International Journal of Epidemiology, 2018, 1645–1657
doi: 10.1093/ije/dyy174
Advance Access Publication Date: 16 August 2018
Original article

Methods

Development and validation of a predictive ecological model for TB prevalence

Sandra Alba,[1]* Ente Rood,[1] Mirjam I Bakker,[1] Masja Straetemans,[1] Philippe Glaziou[2] and Charalampos Sismanidis[2]

[1]KIT Health, KIT Royal Tropical Institute, Amsterdam, The Netherlands and [2]Global TB Programme, World Health Organization, Geneva, Switzerland

*Corresponding author. KIT Health, KIT Royal Tropical Institute, Mauritskade 63, 1092 AD Amsterdam, The Netherlands. E-mail: s.alba@kit.nl

Editorial decision 18 July 2018; Accepted 6 August 2018

Abstract

Background: Nationally representative tuberculosis (TB) prevalence surveys provide invaluable empirical measurements of TB burden but are a massive and complex undertaking. Therefore, methods that capitalize on data from these surveys are both attractive and imperative. The aim of this study was to use existing TB prevalence estimates to develop and validate an ecological predictive statistical model to indirectly estimate TB prevalence in low- and middle-income countries without survey data.

Methods: We included national and subnational estimates from 30 nationally representative surveys and 2 district-level surveys in India, resulting in 50 data points for model development (training set). Ecological predictors included TB notification and programmatic data, co-morbidities and socio-environmental factors extracted from online data repositories. A random-effects multivariable binomial regression model was developed using the training set and was used to predict bacteriologically confirmed TB prevalence in 63 low- and middle-income countries across Africa and Asia in 2015.

Results: Out of the 111 ecological predictors considered, 14 were retained for model building (due to incompleteness or collinearity). The final model retained for predictions included five predictors: continent, percentage retreated cases out of all notified, all forms TB notification rates per 100 000 population, population density and proportion of the population under the age of 15. Cross-fold validations in the training set showed very good average fit (R-sq = 0.92).

Conclusion: Predictive ecological modelling is a useful complementary approach to indirectly estimating TB burden and can be considered alongside other methods in countries with limited robust empirical measurements of TB among the general population.

Key words: predictive ecological modelling, TB prevalence, modelling, predictions, TB prevalence surveys, subnational estimates

© The Author(s) 2018. Published by Oxford University Press on behalf of the International Epidemiological Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/ije/article/47/5/1645/5075187 by guest on 24 April 2024

Key Messages

- Population-based surveys are the gold standard to estimate TB prevalence. These are massive and complex undertakings, so it is important to make the most out data from these surveys.
- One possible application is to use national and subnational TB prevalence estimates to build a predictive ecological model to predict TB prevalence in countries without survey estimates. This is especially useful for countries with a high burden of TB and low case-detection rates, where TB notification data cannot be used directly as a measure of TB burden.
- We compiled a database including all existing TB prevalence survey estimates and ecological predictors for 30 countries and fitted a random-effects multivariable binomial regression model to predict TB prevalence in 63 other low- and middle-income countries without survey data and with an estimated prevalence over 0.1% according to WHO estimates.
- We were able to develop a predictive ecological model for TB prevalence with reasonable internal and external validity. We therefore concluded that this method can provide useful complementary estimates for TB prevalence and can be considered alongside other methods in countries with limited TB data.

Background

Tuberculosis (TB) is a major global health problem and a leading cause of death worldwide alongside the human immunodeficiency virus (HIV). According to the latest estimates, in 2016, 10.4 million people fell ill with TB and 1.3 million people succumbed to the disease.[1] The Sustainable Development Goals (SDGs) for 2030 reflect the scale of the epidemic and its importance as a global priority—one of the health targets (Goal 3) is to end the TB epidemic worldwide.[2] More specifically, the WHO End TB Strategy calls for a 95% reduction in TB deaths and a 90% reduction in the TB incidence rate by 2035 compared with 2015.[3]

Routine and reliable data to monitor time trends in TB disease burden are indispensable to ensure that the right strategies are put in place to achieve these goals and to monitor and evaluate progress towards targets. Although the data available to estimate TB disease burden improved considerably during the millenium development goals era, some data gaps remain, especially in countries with low levels of access to care, weak surveillance and no vital registration systems. The most readily accessible routine data informing TB burden estimation are surveillance data on TB case notifications, compiled annually by all national TB control programmes. Whilst the wealth of data produced by TB control programmes can and should be used at national and subnational levels for planning purposes, they do not lend themselves well to cross-country comparisons. Indeed, the subnationally disaggregated notification data are essential to support rational resource allocation and to ensure that right mix of interventions are put in place. However, since levels of under-reporting and under-diagnosis differ from country to country (and are mostly unknown), notification data are usually not robust or stable enough over time to monitor global trends towards the elimination of the TB epidemic. Most notably, increases in the share of services provided by the private sector can have a great impact on notifications (e.g. if data are no longer notified to the national TB control programmes) without having any impact on TB burden.

The End TB Strategy relies primarily on two global TB disease burden indicators, namely the TB incidence rate and the absolute number of TB deaths.[3] Whilst TB prevalence is no longer a global indicator per se, prevalence surveys remain an invaluable empirical measurement to inform estimations of incidence and, in some cases, mortality.[4] Direct measurements of TB incidence require that TB notifications are reliable proxies, whereas direct measurements of TB mortality require fully functioning vital registration systems. Where that is not the case, estimates of TB prevalence can help to estimate the level of under-reporting and under-diagnosis of detected TB cases and guide adjustments to estimate TB incidence. In turn, mortality can be derived indirectly from incidence and case fatality ratios.

The gold standard for the estimation of TB prevalence consists of nationally representative population-based surveys, as they are the only methodology that can provide precise and unbiased estimates of TB prevalence among surveyed populations.[5] TB prevalence surveys are a massive and complex undertaking, with serious demands on available in-country technical resources and financial implications.[6] Therefore, methods and applications that capitalize on data from these surveys to strengthen global monitoring and evaluation efforts are not only attractive, but also imperative. Examples of such methods include mathematical or statistical predictive models to estimate TB in non-surveyed locations or to make future forecasts. Country-level predictions

Downloaded from https://academic.oup.com/ije/article/47/5/1645/5075187 by guest on 24 April 2024

of TB have traditionally relied on forecasting from time series of notification data,[7–16] with the caveat that these data tend to reflect access to services rather than disease burden, which may not be the same, especially in countries with low case-detection rates. More recently, a number of TB burden prediction models have been developed, which aim to circumvent this issue and rely, amongst other data, on data from TB prevalence surveys as input data. These include both mathematical models (deterministic compartmental models or individual-based stochastic)[17] and Bayesian meta-regression models,[18] which can simultaneously estimate TB mortality, incidence and prevalence.

