmonary & smear positive	1.00		

$P(\beta \neq 0|D)$ = Posterior effect probability.
OR = Odds Ratio.
[a] 95% CI was calculated only for the predictors that had non-zero posterior effect probabilities.
[b] Indicates interaction term between the *Living in apartment* and *Cavitary lesion* in chest X-ray variable.
[c] Indicator variables were created for categorical variable. Its reference category is *Area D*.
[d] Indicator variables were created for categorical variable. Its reference category is *Other countries*.
[e] These are also indicator variables, and their reference category is a case having extra-pulmonary TB.

transformation of the continuous variable (*Age*) was needed, as the analysis of its functional relationship with the logit of the dependent variable by fractional polynomial suggested a linear relationship (data not shown).

Table 3 shows the regression coefficients and posterior effect probabilities of the predictors estimated by BMA. In general, most variables that we considered had strong evidence *against* their effects, as represented by the small values of their posterior effect probabilities. Age of the patients and residential location on the island of Montreal appeared to have modest positive evidence for an effect. Being Canadian or Haitian born appeared to be the strongest predictors of involvement in recent transmission. In addition, HIV infection and the interaction of living in an apartment and having cavitary TB disease showed strong positive evidence.

Fig. 1 presents the classification performance of the model as summarized by the ROC curve. The AUC of 0.65 (95% CI: 0.61–0.68) indicates that the performance is significantly superior to random prediction. The maximized accuracy of the model (the point of the curve closest to the top-left corner of the graph) is at a sensitivity of 0.53 and a specificity of 0.67.

As seen in Table 4, the values of NPV remained substantially higher than the prevalence of unique cases in the original dataset throughout the range of the sensitivity (i.e. 0.8). On the other hand, PPV was lower than the prior probability of clustering (i.e. 0.2) except at very low sensitivity.

4. Discussion

We assessed the feasibility of predicting whether a newly diagnosed TB case reported to the public health department was involved in a chain of recent transmission. The discriminative performance of the model we developed was superior to random prediction, suggesting that the model has some potential to aid

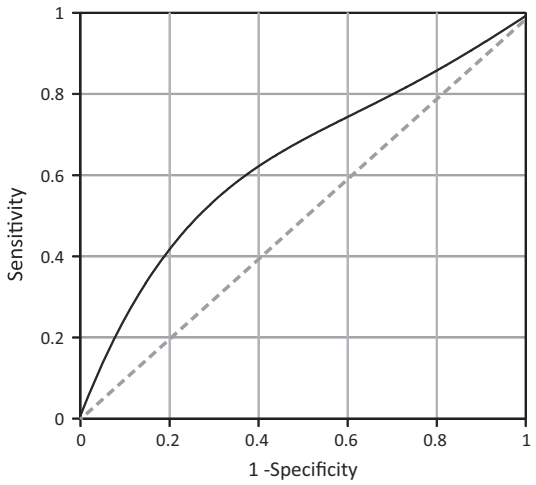

Fig. 1. ROC curve showing the classification performance of the model. Dashed line indicates random prediction. AUC of the ROC was 0.65(95% CI: 0.61–0.68).

Table 4
Specificities, PPVs, and NPVs of the model at various sensitivities.

Sensitivity	Specificity	PPV	NPV
0.95	0.04	0.07	0.92
0.75	0.37	0.08	0.96
0.53	0.67	0.11	0.95
0.5	0.70	0.11	0.95
0.25	0.88	0.14	0.94
0.05	0.99	0.28	0.93

decision-support in a public health practice setting. The probabilistic interpretation of the AUC of our model in the context of TB control is as follows: given two reported TB cases to public health authority, one being involved in recent transmission and the other being an isolated case, there is a 0.65 probability that the model will suggest the prioritization of the case involved in recent transmission. Even though an AUC of 0.65 may represent only a modest improvement over chance discrimination, the classification accuracy of our model is likely to be optimal for the available set of predictor variables at the time of diagnosis. Such a model has the potential to promote accurate and reproducible decision-making in situations where public health practitioners may have not received adequate training and may be forced to prioritize investigation of TB cases due to resource constraint in local health authorities. The strongest predictor of recent transmission is the interaction of cavitary lesion and living in apartment, followed by HIV infection. Age and the area of residence are found to have modestly predictive powers.

Across the range of the classification thresholds considered, the NPV was high, while the PPV was low. A decision-support algorithm based on this model is therefore likely to be most useful in identifying unique cases or excluding recent transmission. The low PPV may be due to a small number of strong predictors, which occur with a low frequency among cases involved in recent transmission.

A challenge associated with predicting recent TB transmission is the influence of many clinical and non-clinical variables on the process of transmission. Prediction of individual clinical outcomes usually considers individual characteristics, such as demographic variables, medical history, clinical observations, and laboratory data. Predicting disease transmission, however, is likely to be improved by the addition of features, such as susceptibility of contacts, dwelling characteristics that facilitate transmission, and the nature of interactions between a person with TB disease and his/her contacts. Although these data would likely improve transmission in our model, they are not captured routinely. Our research focused on attributes of cases that are readily available at the time the case is reported to public health authorities, thus allowing timely decision-making.

