Article

Artificial neural network to predict the effect of obesity on the risk of tuberculosis infection

Alaa Badawi,[1,2] Christina J. Liu,[3] Anas A. Rihem,[4] Alind Gupta[5]

[1]Public Health Risk Sciences Division, Public Health Agency of Canada, Toronto; [2]Department of Nutritional Sciences, Faculty of Medicine, University of Toronto; [3]Department of Pharmacology and Toxicology, Faculty of Medicine, University of Toronto; [4]Ontario Tech University, Oshawa; [5]Cytel Inc., Toronto, Canada

Abstract

Background: Body weight has been implicated as a risk factor for latent tuberculosis infection (LTBI) and the active disease.

Design and Methods: This study aimed to develop artificial neural network (ANN) models for predicting LTBI from body weight and other host-related disease risk factors.

We used datasets from participants of the US-National Health and Nutrition Examination Survey (NHANES; 2012; n=5,156; 514 with LTBI and 4,642 controls) to develop three ANNs employing body mass index (BMI, Network I), BMI and HbA1C (as a proxy for diabetes; Network II) and BMI, HbA1C and education (as a proxy for socioeconomic status; Network III). The models were trained on n=1018 age- and sex-matched subjects equally distributed between the control and LTBI groups. The endpoint was the prediction of LTBI.

Results: When data was adjusted for age, sex, diabetes and level of education, odds ratio (OR) and 95% confidence intervals (CI) for risk of LTBI with increased BMI was 0.85 (95%CI: 0.77 – 0.96, p=0.01). The three ANNs had a predictive accuracy varied from 75 to 80% with sensitivities ranged from 85% to 94% and specificities of approximately 70%. Areas under the receiver operating characteristic curve (AUC) were between 0.82 and 0.87. Optimal ANN performance was noted using BMI as a risk indicator.

Conclusion: Body weight can be employed in developing artificial intelligence-based tool to predict LTBI. This can be useful in precise decision making in clinical and public health practices aiming to curb the burden of tuberculosis, *e.g.*, in the management and monitoring of the tuberculosis prevention programs and to evaluate the impact of healthy weight on tuberculosis risk and burden.

Introduction

Globally, active tuberculosis is a major public health problem and among the leading causes of death from a single infectious pathogen.[1] Although the prevention and treatment of the disease have improved significantly over the past two decades, tuberculosis still responsible for more than one million deaths each year principally in low- and middle-income countries.[1,2] The number of cases with latent tuberculosis infection (LTBI) is over 10 million with approximately 1.5 million deaths – in both HIV-negative and -positive individuals.[1] About 30% of the individuals exposed to *Mycobacterium tuberculosis* infection develop a state of persistent immune response to the pathogen and remain clinically asymptomatic (*i.e.*, LTBI).[3] Only 10% of the latter, however, may progress to active tuberculosis (TB) disease, presenting with clinical signs and symptoms of the disease.[4] Since subjects with LTBI represent a reservoir for active cases; effective prediction, detection, targeted management of this early disease stage were viewed as key components in the World Health Organization's (WHO) "End TB Strategy".[5] This strategy aims at reducing the world rates of TB incidence and mortality by 90% and 95%, respectively, by 2035.[5,6] Although detection of active TB case has been the primary public health response to TB, reducing the LTBI reservoir is viewed as fundamental in reaching the ambitious goal of the "End TB Strategy".[6]

Acute or chronic diminution in body weight was proposed as a risk factor that influence the development of LTBI upon exposure to *M. tuberculosis*.[7-9] For example, recently odds ratio (OR) for tuberculosis was reported to be 4.96 in underweight patients and 0.26 in their obese counterparts.[10] This inverse association between LTBI and body mass index (BMI) was also depicted when a number of studies from Hong Kong, USA, Finland and Norway were collectively and systematically evaluated.[9,11] Furthermore, 3.2-fold increased relative risk of LTBI was noted in people with BMI <18·5 kg/m[12] compared to those with normal

---

Significance for public health

The study aims at developing artificial neural network (ANN) models to predict latent tuberculosis infection (LTBI) from body weight and other host-related disease risk factors in the general population. Three ANNs were developed and trained age- and sex-matched subjects equally distributed between the control and LTBI groups. The predictive accuracy of the three ANNs varied from 75 to 80% with sensitivities ranged from 85% to 94% and specificities of approximately 70%. Areas under the receiver operating characteristic curve (AUC) were between 0.82 and 0.87 with the optimal performance noted using BMI as a risk indicator. Body weight can be employed in developing artificial intelligence-based tool to predict LTBI in the general population and to curb the incidence of active tuberculosis as a major public health problem.

OPEN ACCESS

weight.[9] Irrespective to the nature of interplay between the two conditions and whether body weight causes or is affected by TB, there is a general consensus that higher BMI is linked to lower disease incidence both at population and individual levels.[13] This relationship may be substantiated by observations indicating cardiometabolic risk markers (particularly those associated with obesity such as fasting insulin, total cholesterol, LDL-cholesterol, HDL-cholesterol, and fasting triglycerides) can triple the risk of both LTBI and TB.[14-16] In contrast, some evidence has emerged demonstrating that individuals who are overweight, in close contact with active TB patients and over 50 years of age are all at higher risk of LTBI compared to their counterparts of normal weight.[17] Furthermore, in a population-based study from rural China, overweight and obese subjects were shown to have higher rates of LTBI positivity compared to individuals with normal body weight.[18]

Recently, digital technologies were proposed as effective and efficient tools that can be integrated into the global efforts against TB through aiding in surveillance, patient care, program management and e-learning.[19] The performance and efficiency of digital technology, however, can be augmented by incorporating the innovative approach of artificial intelligence;[20] a branch of computer science concerned with the automation of intelligent behavior.[21] A number of applications were developed in medicine utilizing the framework of artificial neural network (ANN; an artificial intelligence function that imitates the human brain in processing unstructured data) to create patterns that can be used in clinical decision making.[22] For example, artificial intelligence is already embedded in many computer-aided diagnostic platforms and in generating data-driven risk prediction approaches that can be straightforwardly deployed into clinical practices. In this respect, a deep learning algorithm was recently developed and validated to predict and classify clinical abnormalities and pneumonia from chest radiographs at a performance level comparable to practicing radiologists[23] and in the predictions of 30-day unplanned readmission, length of hospitalization and final discharge diagnoses and/or mortality.[24] Within the context of tuberculosis, a number of ANNs were recently introduced for cases of LTBI to predict the risk of developing TB utilizing data on factors such as age, gender, HIV-status, TB history,[25] smoking status, and blood count[26] as well as for classification of the active disease.[27] To our knowledge no study was undertaken to predict acquiring LTBI in healthy subjects assuming exposure to *M. tuberculosis*. Since the association between obesity and tuberculosis suggests a utility for BMI (and the related cardiometabolic risk markers) in identifying subjects at risk of LTBI upon infection[10], the objective of the present study was to examine the relationship between the two conditions and to use ANNs in evaluating the value of BMI in LTBI prediction in a population-based setting.

