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

Prepared by:
Markus C Elze, Elizabeth Allen, Katherine Fielding

Questions for discussion

1. How can we incorporate a complex covariate structure between countries and years into the model?

2. How can we deal with missing data without excluding many observations or, conversely, include explanatory variables that only have data for some observations?

3. How can we model prevalence when we only have 22 prevalence surveys?
WHO consultation on methods to estimate
TB incidence, prevalence and mortality

Predictive statistical ecological model

Markus C Elze\textsuperscript{1}, Elizabeth Allen\textsuperscript{1}, Katherine Fielding\textsuperscript{1}
\textsuperscript{1} London School of Hygiene and Tropical Medicine, London, United Kingdom

Table of contents

Introduction and overall strategy ........................................................................................................... 2
 Objective ............................................................................................................................................... 2
 Status of this document .......................................................................................................................... 2
 Rationale ................................................................................................................................................ 2
 Strategy ................................................................................................................................................ 2
 Incidence and mortality .......................................................................................................................... 2
 Prevalence ............................................................................................................................................ 3
 Mixed effects models .............................................................................................................................. 3
 Discussion ............................................................................................................................................. 3
 Assumptions ......................................................................................................................................... 3
 Model validation ................................................................................................................................... 4
 Advantages and limitations .................................................................................................................... 4
 Modelling approach .............................................................................................................................. 5
 Adjustment of outcome data ................................................................................................................ 5
 Prevalence estimation ........................................................................................................................... 5
 Data sources ......................................................................................................................................... 6
 Predictors .............................................................................................................................................. 6
 Disaggregation ....................................................................................................................................... 7
 Modelling uncertainty .......................................................................................................................... 7
 Software ............................................................................................................................................... 7
 Conclusion ........................................................................................................................................... 8
 Summary ............................................................................................................................................. 8
 Next steps ............................................................................................................................................ 8
 Open questions ..................................................................................................................................... 8
 Acknowledgements .............................................................................................................................. 8
 Appendix ............................................................................................................................................ 9
 Incidence and mortality summary data ................................................................................................. 9
**Introduction and overall strategy**

**Objective**
As part of the MDGs, the WHO Global Task Force on TB Impact Measurement was created to conduct a review of methods for the global estimation of TB incidence, prevalence and mortality from 1990-2015. Our group at the London School of Hygiene and Tropical Medicine was tasked with investigating the use of predictive statistical models using ecological data for estimation. For this work ecological data is taken to be aggregated data, typically on a country-level, that may related to the prevention, diagnosis or treatment of TB, the general state of the health system or even more generally to the wealth, education or infrastructure levels of a country. Given the requirement that the statistical model should be predictive, our goal is not to adjust existing measurements of TB incidence, prevalence and mortality, but rather to estimate them based on reliable data from similar countries together with ecological covariates.

**Status of this document**
This document is a preliminary concept note. Limitations and open questions are discussed and we expect that the discussions at the 3rd meeting of the TB estimates subgroup will be highly beneficial in moving the document forward and fully expect to revise the statistical methods used, modelling strategy and data used as a result of the discussions.

**Rationale**
For some countries high-quality data sources, such as case notifications and vital registration data are available. As a result of efforts by the WHO and others, the quality and availability of these surveillance data have vastly improved. In some countries these data are supplemented using targeted instruments, such as prevalence surveys, inventory studies, and elicitation of expert opinion, to provide a comprehensive picture of TB burden. However, high-quality data are not available for all countries or even years in countries and either adjustment of low-quality data or estimation are required to produce estimates of TB burden for all countries of the world between 1990 and 2015.

**Strategy**
Given that HIV status affects TB incidence, prevalence (via reduced TB duration) and mortality, and MDR status affects prevalence and mortality, we initially propose modelling the HIV-negative, non-MDR population and then we will repeat the analysis for HIV-positive individuals, differentiating by ART treatment status, and again for the MDR TB population.

**Incidence and mortality**
Incidence will be modelled based on case-notification data (219 countries, 4552 country-years) and mortality will be modelled based on vital registration data (127 countries, 2015 country-years). See the Appendix for some summary of these data. Estimation will be necessary for those countries where these data are not available. Data is sparser for HIV-TB
cases as well as MDR TB, but we currently expect that separate models can be fitted for these.