Predictive ecological modelling using TB prevalence as input data represents one other possible avenue to predict TB burden and to make maximum use of existing TB prevalence surveys. Whereas ecological models are very commonly used in epidemiology, they are usually descriptive and explanatory and rarely predictive. An ecological predictive model for TB prevalence offers the possibility to predict prevalence by making use of existing survey data in combination with both TB and non-TB-related information at national and subnational (when possible) levels. TB burden estimates that are not purely dependent on TB data are attractive, as they are less vulnerable to issues of data completeness and bias, which can permeate all TB data in a given country.

The purpose of this study was therefore to explore the feasibility and reliability of predictive ecological modelling to predict prevalence in low- and middle-income countries without national representative TB surveys for countries with an estimated prevalence over 0.1% according to WHO estimates.The relationship between national and (where possible) subnational TB prevalence levels vs TB notification and programmatic data, co-morbidities and socio-environmental factors—in countries where TB prevalence surveys were conducted—was used to predict prevalence in countries where no prevalence surveys were conducted.

Methods

Database compilation

The first step in database compilation was the definition of countries as part of the training set (i.e. whose data are is used to define the predictive equation) and countries for which prevalence was to be estimated.[19] The complete training set initially included all countries where prevalence surveys have been conducted between 1990 and 2015. This included national estimates for 22 countries in which nationally representative surveys were conducted, subnational estimates of an additional eight nationally representative surveys and three district-level surveys in India (Table 1, Figure 1). A thorough review of survey methodologies and results led to the exclusion of three surveys from the initial training set (non-comparable survey methodology or presentation of results) and an additional survey was excluded because no predictor data could be obtained for that country and year. As a result, the final number of data points (survey estimates) available for analyses in the training set was reduced from 54 to 50 (Table 1). Predictions were made for all low- and middle-income countries in Africa and Asia with predicted prevalence of over 0.1% (according to WHO estimates) where national surveys have not been implemented—a total of 63 African and Asian countries. The number of participants per survey is show in Supplementary File 1, available as Supplementary data at *IJE* online.

A conceptual framework was developed for the model based on selected publications on drivers and determinants of TB.[20–27] Four categories of predictors were identified: TB notification data, TB programmatic determinants, co-morbidities and socio-environmental factors. TB notification data include all forms and laboratory-confirmed notified cases of TB, as well as percentage of multi-drug-resistant, percentage retreated and treatment success rate. TB programmatic determinants are health system determinants representing a country's capacity to find and effectively cure all TB cases. Co-morbidities are those that are known to be associated with TB (including poor nutritional status as a broader indicator of impaired health resilience). Socio-environmental factors encompass a broad range of factors that either increase the risk of exposure to TB infection or are linked to impaired host defense against infection.

Whereas our framework represents a theoretical construct to capture a range of potential predictors of TB prevalence, its operationalization was limited to the variables available in openly accessible databases. Predictor variables were matched to prevalence estimates if they were available for the year of the survey. Data at national-level sources of data included: the WHO global TB data collection system,[28] the WHO Global Health Observatory data repository,[29] the World Data Bank,[30] the WHO-UNICEF vaccination coverage estimates,[31] the WorldClim database (1-km spatial resolution climate surfaces for global land areas),[32] as well as source datasets provided by UNAIDS[33] and the International Diabetes Federation Atlas.[34] Sources of data for subnational areas included data from national bureaus of statistics (e.g. censuses), Demographic and Health Surveys (DHS) reports[35] and Multiple Indicator Surveys (MICS) reports.[36] TB notification data at the subnational level were obtained from national TB control programmes (NTPs). A total of 111

Downloaded from https://academic.oup.com/ije/article/47/5/1645/5075187 by guest on 24 April 2024

Table 1. Survey estimates available for the TB prevalence model ($N=54$) and included in the model ($N=50$)

Level	Africa	Asia
National estimates	2011 Ethiopia	1990 China
	2012 Gambia	1990 Republic of Korea[a]
	2012 Rwanda	1991 Thailand[a]
	2012 Thailand	1994 Myanmar
	2012 Tanzania[b]	1995 Republic of Korea
	2013 Ghana	1997 Philippines
	2013 Malawi	2002 Cambodia
	2013 Sudan	2007 Philippines
	2014 Zambia	2008 Bangladesh[a]
	2014 Zimbabwe	2011 Cambodia
	2015 Uganda	2011 Lao
Subnational estimates from national prevalence surveys[c]	2012 Nigeria (6 areas)	2000 China (3 areas)
		2004 Indonesia (3 areas)[b]
		2007 Vietnam (3 areas)
		2009 Myanmar (2 areas)
		2010 China (3 areas)
		2011 Pakistan (6 areas)
		2014 Indonesia (3 areas)[b]
Subnational surveys (India)		2007 Thiruvallur (Tamil Nadu)[a]
		2009 Jabalpur (Madhya Pradesh)[d]
		2009 Bangalore Rural (Karnataka)

[a] Excluded from training set (too few predictor variables, no confidence interval reported or non-standard survey methodology/implementation).

[b] The Tanzania and Indonesia surveys only reported sputum smear positive cases (SS+), so we estimated the number of bacteriologically confirmed cases based on the ratio between SS+ and bacteriologically confirmed from prevalence surveys conducted in the respective regions (WHO defined Africa and South-East Asia region respectively).

[c] See Supplementary File 3, available as Supplementary data at *IJE* online, for details.

[d] Reported estimates were corrected by multiplying them by 1.7 to account for no x-ray in the survey's screening procedure, as suggested by the authors of the study.

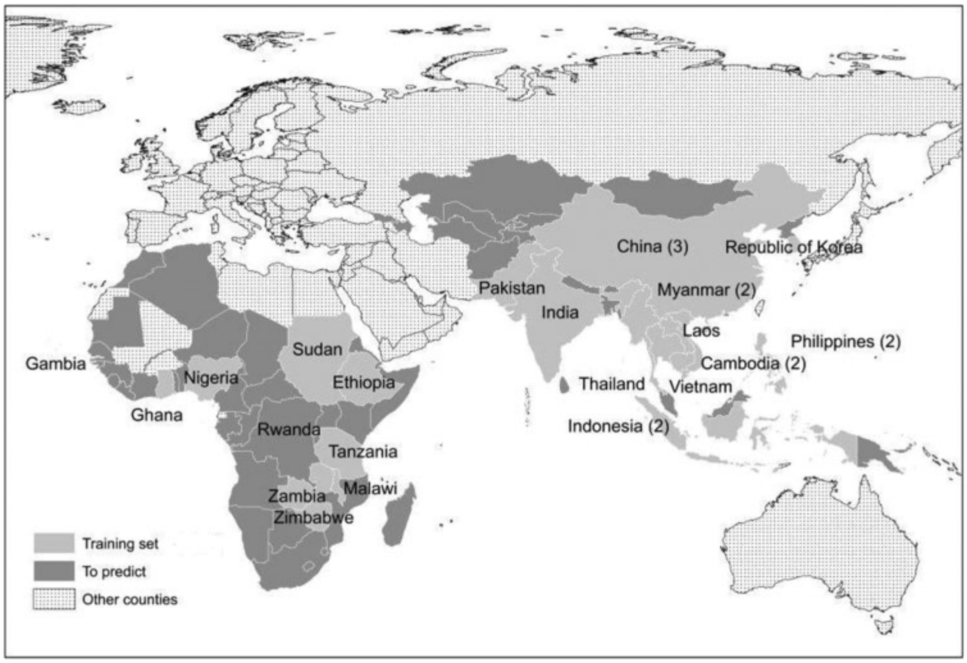


Figure 1. Countries used for TB prevalence prediction (the training set) and countries for which prevalence was predicted.

