Predictive statistical modelling to inform TB incidence, prevalence and mortality estimates: interim report

Prepared by:
Sandra Alba, Ente Rood, Masja Straetemans and Mirjam Bakker
Predictive statistical modelling to inform TB incidence, prevalence and mortality estimates

Background paper – Interim report
Prepared by KIT Biomedical Research
for the WHO Global Task Force on TB Impact Measurement
3rd meeting of methods subgroup
Glion-sur-Montreux
March 31 - April 2 2015

March 2015
Sandra Alba, Ente Rood, Masja Straetemans and Mirjam Bakker

KIT Biomedical Research, Royal Tropical Institute
Amsterdam
Table of Contents
1. Introduction .................................................................................................................................... 3
2. Goal ................................................................................................................................................ 4
3. Conceptual framework ................................................................................................................... 5
4. Database compilation ..................................................................................................................... 6
5. Data preparation ............................................................................................................................ 7
6. Exploratory data analyses .............................................................................................................. 12
7. Selection criteria for entry of predictors in model ......................................................................... 13
8. Model building and model selection approach ............................................................................. 14
9. Final predictive models ................................................................................................................. 15
10. Model validation ........................................................................................................................ 17
11. Predictions ................................................................................................................................ 18
12. Conclusions .............................................................................................................................. 21
1. Introduction

The Global Task Force on Tuberculosis (TB) impact measurement was established in June 2006 by the Global TB programme in the World Health Organization (WHO). The aim of TB impact measurement is to measure the burden of disease caused by TB in terms of the impact indicators TB incidence, prevalence and mortality. It is within the Global Task Force’s mandate to produce a robust, rigorous and widely endorsed assessment of whether the 2015 global targets set for TB control are achieved at global, regional and country levels. In the context of global TB strategies and targets, WHO publishes a global TB report every year since 1997.

In 2015 a global consultation will be organized with the aim to review and update where appropriate current methods for producing estimates of TB disease burden, and an associated strategy for using these methods to assess whether the global TB targets are achieved at global, regional and county levels. Therefore alternative options (compared with current methods used by WHO) need to be investigated.

The Epidemiology group of the Royal Tropical Institute (KIT), KIT Biomedical Research, was appointed to develop 3 sets of ecological predictive models for the 2015 global consultation to estimate TB prevalence, incidence and mortality using data obtained from state of the art estimation procedures (as input for models) and established risk factors.

Approach to model building

This document reports on the methodological approach and preliminary results from the development of predictive ecological models for TB incidence (Task 1), TB prevalence (Task 2) and TB mortality among HIV-negative individuals (Task 3). Predictive models differ substantially from explanatory models, which are most common in epidemiology. Our model building approach followed the guidance outlined by Schmueli in the seminal paper “To explain or to predict?”1. The primary aim of the models is to maximize predictive power, and not to correctly capture the causal pathways between TB burden and risk factors. When going through this document readers are urged to resist the temptation of interpreting results within the “explanatory paradigm” and not to try to ascribe causal relations based on the estimated effects.

Status of this document

The final goal of the predictive models is to enable predictions of TB incidence, prevalence and mortality from 1990 to 2015 for a selection of countries for which the Task Force is mandate to produce estimates, and to identify a set of conditions which warrant or do not warrant the use of these models. This document focuses on the development of the basic database and model structures to enable predictions for 2013, which is what we report on in this document. This report is intended to provide elements of discussion and trigger constructive criticism during the global consultation on meetings for further development of the models. The report includes:

- Presentation of conceptual framework
- Methodology for database compilation and model building process

1 Galit Shmueli, To Explain or to Predict?, Statistical Science 2010, Vol. 25, No. 3, 289–310
• Assessment of the goodness of fit and internal validity of models.
• Predictions of prevalence estimates for 2013 and a comparison with WHO 2013 estimates
• Preliminary reflections on the usefulness of the methods
• Further analyses needed and data wish list in order to refine the models

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; to Dr Ejaz Qadeer, Dr Razia Fatima and Dr Javed Busharat for providing subnational TB notification data and population estimates for Pakistan; to Dr Bin Hoa and Dr Le Van Hoi for details of the Vietnam prevalence survey and for providing subnational TB notification data for Vietnam; to Dr May Thinzar Kyi and Dr Ervin Cooreman for providing subnational TB notification data for Myanmar. We are thankful to Tacilja Ruckert for data management support. Last but not least, many thanks to Babis Sismanidis for insightful discussions and support throughout the process.