Design and Methods

Study population

Data were collected from the US National Health and Nutrition Examination Survey (NHANES), a cross-sectional survey of the noninstitutionalized civilian US resident population. The survey is designed to collect information on the health and wellness as well as nutrition status of the populations. It is conducted by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC) and examines a nationally representative sample of approximately 5,000 individuals of all age

groups each year from all counties across the USA. Study methods were all approved by the NCHS research ethics review board and informed consent was obtained from all the study participants.[28] The study subjects were selected by using a complex multistage sampling design.[28] This survey includes an in-home health interview and a physical examination in a mobile examination centre (MEC) in addition to a follow-up telephone interview. The present study includes data from the 2011/2012 cycles of NHANES, a cycle that includes QuantiFERON®-TB Gold In Tube (QFT-GIT) to measure LTBI.[29] Detailed methods of the NHANES survey construction and sampling strategy have been previously described.[30,31] The cycles are stratified, multistage, probability random-sample designed to represent the noninstitutionalized house-dwelling US civilian population. In this analysis, eligible participants were adults (>18 years) who completed the interview and health examination and had valid QFTGIT (positive/negative) and weight and height results. The total number of participants included in the present study was 5,156 subjects (male:female ratio of 1:1.06). The study participants were further divided to controls (n = 4,642) and LTBI (n = 514) subgroups.

Study measures, metabolic markers, sociodemographic factors and other covariates

As an indicator of obesity, BMI (kg/m²) was assessed as previously described.[31] International classification of adult underweight, normal weight, overweight and obesity statuses according to BMI were £18.50, 18.50 – 24.99, 25.00 – 29.99 and ≥30.00 kg/m², respectively as defined by the World Health Organization (WHO).[12] Assessment for the status of LTBI was carried out by QFT-GIT, analyzed according to manufacturer instructions [QuantiFERON®-TB Gold (QFT®) ELISA; QIAGEN, Germantown, MD, USA - www.quantiferon.com]. Results were interpreted according to guidelines from the CDC for using interferon-gamma release assays (IGRAs).[32] Participants with positive QFT-GIT results were classified as LTBI positive whereas participants with negative QFT-GIT results were classified as LTBI negative (controls). Individuals with indeterminate QFTGIT results and those who self-reported they had ever been told by a health care professional to have active TB were excluded. Samples for QFT-GIT testing were processed at a Clinical Laboratory Improvement Act-certified laboratory as previously described.[33]

A number of metabolic markers were measured including cardiometabolic risk markers (apolipoprotein [Apo] B1 [g/L], LDL-C [mmol/L], HDL-C [mmol/L], T-Chol [mmol/L], T-Chol:HDL-C ratio, triglycerides [mmol/L], and HbA1c [%]); and systolic and diastolic blood pressure (mmHg).[30,31] Diabetes status was defined as self-reported or HbA1c ≥6.5%.[34] Individuals who have already been diagnosed as hypertensive, diabetic, or those who were using antihypertensive drugs were included.[35-37] Insulin resistance was approximated using the homeostatic model assessment (HOMA-IR) formula [(glucose (mmol/L) x insulin (μIU/mL))÷22.5].[38,39] Sociodemographic information was captured through responses to questionnaires given during the structured interview portion of the survey and included age, gender, ethnicity, education, history of injection drug use, and ratio of family income to poverty. Ethnicity was categorized into 4 main subgroups: White, African Americans, Asian (*i.e.*, Korean, Filipino, Japanese, Chinese, South Asian, Southeast Asian, Arab, and West Asian), and Other (*i.e.*, Latin American or mixed racial origins). Ratio of family income to poverty was assessed as determined by the Department of Health and Human Services to be used as a measure of poverty.[40] Self-reported smoking status was categorized into smokers (daily/occasional) and non-smokers.[41]

Approach to the artificial neural networks

Data processing

Before training the neural network architecture, datasets were extracted to contain either BMI, gender and age alone (basic network, Network I) or in addition to HbA1C (as a proxy for diabetes, Network II) and further addition of education (as a proxy for socioeconomic status, Network III). These variables were selected based on the odds ratio (OR) for LTBI with increased BMI (see below). Third, since the ratio of control to LTBI cases was >9:1, we applied age- and sex-matching procedure within each of the 3 networks to balance class distribution for the outcome where each subject in the LTBI group was matched with a counterpart from the control group that has a matching sex and age (within ±5 years). Although the latter matching criterion reduced the number of samples available for training to $n=1018$, it ensured class balance known to improve model performance.[22]

Neural network architecture, hyperparameters and model training

The analysis was done with the Python Programming Language (v.3.7.6; 2019, Python Software Foundation)[42] using the TensorFlow backend.[43] As shown in Figure 1, the feed-forward ANN consists of an input layer, a series of hidden layers with nodes (neurons), and an output layer. Layers are interconnected through feed-forward links between nodes (see below).

The hyperparameters tuned for the ANN models were developed based on the Keras hyperparameter optimization framework.[43] ANN training was undertaken in two consecutive stages: i) a forward pass; and ii) a backward pass as previously described.[46] Each forward pass calculates a weighted sum of inputs $x_i, i=0,\ldots,n$ into a node (y) and passed through an activation function σ as follows:

$$y = \sigma \left(\sum_{i=0}^{n} W_i \, x_i \right),$$

where W_0 represents the bias term. The activation function was chosen to be the leaky rectified linear unit (LReLU) to prevent zero gradients during backpropagation, where $ReLU(x)=\max(a \cdot x, x), 0 < \alpha \leq 1$ where α was chosen to be 0.3 (the default value in Keras). Batch normalization was used prior to application of activation function. The ANN weights were initialized using He uniform variance initialization ("he_uniform" kernel initializer in Keras).[47] This approach accelerates the gradient descent process of learning for a neural network that uses Rectified Linear Unit (ReLU) activations.