Downloaded from https://academic.oup.com/ije/article/47/5/1645/5075187 by guest on 24 April 2024

predictor variables were identified, as summarized in Supplementary File 2, available as Supplementary data at *IJE* online. Most indicators were obtained as part of time series covering years between 1990 and 2015, but with many data gaps. When data for a given country were missing for a year, when prevalence survey data were available, we used the 'first observation carried backward' and 'last observation carried forward' method for up to 5 years prior to or after the available data points to improve coverage.

Model definition

Adjusted numbers of survey participants (*N'*) and adjusted numbers of bacteriologically confirmed (C') individuals, which take into account population weighting, clustering, non-participation and missing data, were estimated based on the final reported prevalence estimates (*p*) and the upper and lower limits of their confidence intervals (CIs) assuming a normal symmetrical interval on either side of the prevalence estimate. We assumed that the adjusted number of bacteriologically confirmed cases in a given country *i* and subnational area *j* arose from the binomial distribution

$$C_{ij}^{'} \sim B(N_{ij}^{'},\ p_{ij})$$

and thus fitted the following multilevel multivariable model:

$$\log it(p_{ij}) = \alpha + \beta_1 x1_{ij} + \cdots + \beta_n xn_{ij} + u_i + e_{ij},$$

where *i* denotes the countries for which survey estimates were available and *j* the subnational estimate within country *i*; α is the estimated baseline logit transformed number of cases C' in country *i* and subnational area *j*; the β parameters β_1 to β_n are the estimated regression coefficients for the independent predictor variables $x1$ to xn; u_i denotes a country-specific error term (normally distributed); and e_{ij} denotes a subnational area-specific error term (normally distributed).

It is worth pointing out here that about half of the countries had only national-level estimates and the other half subnational-level estimates for the outcome variable (number of bacteriologically confirmed cases). For the countries for which subnational-level estimates were available, predictors were at times found at the given subnational level, but at times the national-level predictor value had to be used at all subnational levels if it was not available at the subnational level. As a result, the multilevel model we fitted included a mixture of national and subnational levels. The country-specific error term u_i (random effect) ensures that the subnational estimates belonging to the same country are adequately grouped together and dependencies between them modelled out. This random effect also enables to account for the fact that some countries had repeated surveys (e.g. Korea, China and Indonesia).

Model building

In predictive modelling, the aim is to develop a model to predict new or future observations. Model-building strategies for this type of model focus on association rather than causation and criteria for choosing predictors are the availability of the predictors at the time of prediction as well as the strength of the association between the response and the predictors[19] or predictions. The primary consideration in model building was to avoid over fitting—'the biggest danger to generalization',[19] especially relevant in our case given the small sample size available for modelling. The secondary consideration was to reduce the dimensionality of the data for modelling, to minimize multicollinearity and address multiple testing.

Predictor variables were thus selected for inclusion in the multivariable model based on the following procedure. First, predictors were selected based on completeness in the training dataset—only complete variables were considered due to the limited number of prevalence data points available for modelling. Second, the relationship between the predictors and the outcome count data (C') was explored by means of scatter plots to identify potentially non-linear relationships. Logarithmic and squared transformations of the predictors were included in this step based on the visual inspection of the scatter plots. A climatic score was computed by means of principal component analysis based on a country's average yearly temperature and minimum as well as maximum precipitation. Third, complete predictors were univariately fitted to the outcome data and model fit was assessed based on akaike information criterion (AIC) values. Finally, pairwise correlations were calculated and correlated predictors (Pearson's correlation coefficient $\rho > 0.7$) were dropped based on the lowest relative fit to the outcome data.

Two model-building strategies were pursued for the multivariable model and their performance was compared. The first approach (Model 1) was a purely data-driven algorithm to maximize goodness of fit, whereby the final multivariable model was built by backward elimination of variables with the highest *p*-values (Wald test), starting from the full model with all predictors with a *p*-value < 0.05 in univariate analyse (this low threshold was chosen to limit the number of candidate variables given the small sample size). Elimination was conducted until only five predictors were left in the model, to ensure an approximately 1:10 variable:observation ratio, as variously

Downloaded from https://academic.oup.com/ije/article/47/5/1645/5075187 by guest on 24 April 2024

suggested in applied statistics literature to avoid overfitting.[37,38] The second approach (Model 2) was epidemiologically informed and followed a two-step approach. First, a multivariable model was created by introducing the variable 'continent' (Africa vs Asia) as well as all TB-related variables found to be associated in the univariate models with $p < 0.05$ (here, too, this low threshold was chosen to limit the number of candidate variables given the small sample size) and backward elimination was done to discard redundant variables ($p > 0.05$). This choice was made to ensure that this final model could factor in the fact that prevalence in Asia is on average higher than in Africa; and to ensure that TB notification data (which could be expected to be the most predictive variables for TB prevalence in settings with complete and accurate reporting) would have a place in the final model. Only after this first stage were then other more distantly related ecological predictors added one by one from those predictors associated in univariate analyses with $p < 0.05$. Following the 1:10 variable:observation ratio threshold, introduction of variables was conducted until five predictors were left in the model.

The linear models (Model 1 and Model 2) were used to predict the point estimate $\log it(p_{ij})$, and the standard error of the linear prediction was used to compute a 95% CI for $\log it(p_{ij})$. The point estimate as well as the lower and upper levels of the CIs were then back-transformed to produce the final reported estimates of TB prevalence (p_{ij}). Given that predictions were made for countries without surveys, the parameter $N_{ij}^{'}$ was missing for all countries. It was thus set at 50 000 everywhere, corresponding to the median number of participants in the surveys included in the training set (Supplementary File 1, available as Supplementary data at *IJE* online). In other words, we predicted TB prevalence for a hypothetical survey with 50 000 participants in each country.

Internal validation

Validation consisted of evaluating the degree of overfitting, namely 'evaluating the performance of the model not on the training set, i.e. the data used to build the model, but on a holdout sample which the model "did not see"'.[19] A popular approach when data are scarce is cross-validation,[39,40] of which the leave-one-out cross-validation (LOOCV) procedure is an example. For every observation in the estimating sample, LOOCV estimates the model specified with all but the i^{th} observation, fits the model using the remaining N-1 observations and uses the resulting parameters to predict the value of the dependent variable for the i^{th} observation. LOOCV reports a pseudo-R^2 value that is the square of the correlation coefficient of the predicted and observed values of the dependent variable.