2. Goal

Task 1

The goal of this task is to develop a predictive ecological model for TB incidence that uses available data from direct TB incidence estimates (Method 1 in Table 1 below- mostly used in high income countries, but also in some middle income countries) to predict TB incidence in middle and high income countries without direct method (i.e. using Method 3 and 4).

Note that an alternative definition of the inputs for this model was also considered, i.e. using available data from the other direct TB incidence estimates (Method 2 - mostly middle-low income countries) to predict TB incidence in other middle-low income countries without direct method (i.e. using Method 3 and 4). However method 2 has only be used in 3 countries (Cambodia, Ethiopia and Laos), which is too few to build a predictive model, so this approach was not considered further.

Table 1. Overview of WHO methods for the estimation of TB incidence

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>State of the art TB surveillance, factoring in under-diagnosis and under-reporting (e.g. estimated in capture recapture surveys) (labeled as “high income” or “capture-recapture” in WHO database)</td>
</tr>
<tr>
<td>Method 2</td>
<td>Incidence (I) derived from prevalence surveys (I=prevalence/duration) (duration estimated by mathematical compartmental models) (labeled as “prevalence” in WHO database)</td>
</tr>
<tr>
<td>Method 3</td>
<td>I=mortality/Case Fatality Ratio (labeled as “mortality” in WHO database)</td>
</tr>
<tr>
<td>Method 4</td>
<td>Incidence based on expert opinion, in countries with not very reliable data, I=notification/1-un-reported where un-reporting is estimated from expert opinion (labeled as “Expert opinion” in WHO database)</td>
</tr>
<tr>
<td>Other methods</td>
<td>“neighbour” in West-Bank &amp; Gaza and South Sudan, “Survey” in France, “trends ART” in Bhutan and India</td>
</tr>
</tbody>
</table>
Task 2
The aim of this task was to develop a predictive ecological model for TB prevalence that uses available data from recent national TB prevalence surveys to predict TB prevalence in low and middle-income countries with predicted prevalence of over 0.1% where national surveys have not been implemented.

Task 3
The aim of this model was to develop a predictive model for TB mortality among HIV-negative individuals using data from countries using vital registration to estimate mortality (method labeled as “VR” or “VR imputed” in WHO database) (mostly from middle and high-income countries) to predict mortality in countries without vital registration data (labeled as “Indirect” in WHO database) (mostly low-income countries).

3. Conceptual framework

A literature search was performed (see Annex 1 for list of publications consulted) to identify predictors of TB burden. These can be categorized into our four categories of variables (see Annex 1 for detailed description and potential causal mechanism):

1) TB data
   - TB case notification (active TB in community)
   - MDR-tuberculosis burden: detection, enrolment, and treatment outcomes
2) Programmatic determinants
   - weak health system
   - inappropriate health seeking behavior (level of awareness)
   - poor access to TB services (DOTS coverage)
   - poor treatment outcome
   - BCG vaccination coverage among children
3) Co-morbidities
   - HIV
   - Malnutrition/poor-nutritional status
   - Diabetes
   - Lung diseases
4) Socio environmental factors
   - Humidity
   - Prevalence of high risk groups: prisoners, homeless people, migrants, drug addicts, refugees, displaced populations,
   overarching:
   - urbanization
   - demographic transition
   - migration
   - poverty, low SES, low education
   leading to:
   - population density
- poor water source and sanitation
- crowded living conditions
- poor ventilation
- indoor air pollution (caused by burning of solid fuels) (low level evidence)
- tobacco smoke
- alcoholism
- aging populations
- outdoor air pollution (limited research)

4. Database compilation

The first step in the database compilation was the definition of 2-stage data quality framework and assessment to guide the choice of variables to include in the process based on the OECD Quality framework and review\(^2\). For data compilation variables that were relevant, timely, accessible and interpretable were chosen. See Annex 2 for details

Outcome variables (incidence, prevalence and mortality) were derived from the WHO TB database which consist of data obtained through the WHO Global TB data collection system.\(^3\) TB prevalence subnational estimates were obtained from TB prevalence survey reports as well as from the authors of the report or collaborators in the implementation of the survey.