**Prevalence**

22 prevalence surveys from 15 countries are currently available. This number is insufficient to act as the sole source of observations for fitting a statistical model. The WHO has produced estimates for TB duration for HIV-positive and HIV-negative patients, which carry some uncertainty, but can be used to directly estimate prevalence from incidence. We currently do not see a way to improve on this estimation approach using predictive statistical ecological models. Thus we will continue to use this approach to calculate prevalence from incidence estimates from the ecological model. By making assumptions about the case detection rate and survival rate per year lived with (un)treated TB, we can check the consistency of prevalence estimates with the observed TB mortality.

**Mixed effects models**

We propose mixed effects to allow for variation between countries, regions and years,. Mixed effects models, model an outcome $Y_i \in \mathbb{R}$ based on covariates vectors $X_i \in \mathbb{R}^n$ and $Z_i \in \mathbb{R}^m$ using

$$f(Y_i) = \alpha^T \cdot X_i + A_i^T \cdot Z_i + \epsilon_i,$$

where $\alpha \in \mathbb{R}^n$ is a vector of fixed effects regression coefficients (similar to those found in linear regression), $A_i \in \mathbb{R}^m$ is a vector of normally distributed random effects with mean zero, and $\epsilon_i \in \mathbb{R}$ is a normally distributed error term. A link function $f$ can be used as in generalized linear models for log-linear and logistic regression models.

**Discussion**

**Assumptions**

For the purposes of modelling, we will make the following assumptions:

- Incidence, prevalence and mortality estimates based on reliable (WHO-adjusted) case notification, vital registration, or similar data sources are considered to be correct and no further adjustments beyond those done by the WHO will be considered.
- Any data relating to HIV prevalence and treatment will be used as they are and no adjustments will be made.
- Any covariates used for modelling will be considered to be error-free.
- We will consider WHO estimates for case detection, duration, and case fatality and similar rates to be perfectly accurate and to not depend on country or year.

We try make our model robust against violations of these assumption as much as possible. As WHO estimates for case detection, duration and case fatality usually do not vary by country or year, we try to avoid using them if other options are feasible.


**Model validation**

A primary tool for model validation is Leave-one-out cross-validation, where the model fitting is repeated successively leaving out individual countries with available good quality data. The estimates for these countries from the models are then compared to the actual data.

Additional proposed validation steps make use of the fact that we can arrive at estimates using different modelling strategies:

- When modelling an outcome separately for each year and then interpolating missing values (e.g. where modelling was not possible for one year due to low number of available vital registration data) using cubic splines, results should not differ strongly from the integrated model for all years. Furthermore, effect sizes in the models for each year should not differ strongly.
- Incidence, prevalence and mortality estimates should be reasonably internally consistent using WHO estimates for TB duration and case fatality rates.

**Advantages and limitations**

The modelling strategy outlined above has several advantages:

- The approach is conceptually relatively simple and the calculation as well as the role of all variables in the models can be readily understood.
- Due to the use of separate models for incidence and mortality as well as HIV positive and MDR patients, disaggregated data are readily available and the effect of explanatory variables on each of these is clear.
- Despite our assumptions of error-free observations (see above), the mixed models are robust to certain types of measurement error in the dependent and independent variables. Misspecification of the functional form for an independent variable does not invalidate the model, but limits the predictive power of that particular variable.

However, there are also limitations to this approach:

- The ecological statistical models we are creating will offer a very limited mechanistic understanding of TB. Many covariates incorporated into the models will only have a tangential link to TB outcomes. The use of ecological data do not allow for more mechanistic modelling options, such as multi-compartment models to explicitly model transmission and detection rates.
- The simple statistical models we are proposing here cannot easily incorporate multiple data sources, such as enriching case notification data using targeted surveys.
- Incorporation of expert opinion or adjustment of estimates based on additional information is not easily possible.
These models are not bounded in their estimates. While it may be reasonable to assume that case notifications are a lower bound for incidence, this cannot be easily represented in the model.