External validation

External validation was based on sample predictions made for 2015 for 63 countries and consisted of three steps. First, the coherence and credibility of model predictions were assessed by ascertaining whether the range of predictions (minimum and maximum) was consistent with the training data.

Second, model predictions were compared with WHO 2015 estimates. WHO estimates prevalence for all forms of TB in all ages whereas our model predicted bacteriologically confirmed adults, since they are the input data for the model from prevalence surveys. We converted our model predictions into an estimate of all forms of TB in all ages using the correction factor developed by WHO and applied to their own estimates. The adjustment factor is

$$f = \frac{1 - c + cr}{1 - e},$$

where c is the proportion of the population under the age of 15, r is the prevalence ratio (children/adults) and e is the prevalence proportion of extra-pulmonary (extra-pulmonary/total). We obtained c from the World Data Bank[30] population estimates, whereas r and e were obtained from the completed prevalence surveys: $r = 12.5\%$ (SD 1, 4%) and $e = 10\%$ (SD 0.3%). Prevalence estimates and Model 2 predictions were compared visually by means of an adapted Bland and Altman plot of agreement—comparing the ratio of measures rather than their difference to reduce the influence of countries with very high prevalence rates.

Third, model estimates were compared with actual estimates from 2015 prevalence surveys. This could be done for two surveys conducted in 2015—in Bangladesh and in the Phillipines—which were not included in the training set because the estimates were not available when data management and analysis were performed.

Data management and analyses

All data management and analyses were done using Stata 14. All codes used for analyses are presented in Supplementary File 4, available as Supplementary data at *IJE* online.

Results

The final database included 50 data points in the training set and a total of 111 candidate predictor variables. Prevalence survey estimates for the 50 data points are as

Downloaded from https://academic.oup.com/ije/article/47/5/1645/5075187 by guest on 24 April 2024

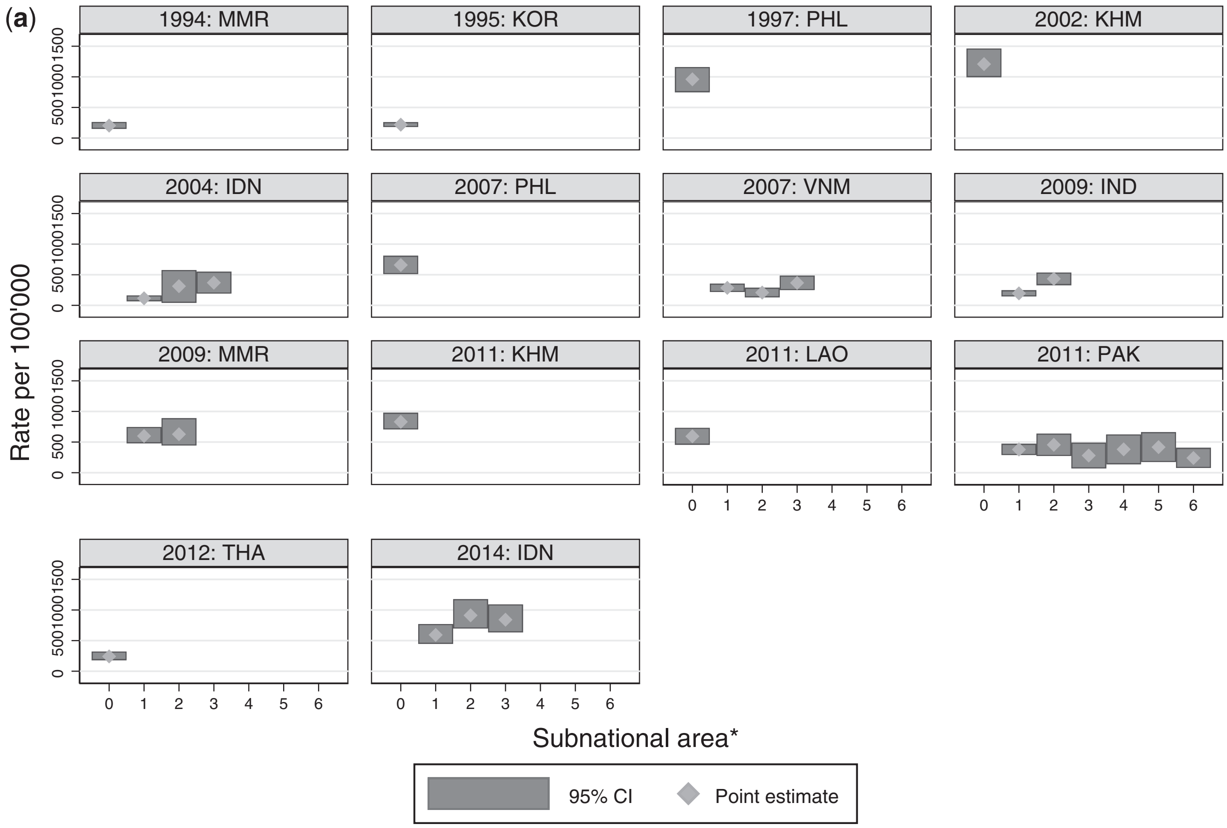
summarized in Figure 2a and b. After variable selection, 14 variables were included as potential predictors for the predictive multivariable model. Predictions were made for 63 countries (3 countries were dropped due to missing predictor variables).

Descriptive statistics by set of countries show that the profile of countries in the training set is similar to those for which predictions are made (Table 2). However, the countries in the training set appear to be much more densely populated and with a greater number of large cities, with higher male-to-female ratios at birth and lower HIV AIDS prevalence. This may partially be explained by the fact that there is a higher proportion of Asian countries in the training set and a larger proportion of African countries in the set to predict combined with period effects (the countries to predict are all from 2015 whereas the training set is from 1990 to 2015). Indeed, the number of large cities has increased since 1990, the male-to-female ratio has declined in Asia and HIV prevalence has increased.

The performance of the two final predictive multivariable models is shown in Table 3. Model 1 based on a data-driven approach to variable selection performed better than Model 2 in terms of measures of internal validity (lower AIC and higher LOOCV cross-validation correlation). However, the estimates from Model 1 were neither credible [maximum prevalence of over 8222 per 100 000 (bacteriologically confirmed cases in adults), over five times the upper CI of the prevalence survey with the highest prevalence in the training set] nor coherent (estimates were on average higher in Africa than in Asia, the opposite of what can be observed in the prevalence surveys). On the other hand, Model 2, resulting from an epidemiologically informed inclusion of variables, provided only slightly lower internal validity measures but much more credible and coherent prevalence predictions. Thus, Model 2 was retained and used for final predictions (Table 4).