Predictor variables were downloaded from openly available data sources to capture the dimensions in the conceptual framework. For data at national level sources of data included: TME, the Global Health Repository (GHR), the World Bank, UNICEF reports (BCG prevalence). Sources of data for subnational areas included data from national bureaus of statistics (e.g. census) and Multiple Indicator Surveys (MICS). See Annex 3 for full details. Datasets including predictor variables were merged to the outcome datasets data based on a country-code (iso3) and a timestamp (year).

Data base specifications and data dictionaries can be found in Annex 3 (including naming conventions for data management and data analysis files).

Some country definitions differed across datasets. The major differences were between the TME data (from which outcome variables were derived) and the World Bank and GHR datasets. Countries for which data in the World Bank data was not available could also not be classified by income status (high, middle and low income), and thus were not used in the tasks which were dependent on classification by income status. The non-availability of data in the WB or GHR impact the ability to make predictions for certain countries, depending on whether these variables were included in the final multivariate models.

No data available in WB dataset in 2013 (but available in TME)

- AIA Anguilla
- MSR Montserrat
- TKL Tokelau
- COK Cook Islands

---


\(^3\) [https://extranet.who.int/tme/](https://extranet.who.int/tme/)
5. Data preparation

The first step in data preparation consisted of defining which countries would be used to develop the predictive models, and for which countries predictions would be made. This involved decisions, made in consultation with the Task Force, about which countries’ data can be considered of highest quality and can be relied upon in order to build a predictive model. The subset of data pertaining to these countries is called the “training set” – it is the set on which the models are developed.
(“trained”) in order to make predictions in the countries whose data is considered less reliable and should be predicted. An overview of input and output data for each of the tasks is presented in Tables 2-4 and maps indicating which countries are in the training set and for which predictions will be made are shown in Figures 1-3. Task specific issues for Task 2 are presented below.

Table 2. Input and outputs of Task 1

<table>
<thead>
<tr>
<th>Input/Output</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent variable</td>
<td>Numerator: Estimated number of incident cases (all forms) (Estimates Method 1 see Table 1) and denominator: Estimated total population number; both from WHO TB burden file</td>
</tr>
<tr>
<td>Independent variables</td>
<td>TB data programmatic, co-morbidities and socio-environmental predictors (See Annex 1 for details)</td>
</tr>
<tr>
<td>Countries used for prediction</td>
<td>In first instance, 1,688 datapoints over 24 years (1990-2013); 72 countries eligible in 2013 (64 high income countries, 4 middle income countries and 4 with missing income status); final model based on 213 datapoints with complete data</td>
</tr>
<tr>
<td>Countries to predict</td>
<td>2013 estimates for 100 middle income countries and 6 high income countries; however, not all of these have a complete set of data to enable prediction.</td>
</tr>
</tbody>
</table>

Figure 1. Countries used for TB incidence prediction (Training set) and countries for which incidence was predicted
Table 3. Input and outputs for Task 2

<table>
<thead>
<tr>
<th>Input/Output</th>
<th>Description</th>
</tr>
</thead>
</table>
| Dependent variable    | Bacteriologically confirmed (BC) TB prevalence from prevalence surveys conducted from 2007 onwards, when the methodology for analysis of TB prevalence surveys was standardized as documented in the WHO “Lime book”  
                           Some TB prevalence surveys present data for subnational areas. Where possible these data were also used. See Annex 4 for details |
| Independent variables | TB, programmatic, co-morbidities and socio-environmental predictors (See Annex 1 for details)  
                           For countries for which subnational estimates of TB are available, predictors were obtained from censuses, Demographic Health Surveys (DHS), Central Bureau of Statistics (CBS) and other representative surveys. TB notification data at subnational level were obtained from national TB control programs (NTPs). Predictors that are available at national AND subnational level can be used fully at disaggregated level, predictors only available nationally were averaged out on all subnational levels. |
| Countries used for prediction | National prevalence survey data for 13 countries, subnational estimates for 5 countries and 2 district level prevalence survey estimates (30 datapoints in total)                     |
| Countries to predict  | 2013 estimates for 25 low and 49 middle income countries without prevalence survey with expected prevalence <0.1% according to WHO estimates                                                                             |