Following the activation function phase, the entire process is repeated with the values from one layer acting as the input for the next layer until the output layer is reached (Figure 1). A sigmoid activation function was used in the final layer of the network to scale predicted values between 0 and 1, *i.e.*, as class probabilities. Based on the binary classification nature of the prediction sought, the binary cross-entropy loss function or sigmoid cross-entropy loss (L) was used to calculate the average of class errors as follows:

$$L(y,\hat{y}) = -\frac{1}{N} \sum_{i=0}^{N} (y_i * \log(\hat{y}_i) + (1 - y_i) * \log(1 - \hat{y}_i))$$

where N is the number of samples, y_i is the true label for instance $i=1,\ldots,N$ and is the predicted probability from the ANN. The backward pass through the ANN involves updating network weights using backpropagation based on the discrepancy between y_i and $\hat{y}_i$ and using the derivative of $L(y, \hat{y})$ with respect to network

parameters W used to calculate $\hat{y}_i$. Hyperparameters tuned for each of the three developed Networks are shown in Supplement Table 1.

Model evaluation

To evaluate the performance of our three Networks, we used the -fold cross validation procedure with $K=10$,[52] *i.e.*, our protocol separates a dataset into 10 groups or folds. Given N folds, there is a corresponding number of validation sessions (n=10). For each session, one of the folds is held out for testing, while the remaining folds are used for training, until the entire dataset has been used. Receiver operating characteristic (ROC) curves and the areas under the curves (AUC) were then generated from the false positive and true positive rates calculated for each of the Networks as previously described.[53] For performance comparison, a series of baseline benchmark machine learning methods were evaluated for the three ANN models. The baseline machine learning methods included: i) random decision forests (RDF) method,[54] ii) support-vector machines (SVMs) using the sigmoid sklearn,[55] and iii) logistic regression (LG) method[56] to model binary dependent variables (using the default cut-off value of 0.5). The performance of the baseline evaluation methods (*i.e.*, specificity, sensitivity and AUC) is shown in Supplementary Table 2 and Supplementary Figure 1.

Statistical analysis

All analyses excluded survey weights and were stratified by LTBI status. Frequency distributions and means (±standard deviation, SD) were used to describe baseline characteristics. Differences between controls and LTB groups for the examined sociodemographic characteristics and levels of biomarkers and cardiometabolic risk factors were determined using *t*-test and χ^2 tests for continuous and categorical variables, respectively. Fisher's exact test was used for categorical data analysis where there was a small sample size. Multivariate logistic regression models were used to estimate adjusted odds ratios (OR) and 95%
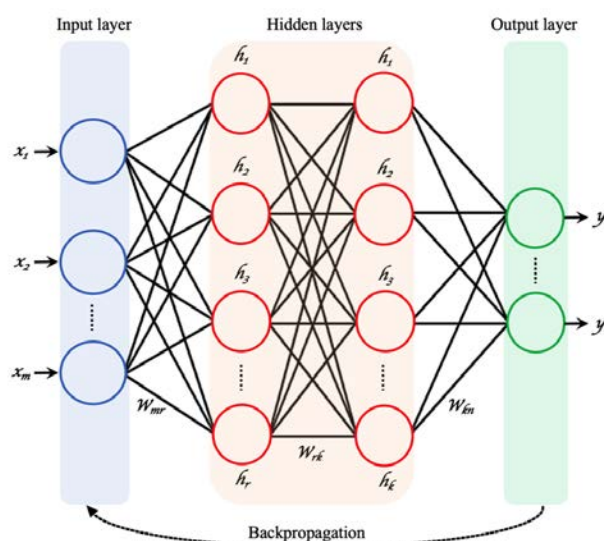

Figure 1. Architecture of multilayer artificial neural network.

Table 1. Characteristics of the study populations.

Characteristic*	Controls (n=4642)	LTBI (n=514)	p#
Males (%)	48.7 (2259)	45.7 (236)	
Age (years)	46.1±18.5 (4642)	55.8±15.7 (514)	<0.001
Ethnicity (%)			
White	39.1 (1816)	14.2 (73)	<0.001
Black	26.2 (1218)	23.8 (122)	
Asian	12.2 (564)	27.8 (143)	
Highest level of education (%)			
Less than grade 12	19.9 (927)	36.5 (190)	<0.001
High-school graduate	19.8 (923)	20.0 (104)	
Post-secondary graduate	54.4 (2530)	41.3 (215)	
Other	5.9 (262)	2.2 (105)	
Ratio of family income to poverty	2.42±1.6 (4282)	2.19±1.5 (454)	0.036
History of intravenous drug use (%)	1.3 (61)	1.7 (9)	
Smoking status (%)			
Daily / Occasional	18.8 (872)	19.6 (102)	0.003
Non-smoker	20.9 (972)	26.5 (138)	
Diabetes (%) - self-reported or HbA1c ≥ 6.5%§	14.1 (653)	24.4 (127)	<0.001
Cardiometabolic risk markers			
Systolic blood pressure (mmHg)	123±18 (4276)	126±19 (477)	<0.001
Diastolic blood pressure (mmHg)	71±13 (4276)	71±13 (477)	
Triglycerides, fasting (mmol/L)	1.42±1.06 (2258)	1.44±0.85 (239)	
Total cholesterol (mmol/L)	4.95±1.08 (4553)	5.04±1.04 (503)	
LDL-C (mmol/L)	2.92±0.91 (2216)	2.93±0.89 (237)	
HDL-C (mmol/L)	1.36±0.38 (4553)	1.32±0.37 (503)	
Total cholesterol:HDL-C ratio	3.87±1.29 (4553)	4.05±1.30 (504)	
Insulin, fasting (pmol/L)	81.6±72.0 (2174)	84.1±69.9 (232)	
Glucose, fasting (mmol/L)	5.93±1.87 (2284)	6.32±2.05 (244)	
HOMA-IR	22.1±26.9 (2172)	31.1±83.9 (234)	0.0423
HbA1c (%)	5.73±1.10 (4630)	6.03±1.23 (514)	
Apolipoprotein B, fasting (g/L)	0.89±0.25 (2259)	0.90±0.23 (239)	
Obesity			
Body mass index (kg/m²)	28.7±7.0 (4642)	28.5±6.7 (515)	
Underweight: BMI: <18.5 kg/m² (%)	2.18 (101)	2.33 (12)	
Normal weight: BMI: 18.5 – 24.9 kg/m² (%)	30.63 (1422)	31.46 (162)	
Overweight: BMI: 25 – 29.9 kg/m² (%)	32.06 (1488)	34.37 (177)	
Obese: BMI: >30 kg/m² (%)	35.13 (1631)	31.66 (163)	

*Percentages are for unweighted frequency: numbers in parentheses represent the number of assessed subjects; #significant difference between control and tuberculosis cases was carried out by χ² test or Student's t-test: Only significant differences are shown; §diabetes is defined as Hb1Ac ≥6.5% or self-reported cases as per survey questionnaire; LTBI, latent tuberculosis infection; LDL-C, low-density lipoprotein; HDL-C, high-density lipoprotein; HOMA-IR, homeostatic model assessment insulin resistance; HbA1c (%), glycosylated haemoglobin.

confidence intervals (CI) between BMI and LTBI and were adjusted for potential confounders. The degree of missing data was assessed for each variable and was considered for multivariable regression model inclusion. If a variable had >80% missing data, it would not be fit to be included in the regression model. All analyses were conducted using SPSS (IBM SPSS Statistics, ver. 21.0. Armonk, NY, USA).

