Predictive statistical modelling to estimate TB disease burden

Markus C Elze, Elizabeth Allen, Katherine Fielding

Improving health worldwide

www.lshtm.ac.uk
Objective

- Predictive statistical models using ecological data
- High-quality data sources available for some countries
- For others, targeted instruments offer additional information
- Data typically aggregated on a country level
- Assuming no high-quality data available to be adjusted, must predict from similar countries and years
- This means trusting high-quality data sources and using them to predict data for countries where such sources are not available
Overall strategy

• Initially exclude HIV and MDR
• Model incidence based on case-notification data (219 countries, 4552 country-years)
• Model mortality based on vital registration data 127 countries, 2015 country-years)
• HIV and ART treatment status assumed to influence incidence, prevalence and mortality while MDR assumed to influence prevalence and mortality only -> create separate models for these
Mortality per 100,000

Number of observations

Reported TB Mortality per 100,000
Case notifications per 100,000
Modelling prevalence

• Only 22 prevalence surveys from 15 countries currently available -> too few for statistical modelling
• Use WHO estimates for TB duration (subgroups HIV\(^+\) and HIV\(^-\) with notified & non-notified cases) instead and estimate prevalence from incidence
• Verify estimates against 22 available prevalence surveys and, using estimates for case detection rates, case notification data
• Prevalence remains a significance open question for estimation
Prevalence per 100,000

TB Prevalence per 100,000

Number of observations
Mixed models

- Classical regression models are fairly robust and suitable to a wide range of applications
- Also want to allow grouping by countries, regions and years

\[ f(Y_i) = \alpha^T \cdot X_i + A_i^T \cdot Z_i + \epsilon_i \]
Assumptions

- Outcomes (incidence, prevalence, mortality) are assumed to be error-free after standard WHO adjustment
- Will use HIV data from UNAIDS without further adjustment
- All covariates use for modelling assumed to be error-free
- Will use some WHO estimates for case detection rate, TB duration, and case fatality rates without further adjustment

Model does have some robustness against violations of these assumptions
Model validation

• Leave-one-out cross-validation allows observing both accuracy of predictions and changes in effect sizes
• In addition, arrive at estimates using different calculations (e.g. mortality also via incidence and CFR) or by cubic splines regression from different years
• Usual checks for violation of modelling assumptions, e.g. normally distributed errors
Advantages

• Conceptually simple
• Established & robust statistical approach
• Flexible in both outcome and covariate choices
• Disaggregated data readily available due to separate models
Limitations

• Limited mechanistic understanding of TB
• Cannot predict multiple outcomes at once
• Cannot easily build upon highly informative data such as case notifications or expert opinions
• No bounding
• Problems modelling outcomes or using covariates with many missing observations

More advanced approaches, such as Bayesian models, could solve some of these issues
Adjustment of outcome data

- Use standard WHO approaches for adjustment and use the adjusted data to predict for countries for which data are unavailable

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

- These calculations may lead to missing data e.g. when coverage is unavailable
Data availability - mortality
Data availability – case notifications

Number of observations

Number of available years per country
Data sources for predictors

• Initial data sources
  – WHO's global tuberculosis database
  – World Bank Open Data

• Other desirable data sources
  – Global Health Observatory
  – USAID’s Demographic and Health Surveys (limited)
  – United Nations Population Division
  – UNAIDS
Candidate predictors examples

• Opinion about presence of TB in country (WHO GTB database)
  – WHO high burden country
  – WHO high TB/HIV burden country

• Spending on TB (WHO GTB database)
  – Government / NGO spending on TB, split by detection, treatment, prevention, ...
  – Number of TB detection labs by type

• TB policy (WHO GTB database, currently 2013 only)
  – Detection / Treatment available free of charge
  – Availability of GeneXpert tests, Certification of labs, MDR testing offered
  – Use of national electronic health records
Candidate predictors examples

• HIV-related variables (UNAIDS)
  – HIV prevalence
  – ART coverage

• Overall state of the healthcare system (World Bank for now)
  – Total government expenditure on health
  – Average distance to next hospital
  – Number of workers at health care facilities
  – Under 5 mortality
  – Immunization percentages

• Overall wealth of country and individuals (World Bank)
  – World Bank income group
  – GDP
  – Human development index
Candidate predictors examples

• WHO estimates (WHO GTB database and publications)
  – Proportion of MDR TB
  – Case detection rate
  – Treatment success rates

• Other social and environmental factors (World Bank for now)
  – Underweight children
  – Literacy
  – Completion of secondary education
Use of previous estimates

- Some data, e.g. classification as high burden country, are based on previous estimates
- In the absence of data for adjustment, these are highly informative for modelling
- Should such data be used or should this kind of circular regression be avoided?
Further modelling considerations

• Can easily disaggregate by HIV and MDR using separate models
• Disaggregation by age and sex is possible using separate models as well (given data availability)
• Uncertainty can be modelled using bootstrap methods (preferable to standard variance-covariance matrix based approaches)
• Software used is R using lme4 package
Conclusions

• Broad strategy so far
• Initial results show fairly low accuracy
• Initial models done using linear models instead of e.g. logistic models
• Many detail questions regarding data availability and modelling strategy still open
Next steps

• Incorporate additional databases
• Check if different functional form should be used
  – For covariates
  – For outcomes
• Modelling of incidence, prevalence
• Data sources for HIV and MDR models
• Model checking and validation
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?
Open questions

• 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?