Figure 2. Countries used for TB prevalence prediction (Training set) and countries for which prevalence was predicted
Table 4. Inputs and outputs for Task 3

<table>
<thead>
<tr>
<th>Input/Output</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent variable</td>
<td>Numerator: Estimated number of deaths from TB (all forms, excluding HIV) from vital registration (VR or VR imputed)</td>
</tr>
<tr>
<td></td>
<td>Denominator: Estimated total population number; both from WHO TB burden file</td>
</tr>
<tr>
<td>Independent variables</td>
<td>TB, programmatic, co-morbidities and socio-environmental predictors (See Annex 1 for details)</td>
</tr>
<tr>
<td>Countries used for prediction</td>
<td>In first instance 3,022 datapoints over 24 years (1990-2013); 126 countries eligible in 2013 (59 high, 62 middle, 2 low income countries and 4 with missing income status) final model based on 307 datapoints with complete data</td>
</tr>
<tr>
<td>Countries to predict</td>
<td>2013 estimates for 11 high, 42 middle and 32 low income countries and 6 countries with missing income status</td>
</tr>
</tbody>
</table>

Figure 3. Countries used for TB mortality prediction (Training set) and countries for which mortality was predicted

Task 2
The training set for this task included prevalence surveys conducted from 2007 onwards. These are the surveys whose data have been analysed according to the “state of the art” methodology as described in the WHO “Lime book”, or a method judged closely similar enough (see Annex 4 for details). This included 13 nationally representative surveys as well as 3 district level surveys conducted in India: 2009 Jabalpur (Madhya Pradesh state) and 2009 Bangalore Rural (Karnataka
state). The 2007 Thiruvallur (Tamil Nadu state) survey was dropped from the database as the methodology used for analyses as reported in the source publication did not appear to be consistent with the standardized “Lime book” methodology.

National prevalence surveys included in the database

- Philippines 2007
- Vietnam 2007
- Bangladesh 2008
- Myanmar 2009
- China 2010
- Pakistan 2011
- Cambodia 2011
- Ethiopia 2011
- Lao People’s Democratic Republic 2011
- Gambia 2012
- Nigeria 2012
- Rwanda 2012
- Thailand 2012

In addition disaggregated subnational data was obtained from the reports or the authors of the reports of the following 5 surveys: Vietnam (3 areas), Myanmar (2), China (3), Pakistan (6), Nigeria (6). See Annex 4 for details. The outcome variable was the prevalence of bacteriologically confirmed (BC) TB from TB prevalence surveys. Candidate models for this task included GLM models for which the numerator and denominator need to be specified explicitly. Prevalence surveys report numerators (BC TB) and denominators (number of participants in survey) for nationwide estimates. However the ratio between these two quantities does not equate the final estimated prevalence as the latter is derived from models which take into account population weighing, clustering, non-participation and missing values. Furthermore for disaggregated subnational data the numerators and denominators were sometimes not available. Therefore, an adjusted number of BC (bc) and participants (n) were estimated based on the final reported prevalence estimates and their confidence intervals (note that the prevalence estimate remains unchanged through these calculations, it is just the numerator and denominator that are adjusted). This was done using a very crude method based on the prevalence (p) and the upper limit (ul) and lower limit (ll) of the confidence interval and assuming a normal symmetrical interval on either side of the prevalence estimate: n1=(p*(1-p))/((ul-p)/1.96)^2, n2=(p*(1-p))/(((ll-p)/1.96)^2), n=(n1+n2)/2.