Results

A total of 5,156 respondents were examined in the present study. The prevalence of LTBI in the study population was approximately 10% (n=514). Baseline sociodemographic characteristics and levels of cardiometabolic risk markers of the study population are shown in Table 1, stratified by LTBI status. Individuals with LTBI were, on average, older than their control counterparts (p<0.001). The control group was predominantly Whites (39.1%) whereas in individuals with LTBI, Black and Asian subjects constituted >50% of the group. There was a significantly higher percentage of subjects with less than grade 12 education in the LTBI compared to controls and lower percentage of those had post-secondary education. Also, the ratio of family income to poverty was significantly lower in the LTBI group (p = 0.036). Approximately 1.7-fold significantly higher (p<0.001) prevalence of diabetes was noted in the LTBI group than controls. No significant differences were shown between LTBI and control group in any of the examined cardiometabolic risk markers except for the levels of fasting triglycerides that was slightly but significantly higher in the LTBI than controls (p<0.001).

We categorized BMI into the ranges defined by the WHO with cut-off points of £18.50, 18.50 – 24.99, 25.00 – 29.99 and ≥30.00 kg/m² for underweight, normal weight, overweight and obese, respectively.[12] Multivariate logistic regression models used to estimate OR adjusted for potential confounders (and 95% CI) for

Table 2. Odds for latent tuberculosis infection with increased body mass index.

Model adjustment*	OR (95% CI)	p#
No LTBI	1 (Ref)	
Crude	0.94 (0.84 – 1.04)	
Age	0.89 (0.80 – 0.98)	0.038
Diabetes§ (Hb1Ac ≥ 6.5%)	0.90 (0.82 – 0.97)	0.049
Age and sex	0.89 (0.80 – 0.99)	0.048
Age and diabetes	0.87 (0.79 – 0.97)	0.019
Age, sex and diabetes	0.88 (0.79 – 0.98)	0.026
Age, sex, diabetes and education	0.85 (0.77 – 0.96)	0.010

*Multivariate logistic regression models were used to estimate the adjusted odds ratios and 95% CI between LTBI and increased BMI; #only significant values are shown; §diabetes is defined as Hb1Ac ≥6.5%; LTBI, latent tuberculosis infection; OR, odds ratio; 95% CI, 95% confidence intervals.

LTBI with increased BMI is shown in Table 2. Age, sex, diabetes and level of education were the main confounders in the association between increased BMI and lower LTBI risk. When adjusted for age, sex and diabetes, OR for LTBI was 0.88 (95%CI: 0.79 – 0.98; p=0.026). When this model was further adjusted for the level of education, the OR was decreased to 0.85 (95%CI: 0.77 – 0.96, p=0.01). The addition of smoking, injection drug use and ethnicity to the model, did not affect the odds of LTBI associated with increased BMI (*data not shown*).

Based on the findings from the multivariate logistic regression and OR for LTBI risk with increased BMI (Table 2), three ANNs were generated from age- (±5 years) and sex-matched set of data to employ BMI (Network I), BMI and HbA1C (Network II) and BMI, HbA1C and education (Network III) in the prediction of LTBI. The models were, therefore, trained on n=1018 subjects equally distributed between the control and LTBI groups. To generate the ANN for the prediction of LTBI from obesity and related factors, a different number of hidden units and hidden layers were used for each of the three trained Networks (Table 3). Hidden units and layers were, respectively, 16 and 1 in Network I; 32 and 2 in Network II and 16 and 3 in Network III. We applied this number of layers in calculating the accuracy, sensitivity and specificity of the three Networks. Accuracies (%) varied from 74.3±9.2 in Network III to 80.3±4.7 in Network I. As shown in Table 3, the sensitivities of the three ANNs fell between 85%-94%. Specifically, when BMI alone was considered, the sensitivity of Network I was 94% that declined to 90% upon the inclusion of HbA1C and further to 85% when a socioeconomic factor such as education was additionally included. Similarly, Network specificity varied from 60-70% based on the obesity-related factor included in the ANN. The maximum specificity, while keeping the sensitivity above 90%, was 70% in Network I, *i.e.*, when BMI was considered alone as a predictor.

The ROC curves for predicting LTBI are shown in Figure 2. The ROC curves of the three Networks were all well above the diagonal line, the line of no-discrimination, representing random guessing. The values of the AUC were 0.86, 0.82 and 0.87 for Network I, II and III, respectively. Values of sensitivity, specificity and AUC of the main trained model were comparable to those from the benchmark machine learning models generated from the RDF, SVMs and LG methods (Supplementary Table 2 and Supplementary Figure 1).

Discussion

This is the first evaluation of multiple artificial neural networks for predicting LTBI based on obesity and its related factors. Optimal Network performance was observed when BMI was considered as a LTBI risk predictor with a model accuracy of 80%, sensitivity of 94%, specificity of 70% and AUC of 0.86. These ANN performance indicators were slightly decreased when diabetes and education were further included into the model. Nevertheless, the two ANNs that utilized diabetes and education as

Table 3. Hyperparameters, specificity and sensitivity of the artificial neural network predicting latent tuberculosis from obesity and related factors.