Scatterplots of model predictions vs observed WHO prevalence estimates in the training set are presented in Figure 3 and other diagnostic plots in Supplementary File 5, available as Supplementary data at *IJE* online. Model 2 parameters on the logit scale, along with standard errors, variance-covariance matrix and data for predictions, are presented in Supplementary File 6, available as Supplementary data at *IJE* online.

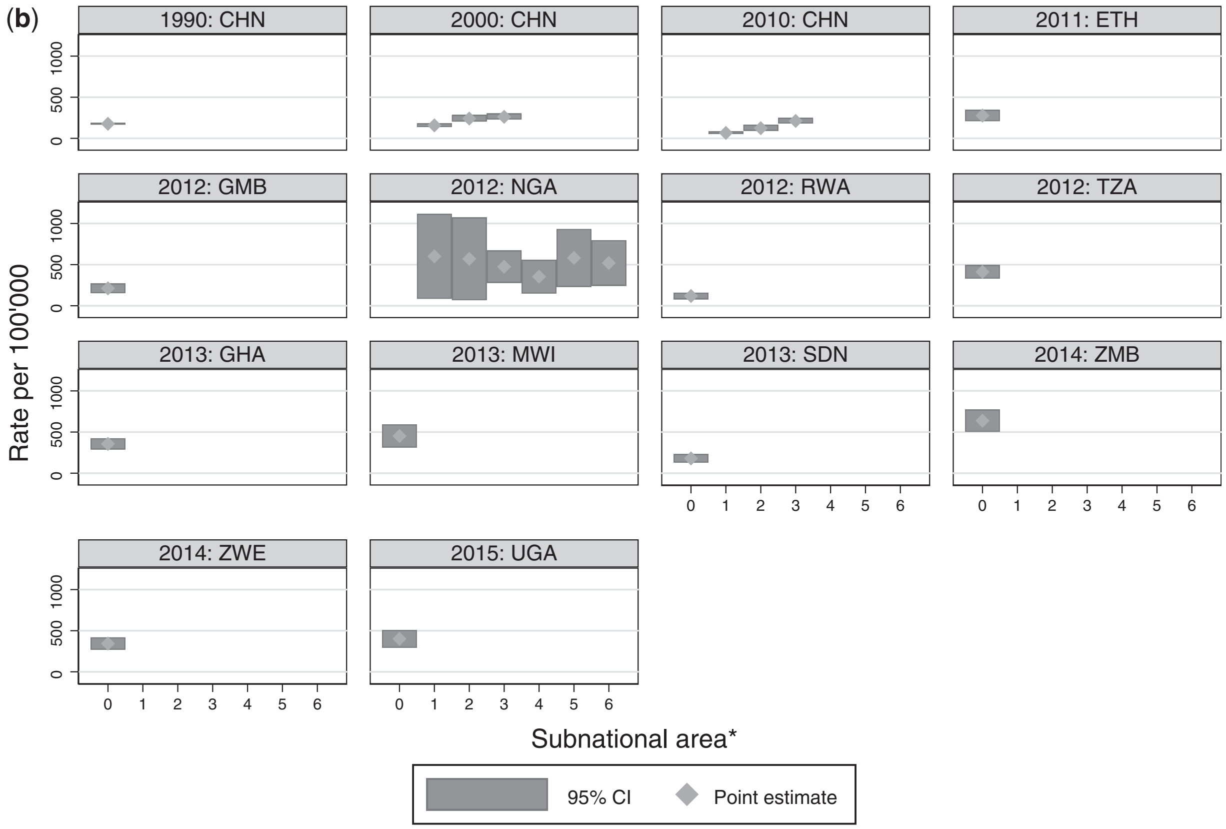
Individual country predictions based on Model 2 along with WHO estimates and Bland-Altman plots of agreement comparing the two estimates can be found in



*Subnational area=0 refers to national estimate

Figure 2. (**a**) Prevalence estimates included in the training set for Asia. (**b**) Prevalence estimates included in the training set for Africa.

Downloaded from https://academic.oup.com/ije/article/47/5/1645/5075187 by guest on 24 April 2024



*Subnational area=0 refers to national estimate

Figure 2. Continued.

Table 2. Descriptive statistics of complete predictors, by set of countries

	Training set (*n* = 50)		Countries to predict (*n* = 63)	
	Mean	SD	Mean	SD
Infant mortality (number of deaths in children under the age of 1 per 1000 live births)	44.9	21.0	43.1	20.5
Proportion of the population under the age of 15	34.2	8.5	36.8	8.0
Population density (pop/km^2)	354.6	553.0	132.7	220.7
Proportion of the population living in an urban setting	37.7	13.6	42.7	19.0
Population living in the largest city (per million)	9.3	5.9	4.9	5.1
Improved sanitation facilities (% of population with access)	50.5	19.2	47.0	26.6
Improved water source (% of population with access)	77.6	14.3	77.5	16.7
Percentage retreated TB cases out of all notified cases	7.3	4.5	10.1	7.5
New all forms TB cases notified (rate per 100 000 population)	106.7	60.8	143.1	96.3
New laboratory-confirmed TB cases notified (rate per 100 000 population)	51.5	31.8	67.4	44.0
HIV prevalence (%)	1.7	3.2	3.6	6.2
BCG coverage (%)	85.2	14.6	87.8	13.0
Climatic score (PCA)	0.01	1.54	−0.03	1.50

Supplementary File 7, available as Supplementary data at *IJE* online. Overall, there was good agreement between our model estimates and WHO estimates. The ratio (WHO estimates)/(Model 2 predictions) averaged over all countries was close to 1 (1.09, 95% CI 0.93–1.25). However, the distribution of the ratio is skewed, with five countries standing out as being more than twice as high according to WHO estimates than model predictions (Guinea Bissau, Liberia, Tanzania, Nigeria and Somalia).

The comparison of Model 2 predictions with 2015 survey estimates from Bangladesh and the Philippines provide very positive confirmations. In Bangladesh, the 2015

Downloaded from https://academic.oup.com/ije/article/47/5/1645/5075187 by guest on 24 April 2024

Table 3. Comparison of multivariable Model 1 vs Model 2

		Internal validity		External validity
	Predictor variables	AIC for full model	LOOCV R-sq	Descriptive statistics of out-of-sample predictions[a]
Model 1	1. Population density 2. BCG coverage 3. New all forms TB notification rate 4. Proportion population under the age of 15 5. Population in largest city	521.9	94%	Asia (*n* = 21): Median (IQR): 448 (307) Min-max: 122–4948 Africa (*n* = 35) Median (IQR): 539 (447) Min-max: 216–8222
Model 2	1. Continent (Africa/Asia) 2. Percentage retreated cases out of all notified 3. New all forms TB notification rate 4. Population density 5. Proportion population under the age of 15	576.4	92%	Asia (*n* = 22): Median (IQR): 542 (256) Min-max: 261–1391 Africa (*n* = 40) Median (IQR): 321 (171) Min-max: 161–1009

[a] Predicted prevalence of bacteriologically confirmed TB per 100 000 adults in 63 countries not included in model building.