This method will have to be revised at a later stage using formulas which adequately capture the asymmetrical nature of a confidence interval for a proportion (e.g. the arcsine transformation). It is interesting to note, in passing, that the adjusted numerators and denominators using this approach yielded quantities approximately half of the number of cases and participants in the survey (mean adjusted n used for modelling=28000, mean actual number of participants across surveys=60277, mean adjusted number of BC used for modelling=77, mean actual number of BC cases found across prevalence surveys=179). This is consistent with a design effect of 2, assumed in many of the surveys.
The final number of datapoints available for analyses in the training set was 30. Estimated prevalence estimates are as summarized in Figure 4 below. Note that the graph includes both national and subnational estimates for countries for which subnational data was analysed for illustrative purposes only.

Figure 4. Prevalence estimates included in training set for Task 2

6. Exploratory data analyses

All tasks
The final database (for all tasks) included 166 predictor variables. An exploratory data analyses (EDA) was conducted to assess the quality of the data by means of a defined set of quality criteria including completeness, accuracy, credibility and coherence (Annex 2).

Predictors in cases where predictor data was only available at set intervals (e.g. every 5 years) missing values were imputed using a linear imputation. Imputed values were calculated using the start and end observations of a gap as anchor points between which missing values were linearly imputed. When predictor data was missing for the most recent years only, a linear trend imputation was used to extrapolate the existing series. A linear trend was fitted to the data using a simple linear model with time in years as a predictor. Only predictor variable which showed a linear trend with a fit of > 90% were extrapolated one year forward maximum. See Annex 5 for an overview of data completeness before and after imputation.
Data quality checks were also performed on the outcome variable to assess the eligibility of the specific observations to be included in the final training dataset. Quality assurance of the outcome data included checks for accuracy coherence and credibility (see Annex 2 for details). Further details are provided in the task specific points below.

**Task 2**
The Bangladesh survey only reported SS+, so estimated BC based on the ratio between SS+ and BC from prevalence surveys conducted WPR and SEA region in 2007 (year of Bangladesh survey). The surveys used for the calculation were thus: China, Cambodia, Lao People’s Democratic Republic, Myanmar, Philippines, Thailand and Viet Nam. The ratio was 0.456, so the prevalence of BC was estimated as follows: prev_bc_100k=prev_sp_100k/0.4565.

The report from the Jabalpur survey concluded that BC estimates from the survey should be corrected by a factor 1.7 to account for no x-ray screening, which was done.

The confidence intervals of 3 Nigeria subnational estimates were very wide. Given the paucity of datapoints for model 2 these were keep for modeling but their impact on model fit was assessed after all modeling.

**7. Selection criteria for entry of predictors in model**

**Task 1 and Task 3**
Predictor variables were selected for inclusion in the multivariate model based on the following criteria and procedure described below:

- Predictors were selected based on completeness in the training dataset (<40% complete were excluded)
- Complete predictors were univariate fitted to the mortality count data and ranked based on deviance scores
- Pairwise correlations were calculated and correlated predictors (> 0.7) were dropped based on the lowest relative fit to the mortality count data.

First, predictors were selected based on completeness of all records. Predictors which were less than 40% complete for all country-years for those countries included in the training dataset were excluded. This resulted in a total set of 62 predictor-variables.

In the second step each predictor-variable was fitted to the mortality counts per country-year using a univariate negative binomial regression model (log link function) including an offset based on the log transformed population size. Model fit was assessed based on AIC values and the relative reduction in residual error as compared to a model including an intercept term only (see Annex 8 for results).

Thirdly, variables were ranked based on the best fit and pairwise Pearson correlation coefficients were calculated. Variables with rho values > 0.70 were excluded in a pairwise fashion based on the lowest relative fit to the outcome measure during the univariate analysis results (see Annex 8 for results). This resulted in a final set of 20 predictor-variables to be included in the subsequent multivariate analysis.
**Task 2**

Given the very small sample of 30 countries available for model fitting, only predictor variables which had complete data for all 30 countries (after the imputation described above) were included as candidate variables in the model building process. As a result, the number of variables dropped from 166 to 68.