Utilization of prediction models to aid public health intervention is uncommon, and its application is largely limited to selective screening of case-finding activities for sexually transmitted infections [31]. A simple analogue of a decision support algorithm for the prioritization of contact investigation is the Syphilis Reactor Grid, a table algorithm to rank the investigation based on the age, sex, and laboratory result of a notified case [32]. Although simple to use, the amount of information incorporated into the decision algorithm is limited. In part, the difficulty of implementing prediction algorithm in daily public health practice is due to the unavailability of relevant patient attributes to public health department at the time of diagnosis. Increasing adoption of EHRs and the progress of health information exchange is opening new opportunities for local health departments to utilize the data in clinical information systems, from automated notification of reportable diseases and syndromic surveillance to beyond [33]. The interoperability between clinical and public health information systems would enable querying relevant clinical and demographic characteristics of reported case of notifiable disease for investigational purposes upon notification of reportable disease to local health department [34]. Such data can be immediately synthesized to generate the probability of recent transmission within the information system in local health departments, thereby promoting timely resource allocation for response, particularly in resource-limited settings.

Strengths of this study include the use of multiple imputation to simulate missing values. Without this method, we would have discarded more than 50% of observations due to the high proportion of missing status for HIV infection, leading to decreased precision and a potentially biased estimate of prediction if the distribution of input and dependent variables are associated with the missingness of the HIV status. In addition, recent transmission in our reference data was objectively determined by DNA fingerprint analysis.

Although specific in identifying clusters, DNA fingerprint-based investigation has its own limitations. Because the method is applicable only to culture-positive active TB cases, individuals who contracted latent infection only, and culture negative cases could not be included. For example, children are less likely to develop culture positive disease, and persons with LTBI who are uninfected by HIV are much less likely to develop active disease than those who are HIV-coinfected. Even though the proportion of active TB cases with missing or negative culture results was relatively small, we did not have systematic data on LTBI in contacts, so we could not use this as a complementary index of transmission for purposes of the present analysis.

Similarly, clusters defined by DNA fingerprinting do not necessarily capture transmission events that are already appropriately managed by successful identification, evaluation and treatment of latently infected contacts, since such contacts do not develop active disease. Therefore, in a setting where contact investigation is generally successful, as in Montréal, transmission events identified by genotyping tend to involve casual contacts, where links to a given index case may not be captured by interview, no matter how thorough. In addition, because active TB cases with well-known risk factors of transmission receive intense intervention (including treatment of infected contacts), further transmission and development of active disease among contacts are more likely to be prevented. Thus, the relationship of the risk factors and transmission is obscured in the presence of active intervention in the training data. Despite the little predictive power of clustering observed in

our model, however, positive sputum smear plays a primary role in assessing the urgency of contact investigation in current public health practice [4]. In practice, the prioritization of investigation should be initially determined based on the established factors of infectiousness, and the probabilistic assessment of transmission by the model will subsequently aid ranking investigation based on other attributes, thus combining existing knowledge and predictive analysis.

Further research is required to measure the added value of the predictive model to a TB control program. The classification performance of the model was better than random prediction, but its performance relative to current practice (i.e. judgment as to the likelihood of transmission by public health practitioners) should be ascertained, preferably in prospective evaluation. How to translate the probabilistic output to the decision making in public health setting would depend on the balance of available public health resource, incidence of active tuberculosis and recent transmission, and the goal of local TB control program.

5. Conclusions

We have demonstrated that patient attributes available at the time of disease reporting can be used to predict whether a TB case reported to public health is involved in recent transmission. Models such as the one we developed have the potential to be embedded in a decision support tool and guide evidence-based public health practice. The use of such models and decision-support tools are likely to become increasingly feasible in the future as clinical and demographic data become rapidly available in an electronic format to public health authorities.

Acknowledgments

The authors thank Ms. Fiona Macintosh and Mr. Carmine Rossi for providing the results of DNA fingerprint analysis and Ms. Kimberly Kotar for maintaining the tuberculosis database. The research in this paper was supported by the Canadian Institute of Health Research (MOP-93587 and MOP-84493).

