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