Very highly correlated variables were also dropped. A cut off values of rho >0.80 was used. Scatter plots and Pearson’s correlation coefficients were calculated for all groups of variables which were thought to be highly correlated as expressing the same concept/construct (see framework in Annex 1), e.g. government expenditure variables, life expectancy, various lags of laboratory confirmed TB notifications etc). When very highly correlated variables were found, the one with the highest correlation with prevalence was selected for entry in the model. Following this step 37 variables were left for model fitting.

---

**8. Model building and model selection approach**

**Task 1 and Task 3**

The final outcome variable predicted by the multivariate model was the total count of TB associated mortalities, corrected for the population size (offset variable). The distribution underlying these data was therefore assumed to be a discrete count, which was estimated using generalized linear models. Poisson, negative binomial and zero inflated (two-stage) distribution models were fitted to the data.

Fitting multivariate modelling requires complete data records for all variables included in the model. Therefore missing values for a single variable led to the omission of a full data record. To test whether records were omitted based on missing completely at random of covariate values (here independent of the outcome), the proportion of records missing were tested against mortality and incidence values. The number of records missing per country and per year and the binomial association with incidence are provided in Annex 6 and with mortality in Annex 8.

**Task 2**

Two types of generalized linear models (GLM) were considered for model fitting, assuming binomial (logistic link) and negative binomial distributions for the number of BC cases. The negative binomial model included an offset based on the log transformed adjusted number of participants in the survey. A random effect was included to account for clustering by country. A histogram of the distribution of cases prevalence estimates is shown in Annex 7.

Univariate models were fitted against all 36 predictor variables and were assessed by means of the Akaike Information Criteria (AIC). The model which provided the best fit to the data in univariate models in terms of the AIC was the negative binomial model.
The aim was to build a multivariate model with 3 predictor variables, given the rule of thumb that there should be 10 observations per covariate in a regression model to avoid overfitting. The 10 covariates with the strongest AIC in univariate analyses were entered in a multivariate negative binomial and binomial models. Variables were dropped by backward elimination based on p-values until only significant variables were left in the model. Principal components analysis was used for variable reduction (see section “Model Selection” for details).

All tasks
The final multivariate model best fitting the data was selected based on the relative fit to the data using the Akaike Information Criterion AIC (Likelihood based) – lower values of this statistic indicate better fit to the data. In this procedure variables were sequentially added and removed from the model and model fit was assessed.

To test which distribution best fitted the data the goodness-of-fit of each distribution was assessed by means of the deviance scores (Equation 1 below) and associated p-values (chi-square). Model fit was assessed by plotting predicted versus actual values, deviance residuals versus predicted values, and histogram of deviance residuals (should be normally distributed).

\[ D = 2 \* \sum_{i=1}^{N} \left\{ y_i \left( \ln \frac{y_i}{\hat{\mu}_i} \right) - (y_i + \theta) \ln \left( \frac{y_i + \theta}{\hat{\mu}_i + \theta} \right) \right\} \]  
(Equation 1)

The results of the final best fitting model were validated using multi-fold cross validation procedure. This procedure splits the data randomly into k partitions, then for each partition it fits the specified model using the other k-1 groups. A pseudo-R-sq value is then calculated as the square of the correlation coefficient of the predicted and actual values of the dependent variable.

9. Final predictive models

Task 1
The results of the best fitting negative binomial model for the prediction of mortality and incidence rates are shown in Table 5 and Table 6. In the null models the zero inflated negative binomial models provided better fit to the data than the standard negative binomial models. However, after entry of all predictor variables in the models, the fits of both models was similar, because the subset of data with complete cases did not contain a large number of zeros in the outcome variables. Thus negative binomial models are presented. Residuals plots are shown in Annex 6 and Annex 8.
Table 5. Estimated coefficients (log scale) of final negative binomial multivariate model for incidence (n=213)