Parameter	Network*		
	I	II	III
Hidden units	16	32	16
Hidden layers	1	2	3
Accuracy ±SD (%)	80.3±4.7	76.3±2.3	74.3±9.2
Sensitivity (%)	94	90	85
Specificity (%)	70	66	60

*All models are age- and sex-matched and include either BMI alone (Model I), BMI and HbA1C (Model II) or BMI, HbA1C and education (Model III).
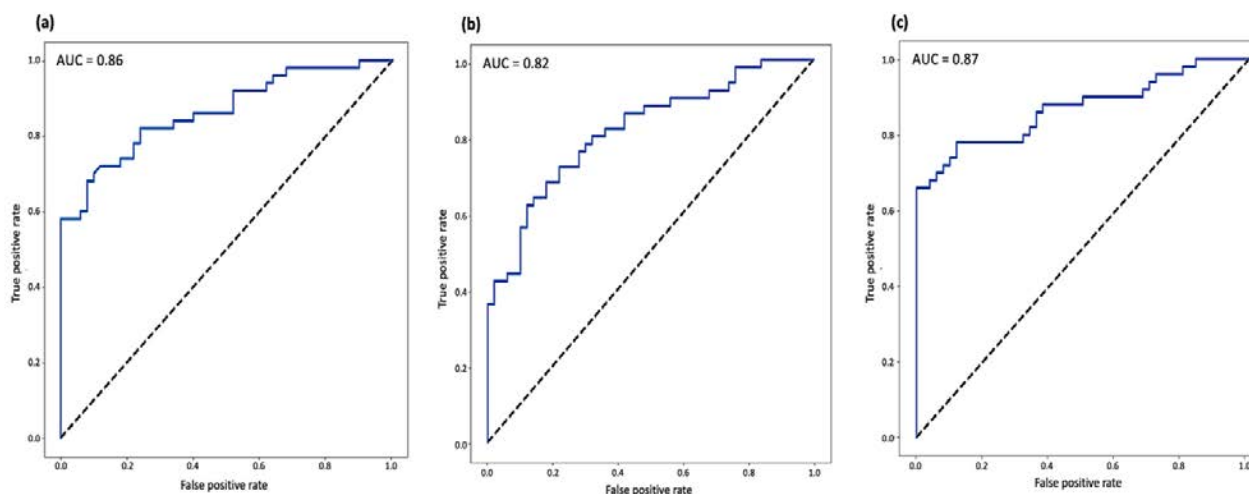

Figure 2. Receiver operating characteristic curves for predicting latent tuberculosis infection. The three trained artificial neural networks were all age- and sex-matched and trained for BMI (a), BMI and HbA1C (b) and BMI, HbA1C and level of education (c). AUC, area under the receiver operating characteristic curve.

LTBI risk predictors still had good performance with accuracies of no less than 74%, sensitivity 85%, specificity 60% and AUC 0.82. The performances obtained from the three models demonstrate a distinct ability of ANNs that include host-related factors (*e.g.*, age, sex, BMI, diabetes, and socioeconomic status) in capturing non-linear interrelationships between the input predictive elements[18,19] as with the RDF and SVM models using a non-linear kernel.[54,55]

In this study, *a priori* statistical evaluation guided our design of the ANN algorithms. The main features in our model training to include HbA1C (as a proxy for diabetes) and education (as a proxy for socioeconomic status) together with applying protocols for age- and sex-matching were all principally based on findings generated from our estimates of ORs for LTBI with increased body weight. It is apparent, therefore, that an initial statistical evaluation of the data can be used to guide the design of ANN algorithms regarding factor inclusion for an effective model planning and engineering.[57] Furthermore, disregarding the effect of sex and gender in ANNs architecture would have generated sub-optimal results, inaccurate predictions, and biased outcomes. The performance of the ANNs observed in the present study validates this assumption and supports the effectiveness of age- and sex- matching protocols in avoiding biases in artificial intelligence used in biomedicine and healthcare, particularly when a small set of data is available for training, *e.g.*, in personalized medicine.[58] It is known that factors such as ethnic origin, marital status, age, history of TB contact, urban residency, socioeconomic status, and metabolic syndrome-related conditions, *e.g.*, obesity and diabetes are all related to LTBI risk and the later development of active disease.[59-63] These genetic, sociodemographic and environmental risk factors exhibit a large inter-individual variation in the general population. Accounting for inter-individual differences of these factors in artificial intelligence, *e.g.*, via stratification and data matching, can both circumvent model biases and facilitate the progress towards individually tailored predictive and preventative measures as well as personalized therapeutic choices.[58]

This study evaluates the utility of factors such as age, sex obesity, diabetes and socioeconomic status in the prediction of LTBI upon exposure to the disease pathogen. Our recent findings[10] and those of others[7-9,64-66] indicate an inverse relationship between BMI and incidence of tuberculosis. When this relationship was adjusted for age, sex, diabetes and education, an overall 15% reduction in LTBI risk with increased BMI observed (OR = 0.85; 95%CI: 0.77 – 0.96, p=0.01), indicating that these factors may influence the relationship between body weight and the risk of LTBI. The present study demonstrates that the inclusion of these factors into ANNs may be valuable in predicting the susceptibility to LTBI. Although more data would have improved the predictive outcome of the ANNs and permitted the inclusion of other disease risk factors,[58,67] our results may shed light on some of the mechanisms mediating the inverse relationship between obesity and tuberculosis. There is a well-characterized interrelationship between levels of cardiometabolic risk markers (*e.g.*, insulin resistance and HbA1C and levels of triglycerides) and increased body weight or adiposity.[68] Adipocytes and the immune cells within the adipose tissue secrete elevated levels of inflammatory mediators[69] that affect both innate and adaptive immune responses and subsequently the immune system capacity to combat tuberculosis and other infectious diseases.[8.70] Increased synthesis of pro-inflammatory cytokines correlates positively with increased body weight both in normal individuals[68] and in patients with tuberculosis[71] to mediate the impact of obesity on response to infection, *e.g.*, with *M. tuberculosis*.[70] Persons with a prolonged persistence to tuberculosis (*i.e.*, LTBI) have elevated insulin resistance and impaired fasting glucose,[72] a profile known to also emerge as BMI increas-

es.[68] This complex interrelationship between obesity, cardiometabolic factors (such as HbA1C, a diabetes risk marker), age and LTBI was noted in a cohort of elderly individuals (>65 years old) who were followed up for 5 years and reported 10% decreased hazard ratio for tuberculosis per unit increase in BMI.[64] In addition to age, obesity and diabetes, we used education level – within an ANN – as a surrogate for socioeconomic status[73] to predict risk of LTBI. Differences in education performance and education level are well-known to be associated with significant differences in socioeconomic status.[73] In support, the ratio of family income to poverty in the population examined here was significantly lower in the LTBI group as was the corresponding level of education. Several studies have linked less education together with low income, crowding and high unemployment to increased rates of tuberculosis.[74-76] Although the inclusion of this socioeconomic factor and diabetes did not improve the LTBI predictive performance of the ANN compared to that obtained when only BMI was included, an overall difference of <10% was observed between the two approaches. The small dataset and the large inter-individual variation may have influenced the lack of significant improvement in ANN performance upon the inclusion of additional LTBI risk factors. Indeed, size and disparity of the trained data are two critical factors known to markedly influence artificial intelligence models for data exploration, learning and accurate predictions.[77] With large datasets, it may be possible to include a tailored set of risk factors – depending on the characteristics of screened individual – to facilitate generating more accurate outcomes of personalized prediction.[58]