We cannot easily incorporate independent variables that are only available for a subset of observations (e.g. data from USAID’s Demographic and Health Surveys). The use of imputation techniques to deal with missing data may introduce bias in this setting as data availability may be strongly associated with the TB situation in the country.

The general advantages and limitations of mixed models are discussed in more detail in the statistical literature. More advanced modelling techniques, such as generalized estimating equations, may alleviate some of the drawbacks of mixed effects models. Furthermore, joint models, such as Bayesian models incorporating incidence, prevalence and mortality together with notification and survey data into a single model, would resolve several of the limitations presented here. Further evaluation of these approaches will be made at a later date.

**Modelling approach**

**Adjustment of outcome data**

Outcome data will be adjusted as they are for the WHO GTB database. Vital registration data will be converted to mortality via

\[
\text{TB Mortality per 100,000} = \frac{\text{Reported TB deaths}}{\text{Coverage} \cdot (1 - \text{Proportion ill-defined})} \cdot \frac{100,000}{\text{Country population}}.
\]

Case notifications will be converted to incidence via

\[
\text{TB Incidence per 100,000} = \frac{\text{Case notifications}}{1 - \text{Underreporting}} \cdot \frac{100,000}{\text{Country population}}.
\]

The estimated amount of underreporting is based on surveys conducted by the WHO.

Note that these formulas sometimes lead to missing data even though the primary information is available, for example a lack of coverage data despite reported TB deaths being available. In such cases, sensible imputation strategies will be investigated.

**Prevalence estimation**

Only a very limited amount of direct prevalence measurements are available. Thus, prevalence is calculated using the usual WHO method as the product of estimated incidence and average duration summed over the four subgroups of HIV-positive and HIV-negative patients with notified and non-notified cases. The WHO provides global estimates for the duration for these four categories, so that we need to estimate the incidence for all four groups. We can arrive at the notified cases simply by using the GTB database; the non-notified cases are the estimated incidence minus the notified cases.
Data sources
For this preliminary report, in order to limit the amount of time needed to integrate different data sources, we focused on two databases:

- WHO’s global tuberculosis database
- World Bank Open Data

It would be desirable to integrate additional data sources, such as the Global Health Observatory database, USAID’s Demographic and Health Surveys and United Nations Population Division data, for the final models. In addition, some data available from the current data sources, such as data on HIV prevalence or infant mortality, should potentially be directly obtained from the relevant monitoring body.

Predictors
Candidates for fixed effects include the following subject areas. A preliminary model incorporating some of these variables at a country level will be presented at the meeting:

- Opinion about presence of TB in country (e.g. WHO high burden country, WHO high TB/HIV burden country). From the WHO GTB database.
- Spending on TB (e.g. government / NGO spending on TB, spending split by detection, treatment, prevention, etc., number of TB detection labs by type). Primarily from the WHO GTB database.
- TB policy (e.g. detection available free of charge, treatment available free of charge, availability of GeneXpert tests, certification of labs, MDR testing offered, use of national electronic health records). From the WHO GTB database, currently only available for 2013.
- HIV-related variables (e.g. HIV prevalence, ART coverage). From USAIDS.
- Overall state of the healthcare system (e.g. total government expenditure on health, average distance to next hospital, number of workers at health care facilities, under 5 mortality, immunization percentages). Primarily from the World Bank for now, other data sources are being investigated.
- Overall wealth of country and individuals (e.g. World Bank income group, GDP, human development index). Primarily from the World Bank.
- WHO estimates (e.g. proportion of MDR TB, case detection rate, treatment success rates). From the WHO GTB database and publications.
- Other social and environmental factors (e.g. underweight children, literacy, completion of secondary education). Primarily from the World Bank for now, other data sources are being investigated.

These variables are usually available for each year from 1990 until at least 2012 for most countries of the world. Covariates for some fixed effects variables may need to be allowed to
vary by region and / or time, but the available data will likely only allow this for a small number of covariates.

The use of data based in part on previous estimates of TB burden (such as the classification as high-burden country) should be discussed. While this makes the estimation somewhat circular, there seems to be good agreement on these classifications in the literature. Due to this good agreement and the high relevance of these classifications for TB assessment, it may be inadvisable to ignore these data.