Table 4. Final multivariable Model 2 (*n* = 50)[a]

Predictor	OR (95% CI)	*p*-value
Continent (Africa vs Asia)	0.52 (0.37–0.72)	<0.001
Percentage retreated out of all notified cases	1.03 (1.02–1.04)	<0.001
New all forms TB notification rate (per 10-unit increase)	1.04 (1.02–1.05)	<0.001
Population density (per 100 people/km^2 increase)	0.96 (0.95–0.97)	<0.001
Proportion population under the age of 15	1.03 (1.01–1.04)	<0.001

[a] These are exponentiated model coefficients; coefficients on the logit scale, along with standard errors, variance-covariance matrix and data for predictions, are presented in Supplementary File 6, available as Supplementary data at *IJE* online.

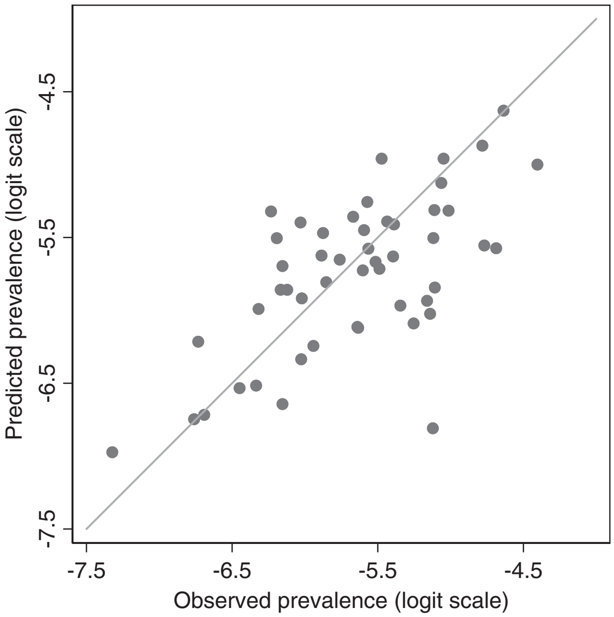


Figure 3. Predicted (Model 2) vs observed (WHO prevalence survey estimates) TB prevalence estimates in training set (*n* = 50) (all years in training set from 1991 to 2014).

prevalence survey yielded and estimated a prevalence of 260 per 100 000 (all forms all ages), fully within the 95% CI of our model predictions: 216 (95% CI 168–277). The survey in the Philippines on the other hand yielded an estimate of 980 per 100 000 for all forms and all ages. This was a much higher prevalence than anticipated by WHO. Whereas it is also above the 95% CI of Model 2 estimates (639, 95% CI 480–848), our predictions (Bland and Altman plot in Supplementary File 7, available as Supplementary data at *IJE* online) also suggested that the WHO estimates were lower than what would be expected based on the countries' TB ecological profiles.

The map presented in Figure 4 enables comparison of the geographical distribution of the WHO estimates and Model 2 predictions. In Africa, the global patterns are similar, with southern Africa generally displaying higher prevalence levels than Saharan African countries (a notable difference is the absence of predictions for the Democratic Republic of Congo and South Sudan, for which predictions could not be made due to lack of covariate data—see Supplementary file 7, available as Supplementary data at

Downloaded from https://academic.oup.com/ije/article/47/5/1645/5075187 by guest on 24 April 2024

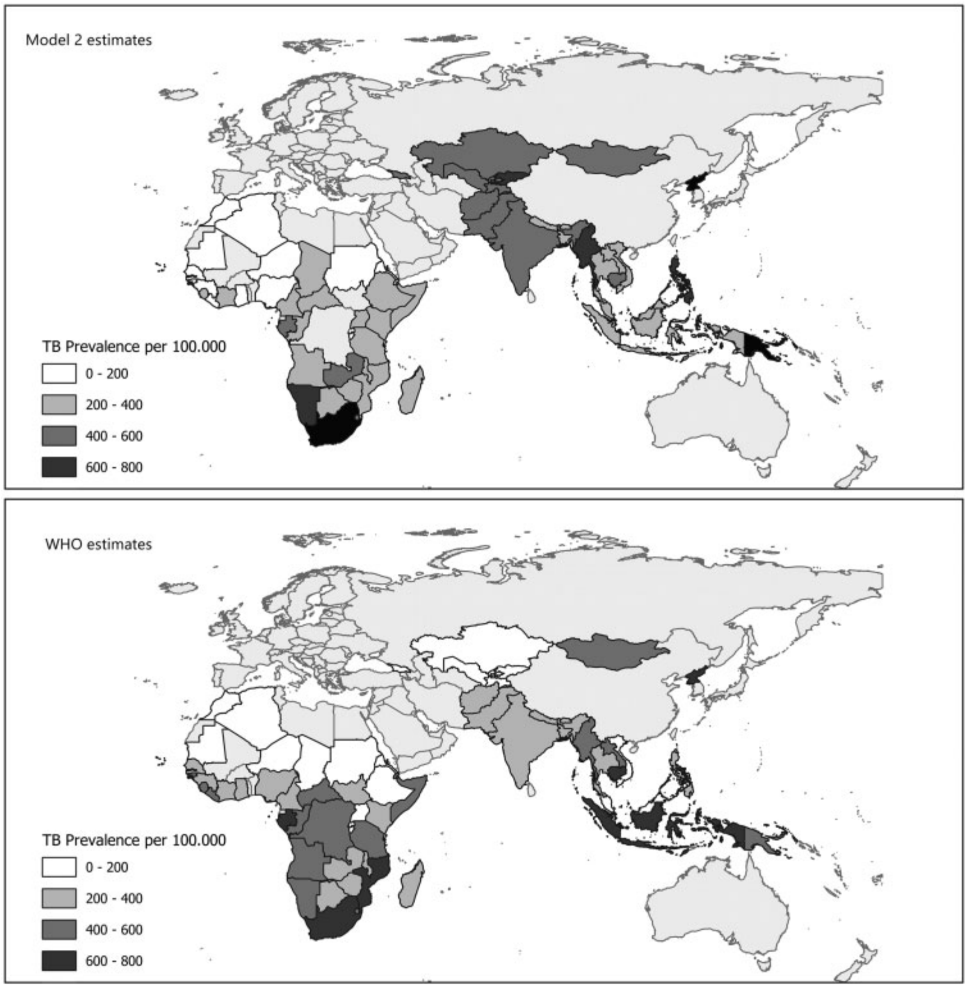


Figure 4. Maps of Model 2 predictions and WHO estimates.

IJE online, for details). With regard to Asia, Model 2 predictions for Central Asia are much higher than the WHO estimates, and estimates in India, Pakistan and Afghanistan are also higher, though the difference is not as stark. Interestingly, Model 2 predictions have corrected for the Indonesia 2014 survey estimates by lowering the prevalence as opposed to the WHO estimates, which kept the very high estimates of the survey into 2015.