References

[1] Tuberculosis in Canada 2009 – pre-release [Internet]; 2009. <http://publications.gc.ca/collections/collection_2012/aspc-phac/HP37-5-1-2009-eng.pdf>.
[2] Raviglione MC. Tuberculosis. The essentials: Lung Biology in Health and Disease, 4th ed.; 2009.
[3] Nguyen LN, Gilbert GL, Marks GB. Molecular epidemiology of tuberculosis and recent developments in understanding the epidemiology of tuberculosis. Respirology 2004;9(3):313–9.
[4] Centers for Disease Control and Prevention. Guidelines for the investigation of contacts of persons with infectious tuberculosis; recommendations from the National Tuberculosis Controllers Association and CDC, and Guidelines for using the QuantiFERON®-TB Gold test for detecting Mycobacterium tuberculosis infection, United States. MMWR, vol. 54(RR-15); 2005.
[5] Nava-Aguilera E, Andersson N, Harris E, Mitchell S, Hamel C, Shea B, et al. Risk factors associated with recent transmission of tuberculosis: systematic review and meta-analysis. Int J Tuberc Lung D 2009;13(1):17–26.
[6] Tamblyn R, Eguale T, Buckeridge DL, Huang A, Hanley J, Reidel K, et al. The effectiveness of a new generation of computerized drug alerts in reducing the risk of injury from drug side effects: a cluster randomized trial. J Am Med Infor Assoc: JAMIA 2012.
[7] Van Soolingen D. Molecular epidemiology of tuberculosis and other mycobacterial infections: main methodologies and achievements. J Intern Med 2001;249(1):1–26.
[8] van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, et al. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol 1993;31(2):406–9.
[9] Goyal M, Saunders NA, van Embden JD, Young DB, Shaw RJ. Differentiation of Mycobacterium tuberculosis isolates by spoligotyping and IS6110 restriction fragment length polymorphism. J Clin Microbiol 1997;35(3):647–51.
[10] Kulaga S, Behr M, Musana K, Brinkman J, Menzies D, Brassard P, et al. Molecular epidemiology of tuberculosis in Montreal. CMAJ 2002;167(4):353–4.
[11] Alland D, Kalkut GE, Moss AR, McAdam RA, Hahn JA, Bosworth W, et al. Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. N Engl J Med 1994;330(24):1710–6.
[12] Farnia P, Mohammadi F, Masjedi MR, Varnerot A, Zarifi AZ, Tabatabee J, et al. Evaluation of tuberculosis transmission in Tehran: using RFLP and sploigotyping methods. J Infect 2004;49(2):94–101.
[13] Gutierrez MC, Vincent V, Aubert D, Bizet J, Gaillot O, Lebrun L, et al. Molecular fingerprinting of Mycobacterium tuberculosis and risk factors for tuberculosis transmission in Paris, France, and surrounding area. J Clin Microbiol 1998;36(2):486–92.
[14] Long R. Canadian tuberculosis standards, 6th ed.; 2007.
[15] Vergouwe Y, Royston P, Moons KG, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol 2010;63(2):205–14.
[16] Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999;8(1):3–15.
[17] van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007;16(3):219–42.
[18] van Buuren S, Groothuis-Oudshoorn K. MICE: multivariate imputation by chained equations in R. J Stat Softw 2011;45(3).
[19] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.
[20] Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 1996;49(11):1225–31.
[21] Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. Cancer 2001;91(8 Suppl):1636–42.
[22] Dreiseitl S, Ohno-Machado L, Kittler H, Vinterbo S, Billhardt H, Binder M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. J Biomed Inform 2001;34(1):28–36.
[23] McLaren CE, Chen WP, Nie K, Su MY. Prediction of malignant breast lesions from MRI features: a comparison of artificial neural network and logistic regression techniques. Acad Radiol 2009;16(7):842–51.
[24] Al Housseini A, Newman T, Cox A, Devoe LD. Prediction of risk for cesarean delivery in term nulliparas: a comparison of neural network and multiple logistic regression models. Am J Obstet Gynecol 2009;201(1). 113(e1–6).
[25] Raftery AE. Bayesian Model Selection in Social Research. MARSDEN PV, editor: CAMBRIDGE; 1995.
[26] Sauerbrei W, Meier-Hirmer C, Benner A, Royston P. Multivariable regression model building by using fractional polynomials: description of SAS, STATA and R programs. Comput Stat Data Anal 2006;50(12):3464–85.
[27] Hoeting Jennifer A, Madigan DM, Raftery Adrian E, Volinsky Chris T. Bayesian model averaging: a tutorial. Stat Sci 1999;14(4):382–417.
[28] Genell A, Nemes S, Steineck G, Dickman PW. Model selection in medical research: a simulation study comparing bayesian model averaging and stepwise regression. BMC Med Res Methodol 2010;10(1):108.
[29] Prost L, Makowski D, Jeuffroy M-H. Comparison of stepwise selection and Bayesian model averaging for yield gap analysis. Ecol Model 2008;219(1–2):66–76.
[30] Adrian Raftery JH, Chris Volinsky, Ian Painter, Ka Yee Yeung. BMA: Bayesian model averaging. Version 3.15.1; 2012.
[31] Marcus JL, Katz MH, Katz KA, Bernstein KT, Wolf W, Klausner JD. Prediction model to maximize impact of syphilis partner notification – San Francisco, 2004–2008. Sex Transm Dis 2010;37(2):109–14.
[32] Schaffzin JK, Koumans EH, Kahn RH, Markowitz LE. Evaluation of syphilis reactor grids: optimizing impact. Sex Transm Dis 2003;30(9):700–6.
[33] Friedman DJ, Parrish RG, Ross DA. Electronic health records and US public health: current realities and future promise. Am J Public Health 2013;103(9):1560–7.
[34] Shapiro JS. Evaluating public health uses of health information exchange. J Biomed Inform 2007;40(6 Suppl):S46–9.