Random effects are based on geography and time and capture components of the unexplained variation in the TB burden. If the model explains the observed data well, both the country- and year-level random effects should remain fairly small:

- Region
- Country (random effects nested within region)
- Year

Disaggregation
Disaggregation of estimates by HIV status is an automatic result of the modelling strategy. Disaggregation by age and sex is possible by modelling the various subgroups separately (using disaggregated case notification and vital registration data). As yet we have not investigated the data availability for this.

Modelling uncertainty
While it is possible to calculate confidence intervals for predictions based on the variance-covariance matrix of the parameters and a normal approximation, bootstrap methods are increasingly becoming the preferred option for mixed models. A bootstrap approach that allows calculation of confidence intervals for both fixed effects and predictions is implemented in the software we use.

Software
We used the R Language and Environment for Statistical Computing, version 3.0.3 64-bit, with the lme4 package for linear mixed-effects models using Eigen and S4, version 1.1-6. The mixed models presented here can be readily fitted using the lmer (linear mixed models) or glmer (generalized mixed effects models) commands. We used the default numerical optimization settings unless we encountered numerical issues.
Conclusion

Summary
This document outlined a broad strategy for the predictive statistical ecological modelling of TB burden using mixed models based on high-quality surveillance data, but does not provide a specific list of variables used. Initial models will be presented at the 3rd meeting of the TB estimates subgroup.

Next steps
The incorporation of additional databases and covariates together with an exploration of the correct functional form of the models and appropriate transformations of the covariates will be the next step. As most covariates are only tangentially linked to TB burden, the performance of the models may remain limited until highly predictive covariates have been identified. The use of alternative modelling approaches is currently being investigated in order to accommodate a more complex correlation structure between outcome variables (e.g. correlation based on temporal and geographical proximity).

Open questions
- Can intelligent transformations or discretizations of covariates improve the variation explained by the model?
- Can we incorporate boundaries in our models (e.g. estimated incidence must be greater than case notifications)?
- Can we specify the correlation between countries and years more clearly using generalized estimating equations or similar techniques?
- Can we produce a joint model that can incorporate case notification, vital registration and prevalence survey data as well as expert opinion at the same time?
- Can we use covariates other than the ones mentioned above?
  - Are data available for enough countries? Can we deal with missing data (e.g. by creating a submodel just for those countries where data are available)?
  - Are data sufficiently easily accessible and usable?
- Is there a way to model prevalence using the available prevalence survey data? Are there other sources of prevalence data that could be incorporated?
- Are there political limitations to the use of certain data or modelling approaches?

Acknowledgements
We would like to acknowledge the support of several groups and individuals in getting this far in the available time:

- The TB Modelling Group at LSHTM as well as the wider TB community at LSHTM
- The WHO Global Task Force on TB Impact Measurement
- Carel Pretorius (Avenir Health)
Appendix

Incidence and mortality summary data

We are able to use all 4552 country-years of case notification data available to us. Basic summaries are as follows.

Absolute number of non-HIV TB case notifications (new cases or relapse):

<table>
<thead>
<tr>
<th>Minimum</th>
<th>1st quartile</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Quartile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>262</td>
<td>2,385</td>
<td>22,130</td>
<td>10,480</td>
<td>1,555,000</td>
</tr>
</tbody>
</table>

Number of non-HIV TB case notifications (new cases or relapse) relative to 100,000 citizens:

<table>
<thead>
<tr>
<th>Minimum</th>
<th>1st quartile</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Quartile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>14</td>
<td>40</td>
<td>73</td>
<td>88</td>
<td>855</td>
</tr>
</tbody>
</table>

In the current analysis, we have 1874 country-years of vital registration data (due to some missing coverage estimates etc.).

Absolute number of non-HIV TB deaths:

<table>
<thead>
<tr>
<th>Minimum</th>
<th>1st quartile</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Quartile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>27</td>
<td>151</td>
<td>1,604</td>
<td>744</td>
<td>100,500</td>
</tr>
</tbody>
</table>

Number of non-HIV TB deaths relative to 100,000 citizens:

<table>
<thead>
<tr>
<th>Minimum</th>
<th>1st quartile</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Quartile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>205</td>
</tr>
</tbody>
</table>