Discussion

Predictive ecological modelling can provide useful complementary estimates for TB prevalence and can be considered alongside other methods in countries with limited TB data. Indeed, despite limited TB data in the countries selected for prediction, a reasonable number of ecological predictors of TB burden could be obtained from openly available databases such as the World DataBank and the Global Health Observatory data repository. Furthermore, many of those predictors could also be found at subnational levels from nationally representative surveys such as the DHS and MICS, as well as the national NTPs. By including all available subnational estimates, we were able to achieve a near 2-fold increase the number of data points in the training set. Even with a very limited number of data points in the training set, the predictive models were able to show high internal validity (cross-validations in the training set) as well as reasonably good external validity (coherence and credibility of sample predictions). The ultimate validation of our model was the comparison of 2015 predictions with actual estimates from the 2015 prevalence surveys in Bangladesh and the Philippines, which provided a very positive confirmation of the validity of our approach.

For the countries where there is good agreement between WHO estimates and our own model predictions, this modelling exercise suggests that the WHO estimates are consistent with the broader ecological landscape of those countries. For the countries where there is a wide discrepancy, the model predictions can be used as one of the sources of information to prioritize the implementation of a TB prevalence

Downloaded from https://academic.oup.com/ije/article/47/5/1645/5075187 by guest on 24 April 2024

survey or a review of the assumptions used for the estimation of TB burden. For example, in countries of the former Soviet Union (Kazakhstan, Kyrgyzstan, Georgia Tajikistan and Uzbekistan), a consistent spatial pattern was observed, with model predictions between twice and three times higher than WHO estimates (Figure 4; Supplementary File 7, available as Supplementary data at *IJE* online). All these countries have higher-than-average retreatment rates (Table 2; Supplementary File 6, available as Supplementary data at *IJE* online) and higher rates of drug-resistant TB. Whereas retreatment rates are explicitly factored in the model predictions, the WHO estimates of prevalence used in this study do not account for the frequency of retreatment and drug resistance.

To the best of our knowledge, the models presented here are the first attempt to capitalize on estimates provided by national prevalence surveys to inform estimates—in countries where surveys have not been conducted—using predictive ecological modelling. In our approach, we chose to build upon an epidemiological framework of TB burden, although conceptually at odds with a pure predictive modelling approach. Taken to an extreme, predictive modelling can be seen as a process of data mining where the only consideration is predictive accuracy. In this study, we pursued two strategies for model building: one based purely on data considerations and maximizing fit according to the AIC and one based on epidemiological judgement of which variables should figure in a model that aims to predict TB based on putative causal relationships. The latter appears to perform better by providing more coherent and credible out-of-sample predictions. This suggests that the traditionally strict predictive modelling approach may not always be the best option to predict complex disease outcomes.

The major strength of our model is that we made maximum use of all the information on TB burden available from TB prevalence survey reports, including all available subnational estimates. The binomial model we fitted implicitly weighs smaller surveys less than larger ones as coefficients and CIs are estimated by maximum likelihood estimation, where the likelihood is a function of the number of survey participants. In addition, since the number of participants used for modelling was in effect an adjusted number of participants based on the precision of estimates (*N*'), less precise estimates were also implicitly given less weight. As a result, the CIs of our predictions take into account the imprecision associated with all estimates in the training set.

The model presented here can be improved in a number of ways. The main limitation of the model is the paucity of data points available for modelling, which prevented us from fitting more complex models, since these would have resulted in overfitting and thus limited predictive power. We were not able to include non-linear relationships in the final multivariable model (although these were investigated graphically and logarithmic transformations tested univariably), nor any time trends. Therefore, first and foremost, future models will be able to benefit from the inclusion of further data points—either as the number of implemented TB prevalence surveys increases or if datasets from existing surveys are made available to derive subnational estimates where feasible and appropriate. Second, the model fitted here did not take gender into account, although TB prevalence surveys always present estimates disaggregated by sex, and a number of predictor variables (total population counts, HIV prevalence, diabetes prevalence, mortality, life expectancy, literacy, etc.) are also available disaggregated by sex. The inclusion of this level of stratification in the model would enable both an increase in the number of data points as well as accounting for gender effects and differences. Last but not least, future modelling exercises could take into account the spatial dependencies in the data more explicitly by fitting geo-statistical models.

Conclusions

National TB cross-sectional surveys provide relatively unbiased estimates of TB prevalence among surveyed populations, but also represent a major undertaking of financial and human resources. Models presented here show that TB prevalence surveys contain very useful information beyond the borders of the country in which it has been implemented. Combined with (sub)national predictors of TB, they can be used to inform TB prevalence estimates in other countries by leveraging TB notification data and socio-demographic indicators within the framework of an ecological predictive model. As the number of completed TB prevalence surveys increases, refinements to the methodology presented here could be made to increase the validity and usefulness of predictions for countries with limited TB data. This process could be facilitated and encouraged by countries and WHO making datasets publicly available for interested researchers.

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

This work was supported by funding from the WHO Global Task Force on TB Impact Measurement.

Downloaded from https://academic.oup.com/ije/article/47/5/1645/5075187 by guest on 24 April 2024

Acknowledgements

We are grateful to collaborators at KNCV for insights into the TB prevalence surveys in Vietnam, Pakistan and Nigeria, and for facilitating data acquisition by liaising with the National TB Control Programs of Myanmar and Vietnam. Many thanks to Ejaz Qadeer, Razia Fatima and Basharat Javed Khan for providing subnational TB notification data and population estimates for Pakistan. We are very thankful to Tacilja Ruckert for data-management support and thorough review of the manuscript and to Federica Giardina for insightful discussions on the statistical methods.

Conflict of interest: None declared.