|                | Estimate | Std. Error | Pr(>|z|) |
|----------------|----------|------------|----------|
| (Intercept)    | -8.02700 | 4.40E-01   | 0.00000  |
| totexphaer     | -0.00008 | 3.35E-05   | 0.01352  |
| govexphgdp     | -0.10510 | 2.77E-02   | 0.00015  |
| lifeexp60m     | -0.08271 | 1.84E-02   | 0.00001  |
| pop60          | 0.02053  | 7.16E-03   | 0.00416  |
| tmax           | 0.00311  | 6.25E-04   | 0.00000  |
| ret_af         | 0.00054  | 8.47E-05   | 0.00000  |
| prec           | 0.00027  | 4.91E-05   | 0.00000  |
| hivtest_pos    | 0.00141  | 2.22E-04   | 0.00000  |
| perc_mdr_new   | 0.00466  | 2.13E-03   | 0.02835  |
| gnipp          | 0.00001  | 2.39E-06   | 0.00002  |
| hivtest        | -0.00005 | 1.81E-05   | 0.00447  |
| DM_prev        | -2.31000 | 1.25E+00   | 0.06536  |
| perc_ret       | 0.02368  | 8.49E-03   | 0.00531  |

Table 6. Estimated coefficients (log scale) of final negative binomial multivariate model for mortality (n=307)

|                | Estimate  | Std. Error | Pr(>|z|) |
|----------------|-----------|------------|----------|
| (Intercept)    | -10.5774  | 0.04759    | < 2e-16  |
| lifeexpbm      | -0.18783  | 0.06543    | 0.00404  |
| new_afr        | 0.71356   | 0.05275    | < 2e-16  |
| urbanpop       | 0.14584   | 0.04478    | 0.001128 |
| perc_mdr_new   | 0.05991   | 0.02831    | 0.034283 |
| govexphgdp     | -0.13399  | 0.03999    | 0.000806 |
| DM_prev        | -0.09344  | 0.05279    | 0.076741 |
| hivtest_pos    | 0.16718   | 0.08158    | 0.040442 |
| bcg            | 0.17497   | 0.04047    | 1.54E-05 |
| perc_ret       | 0.15697   | 0.04331    | 0.00029  |
| prec           | 0.11985   | 0.03476    | 0.000564 |
| c_ret_tsr      | -0.21212  | 0.04269    | 6.75E-07 |
**Task 2**

The full results from the univariate binomial models are presented in Annex 7. The final chosen model was the negative binomial model – despite showing higher AIC values the correlation between predicted and observed values in the final multivariate models was much better in the binomial compared to the negative binomial model.

Principal components analysis was used for variable reduction by combining the following candidate variables from the univariate models: mean temperature in the coldest month of the year (tmin), mean temperature in the year (tmean) and precipitation (prec) in one overall “climatic score” (clim_score). The score was calculated on the full dataset consisting of the training observations together with the observations in the “to predict” set of observations. The first component was shown to explain 77% of the total variance. Higher values of this score characterize countries with warmer temperatures and average rain, i.e. tropical and sub-tropical countries. See Annex 7 for details.

The final multivariate models were fitted with and without the 3 subnational estimates in Nigeria with very large confidence interval (North Central, North West and South South). Since the AIC were consistently higher when excluding these three datapoints, the final model excluding them was retained. Results are shown in Table 6 below and residuals plots are shown in Annex 7.

### Table 7. Estimated coefficients (log scale) of final binomial multivariate model for incidence (n=27)

|                | Estimate | Std. Error | Pr(>|z|) |
|----------------|----------|------------|---------|
| (Intercept)    | -3.035882| 0.907357   | 0.001   |
| new_labconfr   | 0.008118 | 0.002236   | <0.001  |
| clim_score     | 0.160393 | 0.075830   | 0.034   |
| bcg            | -0.036120| 0.010205   | <0.001  |

10. **Model validation**

**Task 1 and Task 3**

The results of the final best fitting model were validated using a 5-fold cross validation procedure. In this procedure the dataset was randomly split into 5 mutually exclusive subsets. The cross validation showed that the model was able to predict the incidence test data well with an average pseudo-R2 of 0.94 (See Annex 6 for details) and the mortality test data with an average pseudo-R2 of 0.89 (See Annex 8 for details).