A critical component in achieving the goal of the WHO's "End TB Strategy" is the prediction and detection of early disease stages; *i.e.*, LTBI.[5,19] The present study demonstrates the utility of ANN-based applications in predicting LTBI risk upon exposure to the pathogen – taking into consideration a more personalized rather than the one-size-fits-all approach.[78] In achieving this, the engineered ANNs maintained a high sensitivity while accounting for the large interindividual variation of prognostic host-related factors known to present in the general population. Artificial intelligence-based applications were introduced to support efforts against tuberculosis at several levels such as patient care (*e.g.*, adhesion to medication), surveillance (*e.g.*, recording and tracking patient information electronically), program management (*e.g.*, handling of diagnostic data) and e-learning (*e.g.*, customizing the approaches of knowledge acquisition).[20] Additionally, artificial intelligence can contribute to basic research. For example, the present study further establishes the inverse relationship between increased body weight and incidence of LTBI and substantiates the feasibility of employing this host-related factor in disease risk prediction.[8-10] In this context, the output of the ANN models developed here, *i.e.*, risk prediction, is one of the main outputs expected from artificial intelligence that also include diagnosis of medical conditions or recommendations of treatments.[58]

The present report has several limitations. The studied population has 1:9 ratio of LTBI to control. The small number of subjects in the LTBI may have led to a high level of heterogeneity and large inter-individual variation. This may have also resulted in the lack of expected significant difference between controls and LTBI groups in a number of the risk factors assessed here, *e.g.*, history of intravenous drug use and a number of cardiometabolic factors. Furthermore, the small dataset employed in the engineering of the ANNs may have introduced some level of imprecision to the Networks' outcome as artificial intelligence machine learning methods utilize inputs of large data sets to have the capacity to learn the desired function and improve model accuracy for a precision outcome.[79] To overcome this limitation, the techniques used

here in the model training phase were structured to work with a small set of data. We used a relatively shallow network and regularization using dropout to minimize the number of effective parameters in the model, considering the small sample size available for training,[80] hence enhancing the existing ANN architecture. Another limitation is that we did not consider the interaction of LTBI and obesity with cardiometabolic risk markers other than HbA1C (proxy for diabetes) despite the well-established relationships between the increased body weight and other metabolic syndrome risk markers and related chronic diseases.[81] Using a surrogate such as HbA1C for diabetes may have introduced measurement bias (that occurs when measured data are often proxies for some ideal features) into our ANN model.[58,82] Algorithmic biases may have also been introduced into our ANN models when education was used as a proxy for socioeconomic status without correcting for ethnicity and/or inequalities in health access.[83] However, introducing stratification of data by age and sex avoided a key bias known to occur in artificial intelligence applications used in biomedicine when a small dataset is available for training.[58] Lastly, the inverse relationship between obesity and LTBI observed here merely reflects an association between the two conditions and does not substantiate an inference to the causality. We did not explore underlying mechanisms for the effect of body weight on the LTBI risk such as the role and effect of malnutrition,[8,84] adiposity,[70] synthesis of pro-inflammatory cytokines[85] or plasma leptin.[86] These factors were all proposed to influence the relationship between LTBI and body weight in human populations.

In conclusion, the present study further underlines the role of body weight as a risk factor in LTBI where underweight is linked to increased disease risk. Factors that may influence this relationship include age, sex, diabetes and socioeconomic status. Utilizing these disease risk factors, we developed a new risk prediction artificial intelligence-based tool for LTBI using ANN approach. The models predicted the risk of LTBI with high accuracy and sensitivity and permitted further demonstration of the complex interrelationship between body weight and disease risk. This study indicated the feasibility and effectiveness of the artificial intelligence algorithm models as useful tools for precise decision making in clinical and public health practices aiming to curb the burden of tuberculosis. This tool can be introduced, for example in the management and monitoring of tuberculosis prevention or eradication programs by evaluating the potential impact of improving nutritional status on tuberculosis risk in a given population. In this context, eradicating malnutrition and mitigating the effects of other disease risk factors were estimated to further lower the global tuberculosis incidence in 2035 by 33% than the present rate of decline.[87] Our results, however, warrant developing further studies aiming at improving the performance of the artificial neural network models engineered here either via analyzing and training larger datasets, using ideal features and labels rather than proxies and correcting for existing inequalities in healthcare access.

Correspondence: Alaa Badawi, Ph.D., Public Health Risk Sciences Division, Public Health Agency of Canada, 180 Queen Street West, Toronto, ON M5V 3L7; Department of Nutritional Sciences, Faculty of Medicine, University of Toronto, Medical Sciences Building, 1 King's College Circle, Toronto, ON, M5S 1A8, Canada. E-mail: alaa.badawi@canada.ca - alaa.badawi@utoronto.ca

References

1. GBD Tuberculosis Collaborators. Global, regional, and national burden of tuberculosis, 1990-2016: results from the Global Burden of Diseases, Injuries, and Risk Factors 2016 Study. Lancet Infect Dis 2018;18:1329-49.

2. GBD Tuberculosis Collaborators. The global burden of tuberculosis: results from the Global Burden of Disease Study 2015. Lancet Infect Dis 2018;18:261–84.

3. Kahwati LC, Feltner C, Halpern M, et al. Screening for latent tuberculosis infection in adults: An evidence review for the U.S. Preventive Services Task Force. Agency for Healthcare Research and Quality (US). Evidence Syntheses, No. 142; 2016. Accessed: 18 July 2020. Available from: https://www.ncbi.nlm.nih.gov/books/NBK385124/

4. Getahun H, Matteelli A, Chaisson RE, et al. Latent Mycobacterium tuberculosis infection. N Engl J Med 2015; 372:2127–35.

5. Uplekar M, Weil D, Lönnroth K, et al. WHO's new End TB Strategy. Lancet 2015;385:1799-801.

6. Dye C, Glaziou P, Floyd K, et al. Prospects for tuberculosis elimination. Annu Rev Public Health 2013;34:271-86.

7. Tverdal A. Body mass index and incidence of tuberculosis. Eur J Respir Dis 1986;69:355-62.

8. Cegielski JP, McMurray DN. The relationship between malnutrition and tuberculosis: Evidence from studies in humans and experimental animals. Int J Tuberc Lung Dis 2014;8:286-98.