References

1. World Health Organisation. *Global Tuberculosis Report 2017 [Internet]*. Geneva, 2017. http://www.who.int/tb/publications/global_report/en/ (29 March 2018, date last accessed).
2. *Health—United Nations Sustainable Development [Internet]*. http://www.un.org/sustainabledevelopment/health/ (30 December 2016, date last accessed).
3. World Health Assembly EB134.R4. *Global Strategy and Targets for Tuberculosis Prevention, Care and Controls After 2015 [Internet]*. Geneva, 2014. http://apps.who.int/gb/ebwha/pdf_files/EB134/B134_R4-en.pdf? ua=1 (30 December 2016, date last accessed).
4. Glaziou P, Sismanidis C, Zignol M, Floyd K. *Methods Used by WHO to Estimate the Global Burden of TB Disease—Technical Appendix to the Global Tuberculosis Report 2016 [Internet]*. Geneva: Global TB Programme, World Health Organization, 2016. http://www.who.int/tb/publications/global_report/gtbr2016_online_technical_appendix_global_disease_burden_estimation.pdf? ua=1 (26 July 2018, date last accessed).
5. World Health Organisation. *Tuberculosis Prevalence Surveys: A handbook [Internet]*. Geneva, 2011. http://apps.who.int/iris/bitstream/10665/44481/1/9789241548168_eng.pdf? ua=1&ua=1 (26 July 2018, date last accessed).
6. TB Team, Office of Health, Infectious Disease and Nutrition, Global Health Bureau, United States Agency for International Development. *Independent Assessment of National TB Prevalence Surveys Conducted Between 2009 – 2015 [Internet]*. 2016. http://www.who.int/tb/advisory_bodies/impact_measurement_taskforce/meetings/tf6_background_4f_prevalence_surveys_review.pdf (6 August 2018, date last accessed).
7. Woodruff RS, Winston CA, Miramontes R. Predicting U.S. tuberculosis case counts through 2020. *PLos One* 2013;**8**:e65276.
8. Zhang G, Huang S, Duan Q *et al*. Application of a hybrid model for predicting the incidence of tuberculosis in Hubei, China. *PLos One* 2013;**8**:e80969.
9. Zheng Y-L, Zhang L-P, Zhang X-L, Wang K, Zheng Y-J. Forecast model analysis for the morbidity of tuberculosis in Xinjiang, China. *PLoS One* 2015;**10**:e0116832.
10. Moosazadeh M, Khanjani N, Nasehi M, Bahrampour A. Predicting the incidence of smear positive tuberculosis cases in Iran using time series analysis. *Iran J Public Health* 2015;**44**:1526–34.
11. Ade S, Békou W, Adjobimey M *et al*. Tuberculosis case finding in Benin, 2000–2014 and beyond: a retrospective cohort and time series study. *Tuberc Res Treat* 2016;**2016**:3205843.
12. Azeez A, Obaromi D, Odeyemi A, Ndege J, Muntabayi R. Seasonality and trend forecasting of tuberculosis prevalence data in Eastern Cape, South Africa, using a hybrid model. *Int J Environ Res Public Health* 2016;**13**:757.
13. Bras AL, Gomes D, Filipe PA, de Sousa B, Nunes C. Trends, seasonality and forecasts of pulmonary tuberculosis in Portugal. *Int J Tuberc Lung Dis* 2014;**18**:1202–10.
14. Kumar V, Singh A, Adhikary M, Daral S, Khokhar A, Singh S. Seasonality of tuberculosis in delhi, India: a time series analysis. *Tuberc Res Treat* 2014;**2014**:514093.
15. Klotz A, Harouna A, Smith AF. Forecast analysis of the incidence of tuberculosis in the province of Quebec. *BMC Public Health* 2013;**13**:400.
16. Liao C-M, Hsieh N-H, Huang T-L *et al*. Assessing trends and predictors of tuberculosis in Taiwan. *BMC Public Health* 2012;**12**:29.
17. Houben RMGJ, Menzies NA, Sumner T *et al*. Feasibility of achieving the 2025 WHO global tuberculosis targets in South Africa, China, and India: a combined analysis of 11 mathematical models. *Lancet Glob Health* 2016;**4**:e806–15.
18. Murray CJL, Ortblad KF, Guinovart C *et al*. Global, regional, and national incidence and mortality for HIV, tuberculosis, and malaria during 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet Lond Engl* 2014;**384**:1005–70.
19. Shmueli G. To explain or to predict? *Stat Sci* 2010;**25**:289–310.
20. Lönnroth K, Jaramillo E, Williams BG, Dye C, Raviglione M. Drivers of tuberculosis epidemics: the role of risk factors and social determinants. *Soc Sci Med* 2009;**68**:2240–46.
21. Lönnroth K, Castro KG, Chakaya JM *et al*. Tuberculosis control and elimination 2010–50: cure, care, and social development. *Lancet Lond Lancet* 2010;**375**:1814–29.
22. Suthar AB, Lawn SD, del Amo J *et al*. Antiretroviral therapy for prevention of tuberculosis in adults with HIV: a systematic review and meta-analysis. *PLoS Med* 2012;**9**:e1001270.
23. Narasimhan P, Wood J, MacIntyre CR, Mathai D. Risk factors for tuberculosis. *Pulm Med* 2013;**2013**:828939.
24. Lin H-H, Suk C-W, Lo H-L, Huang R-Y, Enarson DA, Chiang C-Y. Indoor air pollution from solid fuel and tuberculosis: a systematic review and meta-analysis. *Int J Tuberc Lung Dis* 2014;**18**:613–21.
25. Roy A, Eisenhut M, Harris RJ *et al*. Effect of BCG vaccination against Mycobacterium tuberculosis infection in children: systematic review and meta-analysis. *BMJ* 2014;**349**:g4643.
26. Falzon D, Mirzayev F, Wares F *et al*. Multidrug-resistant tuberculosis around the world: what progress has been made? *Eur Respir J* 2015;**45**:150–60.
27. Fares A. Seasonality of tuberculosis. *J Glob Infect Dis* 2011;**3**:46–55.
28. World Health Organisation. *The WHO Global TB Data Collection System [Internet]*. https://extranet.who.int/tme/ (2 January 2017, date last accessed).
29. World Health Organisation. *Global Health Observatory Data Repository [Internet]*. http://apps.who.int/gho/data/node.home (2 January 2017, date last accessed).
30. The World Bank. *The World Bank DataBank [Internet]*. http://databank.worldbank.org/data/databases.aspx (2 January 2017, date last accessed).
31. WHO-UNICEF. *WHO UNICEF Coverage Estimates WHO World Health Organization: Immunization, Vaccines And Biologicals. Vaccine Preventable Diseases Vaccines Monitoring System 2016 Global Summary Reference Time Series: BCG*

Downloaded from https://academic.oup.com/ije/article/47/5/1645/5075187 by guest on 24 April 2024

[Internet]. http://apps.who.int/immunization_monitoring/globalsummary/timeseries/tswucoveragebcg.html (2 January 2017, date last accessed).

32. GeoNames [Internet]. http://www.geonames.org/statistics/ (2 January 2017, date last accessed).
33. UNAIDS. *AIDSinfo [Internet]*. http://aidsinfo.unaids.org/ (2 January 2017, date last accessed).
34. International Diabetes Federation. *Diabetes Atlas [Internet]*. http://www.idf.org/diabetesatlas/ (2 January 2017, date last accessed).
35. Demographic and Health Surveys. *The DHS Program—Publication Search [Internet]*. https://dhsprogram.com/publications/publication-search.cfm (2 January 2017, date last accessed).
36. UNICEF. *Surveys—UNICEF MICS [Internet]*. http://mics.unicef.org/surveys (2 January 2017, date last accessed).
37. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;**49**:1373–9.
38. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med* 2004;**66**:411–21.
39. Stone M. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc Ser B Methodol* 1974;**36**:111–47.
40. Geisser S. The predictive sample reuse method with applications. *J Am Stat Assoc* 1975;**70**:320–28.

Downloaded from https://academic.oup.com/ije/article/47/5/1645/5075187 by guest on 24 April 2024