**Task 2**

The results of the final binomial multivariate model were validated using a 2-fold cross validation procedure given the small sample size used for model fitting, repeated 5 times. The median R-sq from the 10 crossfold validations was 0.76 (See Annex 6 for details).
11. Predictions

All tasks

Scatterplots of model predictions versus observed WHO estimates for all three models in the training set are presented below in Figures 5-7. Predictions on the log scale are presented in the Annexes and show better fit especially for the incidence and mortality estimates because of the over dispersion caused by very largest countries (China, Brazil).

Figure 5. Scatterplot of predicted TB incidence (negative binomial model) vs observed (WHO estimates) in training set (n=214)

Figure 6. Scatterplot of predicted TB prevalence (binomial model) vs observed (WHO estimates) estimates in training set (n=27)
Figure 7. Scatterplot of predicted TB mortality (negative binomial model) vs. observed (WHO estimates) in training set (n=307)

Task 1 and 3
Out of sample predictions could not be made because of the large number of missing data on the complete set of predictors (See plots in Annexes 6 and 8). The missing values patterns need to be thoroughly investigated and stronger assumptions than those made so far and documented in this report need be made to enable predictions. It is important to note however that Missing data were not significantly associated with the outcome. It therefore seems reasonable to assume that data are missing at random, and although missingness affects the ability to make predictions it does not imply that models are biased.

Task 2
Out of sample predictions for 2013 are presented for the binomial model estimating TB prevalence.. Three countries stand out – Central African Republic (CAF), Somalia (SOM) and Niger (NER), which are outliers in terms of the 3 predictor variables, especially very low BCG vaccination rates. See Annex 7 for full details. The scatter plot excluding these three datapoints is also shown below.

Bland-Altman plots of agreement comparing the two measures show that on average, the estimations using the binomial model are higher than the WHO estimates (average difference above “zero difference line”, mean difference=44 cases) but with a random scatter around this difference. Predictions could only be made for 65 out of 74 countries in the set of countries for which data needed to be predicted because of predictor missing data.
Figure 8. Out of sample predictions of TB prevalence for 2013 (left: all data points to predict n=65, right: excluding predictions for CAF, SOM and NER, n=62)

Figure 9. Bland and Altman plots of agreement between WHO TB prevalence estimates and model predictions excluding CAF, SOM and NER (n=62)
12. Conclusions

The preliminary results presented in this document show that predictive models of TB Models could be fitted for all three tasks. Goodness of fit as assessed by deviance tests and cross validations was satisfactory however further refinement of the models and database are necessary before out of sample predictions can be made based on our models. More specifically further analyses include:

- Further explore missing data imputation methods and large amount of zero data in the database.
- Task 1 and 3
  - Include random effects for countries or income status
  - build one model just on high income countries and one just on middle income countries and compare with variables/coefficients
- Task2:
  - Include survey specific variables (e.g. coverage, participation rate) as random effects to filter out nuisance variability induced by these factors
  - Consider fitting two models, one for Asia, one for Africa when further datapoints from recently conducted prevalence surveys which could not be included in our model (Tanzania, Ghana, Indonesia, Malawi, Sudan, Zambia) are available,
- Explore possibility of stratified by age, sex and HIV status.
- Explore possibility of including time trends, and time lags
- Explore methods for the propagation of error
- Explore need and impact of weighing country estimates by total population size.
- Produce estimates for all years between 1990 and 2015.

Data wishlist
- 2010 UNICEF MICS survey reports for Sindh Province), Pakisitan (could only access the reports for survey s in Balochistan and Punjab.
- China NTP data: total number of new laboratory confirmed and new all forms cases in 2008, 2009 and 2010 for the three subnational areas of the 2010 prevalence survey (Eastern, Central and Western china)
- TME data: In the document “global TB database variables 2014” a data file called _finance_bf_services (providing data on the final budget and service utilization forms for 2014) and a file called _finance_exp (providing data on the expenditure for 2013) are listed. These files were not among the csv files that provided to us and we were not able to download these from the WHO site either. We are interested in the variable summing the total funding/expenditure for TB by year by country.
- Original report for India district level surveys conducted in Wardha, Agra (Jalma) and Faridabad districts