9. Lönnroth K, Williams BG, Cegielski P, et al. A consistent log-

linear relationship between tuberculosis incidence and body mass index. Int J Epidemiol 2010;39:149-55.

10. Badawi A, Gregg B, Vasileva D. Systematic analysis for the relationship between obesity and tuberculosis. Pub Health 2020;186:246-56.

11. Roth J, Sahota N, Patel P, et al. Obesity paradox, obesity orthodox, and the metabolic syndrome: An approach to unity. Mol Med 2016; 2:873-85.

12. Nuttall FQ. Body mass index: Obesity, BMI, and health: A critical review. Nutr Today 2015;50:117-28.

13. Critchley JA, Restrepo BI, Ronacher K, et al. Defining a research agenda to address the converging epidemics of tuberculosis and diabetes: Part 1: Epidemiology and clinical management. Chest 2017;152:165-73.

14. Lönnroth K, Roglic G, Harries AD. Improving tuberculosis prevention and care through addressing the global diabetes epidemic: From evidence to policy and practice. Lancet Diabetes Endocrinol 2014;2:730-9.

15. Riza AL, Pearson F, Ugarte-Gil C, et al. Clinical management of concurrent diabetes and tuberculosis and the implications for patient services. Lancet Diabetes Endocrinol 2014;2:740-53.

16. Badawi A, Sayegh S, Sallam M, et al. The global relationship between the prevalence of diabetes mellitus and incidence of tuberculosis: 2000-2012. Glob J Health Sci 2014;7:183-91.

17. Cubilla-Batista I, Ruiz N, Sambrano D, et al. Overweight, obesity, and older age favor latent tuberculosis infection among household contacts in low tuberculosis-incidence settings within Panama. Am J Trop Med Hyg 2019;100:1141-4.

18. Zhang H, Li X, Xin H, et al. Association of body mass index with the tuberculosis infection: A population-based study among 17796 adults in rural China. Sci Rep 2017;7:41933.

19. WHO, European Respiratory Society. Digital health for the End TB Strategy: An agenda for action (WHO/HTM/TB/2015.21). 2015. Accessed: 19 June 2020. Available from: https://www.who.int/tb/areas-of-work/digital-health/Digital_health_EndTBstrategy.pdf

20. Doshi R, Falzon D, Thomas BV, et al. Tuberculosis control, and the where and why of artificial intelligence. ERJ Open Res 2017;3:56.

21. Luger GF. Artificial intelligence: Structures and strategies for complex problem solving. 5th ed. Essex: Pearson Education Ltd.; 2005.

22. Cui S, Tseng HH, Pakela J, et al. Introduction to machine and deep learning for medical physicists. Med Phys 2020;47:e127-47.

23. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018;15:e1002686.

24. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18.

25. Khan MT, Kaushik AC, Ji L, et al. Artificial neural networks for prediction of tuberculosis disease. Front Microbiol 2019;10:395.

26. Er O, Temurtas F, Tanrikulu AC. Tuberculosis disease diagnosis using artificial neural networks. J Med Syst 2010;34:299-302.

27. Lakhani P, Sundaram B. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017;284:574-82.

28. Mohadjer LMJ, Montaquila J, Waksberg J, et al. National Health and Nutrition Examination Survey III: Weighting and examination methodology. Hyattsville, MD; 1996. Available from: https://ceb.nlm.nih.gov/proj/dxpnet/nhanes/docs/doc/nhanes_analysis/wgt_exec.pdf

29. Barron MM, Shaw KM, Bullard KM, et al. Diabetes is associated with increased prevalence of latent tuberculosis infection: Findings from the National Health and Nutrition Examination Survey, 2011-2012. Diabetes Res Clin Pract 2018;139:366-79.

30. Curtin LR, Mohadjer LK, Dohrmann SM, et al. National Health and Nutrition Examination Survey: sample design, 2007-2010. Vital Health Stat2 2013;160:1-23.

31. Johnson CL, Dohrmann SM, Burt VL, Mohadjer LK. National health and nutrition examination survey: sample design, 2011-2014. Vital Health Stat2 2014;162:1-33.

32. Mazurek GH, Jereb J, Vernon A, et al. Updated guidelines for using Interferon Gamma Release Assays to detect Mycobacterium tuberculosis infection - United States, 2010. MMWR Recomm Rep 2010;59:1-25.

33. Miramontes R, Hill AN, Yelk Woodruff RS, et al. Tuberculosis infection in the United States: Prevalence estimates from the National Health and Nutrition Examination Survey, 2011-2012. PLoS One 2015;10:e0140881.

34. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diab Care 2010;33:S62-9.

35. Grundy SM, Cleeman JI, Daniels SR, et al. Diagnosis and management of the metabolic syndrome: An American Heart Association/National Heart, Lung, and Blood Institute scientific statement. Circulation 2005;112:2735-52.

36. Brenner DR, Arora P, Garcia-Bailo B, et al. Plasma vitamin D levels and risk of metabolic syndrome in Canadians. Clin Invest Med 2011;34:E377.

37. Setayeshgar S, Whiting SJ, Vatanparast H. Prevalence of 10-year risk of cardiovascular diseases and associated risks in Canadian adults: The contribution of cardiometabolic risk assessment introduction. Int J Hyper 2013;2013:276564.

38. Matthews DR, Hosker JP, Rudenski AS, et al. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia 1985;28:412-9.

39. Badawi A, Sayegh S, Sadoun E, et al. Relationship between insulin resistance and plasma vitamin D in adults. Diabetes Metab Syndr Obes 2014;7:297-303.

40. CDC. National Health and Nutrition Examination Survey: 2011–2012 data documentation, codebook, and frequencies. Atlanta: National Center for Health Statistics; 2013.

41. Badawi A, Di Giuseppe G, Arora P. Cardiovascular disease risk in patients with hepatitis C infection: Results from two general population health surveys in Canada and the United States (2007-2017). PLoS One 2018;13:e0208839.

42. Raschka S and Mirjalili V. Python Machine Learning, 2nd Edition. Birmingham: Packt Publishing Ltd.; 2017.

43. Chollet F. Xception: Deep learning with depthwise separable convolutions. arXiv 2017. 1610.02357v3.

44. Valueva MV, Nagornov NN, Lyakhov PA, et al. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. Math Comp Simulation 2020;177:232-43.

45. Dupond S. A thorough review on the current advance of neural network structures. Ann Rev Control 2019;14:200-30.

46. Han SH, Kim KW, Kim S, et al. Artificial neural network: Understanding the basic concepts without mathematics. Dement Neurocogn Disord 2018;17:83-9.

47. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classifica-

tion. arXiv 2015;1502.01852.

48. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv 2015;1502.03167.

49. Santurkar S, Tsipras D, Ilyas A, et al. How does batch normalization help optimization? arXiv 2019;1805.11604v5.

50. Nwankpa C, Ijomah W, Gachagan A, et al. Activation functions: Comparison of trends in practice and research for deep learning. arXiv 2018;1811.03378.

51. Goodfellow I, Bengio Y, Courville A. Deep Learning. Chapter 6: Deep feedforward Networks. MIT Press; 2016. pp. 164-223.

52. Riley RD, Ahmed I, Derbay TPA, et al. Summarising and validating test accuracy results across multiple studies for use in clinical practice. Stat Med 2015;34:2081-103.

53. Trevethan R. Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. Front Publ Health 2017;5:307.

54. Denisko D, Hoffman MM. Classification and interaction in random forests. Proc Natl Acad Sci USA 2018; 15:1690-2.

55. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273-97.

56. Tolles J, Meurer WJ. Logistic regression: Relating patient characteristics to outcomes. JAMA 2016;316:533–4.

57. Yu B, Kumbier K. Artificial intelligence and statistics. Front Inf Technol Electronic Eng 2018;19:6-9.

58. Cirillo D, Catuara-Solarz S, Morey C, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit Med 2020;3:81.

59. Zhang X, Jia H, Liu F, et al. Prevalence and risk factors for latent tuberculosis infection among health care workers in China: A cross-sectional study. PLoS One 2013;8:e66412.

60. Sarivalasis A, Zellweger JP, Faouzi M, et al. Factors associated with latent tuberculosis among asylum seekers in Switzerland: a cross-sectional study in Vaud County. BMC Infect Dis 2012;12:285.

61. Lule SA, Mawa PA, Nkurunungi G, et al. Factors associated with tuberculosis infection, and with anti-mycobacterial immune responses, among five-year old BCG-immunised at birth in Entebbe, Uganda. Vaccine 2015;33:796-804.

62. Martínez-Aguilar G, Serrano CJ, Castañeda-Delgado JE, et al. Associated risk factors for latent tuberculosis infection in subjects with diabetes. Arch Med Res 2015;46:221-7.

63. Kizza FN, List J, Nkwata AK, et al. Prevalence of latent tuberculosis infection and associated risk factors in an urban African setting. BMC Infect Dis 2015;15:165.

64. Leung CC, Lam TH, Chan WM, et al. Lower risk of tuberculosis in obesity. Arch Intern Med 2007;167:1297-304.

65. Prospective Studies Collaboration, Whitlock G, Lewington S, et al. Body-mass index and cause-specific mortality in 900,000 adults: collaborative analyses of 57 prospective studies. Lancet 2009;373:1083-96.

66. Pednekar MS, Hakama M, Hebert JR, et al. Association of body mass index with all-cause and cause-specific mortality: Findings from a prospective cohort study in Mumbai (Bombay), India. Int J Epidemiol 2008;37:524-35.

67. Qin ZZ, Sander MS, Rai B et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. Sci Rep 2019;9:15000.

68. Da Costa LA, Arora P, García-Bailo B, et al. The association between obesity, cardiometabolic disease biomarkers, and innate immunity-related inflammation in Canadian adults. Diabetes Metab Syndr Obes 2012;5:347-55.

69. Park HS, Park JY, Yu R. Relationship of obesity and visceral adiposity with serum concentrations of CRP, TNF-alpha and IL-6. Diabetes Res Clin Pract 2005;69:29-35.

70. Karlsson EA, Beck MA. The burden of obesity on infectious disease. Exp Biol Med (Maywood) 2010;235:1412-24.

71. Lamas O, Marti A, Martínez JA. Obesity and immunocompetence. Eur J Clin Nutr 2002;56:S42-5.

72. Philips L, Visser J, Nel D, et al. The association between tuberculosis and the development of insulin resistance in adults with pulmonary tuberculosis in the Western sub-district of the Cape Metropole region, South Africa: A combined cross-sectional, cohort study. BMC Infect Dis 2017;17:570.

73. Thomson S. Achievement at school and socioeconomic background - an educational perspective. npj Science Learn 2018;3:5.

74. Olson NA, Davidow AL, Winston CA, et al. A national study of socioeconomic status and tuberculosis rates by country of birth, United States, 1996-2005. BMC Publ Health 2012;12:365.

75. Cantwell MF, McKenna MT, McCray E, et al. Tuberculosis and race/ethnicity in the United States: Impact of socioeconomic status. Am J Respir Crit Care Med 1998;157:1016-20.

76. Barr RG, Diez-Roux AV, Knirsch CA, et al. Neighborhood poverty and the resurgence of tuberculosis in New York City, 1984–1992. Am J Publ Health 2001;91:1487-93.

77. Oliveira AL. Biotechnology, Big data and artificial intelligence. Biotechnol J 2019;14:1800613.

78. Ginsburg GS, Phillips KA. Precision medicine: from science to value. Health Aff 2018;37:694-701.

79. Fan W, Bifet A. Mining big data: current status, and forecast to the future. ACM SIGKDD Explor Newsl 2014;16:1-5.

80. Caruana R, Lawrence S, Giles L. Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. In: Proc 13th Int Conf Neural Info Process Sys 2000; MIT Press, Cambridge, MA, USA. pp. 381-7.

81. Badawi A, Drebot M, Ogden NH. Convergence of chronic and infectious diseases: a new direction in public health policy. Can J Publ Health 2019;110:523-4.

82. Barajas A, Ochoa S, Obiols JE et al. Gender differences in individuals at high-risk of psychosis: A comprehensive literature review. Sci World J 2015;2015:430735.

83. Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447-53.

84. James WP, Ferro-Luzzi A, Waterlow JC. Definition of chronic energy deficiency in adults. Report of a working party of the International Dietary Energy Consultative Group. Eur J Clin Nutr 1988;42:969-81.

85. Moreira-Teixeira L, Mayer-Barber K, Sher A, et al. Type I interferons in tuberculosis: Foe and occasionally friend. J Exp Med 2018;215:1273-85.

86. Procaccini C, Lourenco EV, Matarese G, et al. Leptin signaling: A key pathway in immune responses. Curr Signal Transduct Ther 2009;4:22-30.

87. Odone A, Houben RM, White RG, et al. The effect of diabetes and undernutrition trends on reaching 2035 global tuberculosis targets. Lancet Diabet Endocrinol 2014;2:754-64.