Hack TB Methods

P. MacPherson, D. Shaweno, P.J. Dodd

Reproducibility

Our analyses have been conducted in R using literate programming principles and have generated html reports which can be scrutinized for logic and results. These are numbered by order of implementation: 0x for data preparation; 1x for statistical modelling; 2x for processing final results. The code is under a CC-BY 4.0 license.

Data sources and linkage

Spatial data, census data and prevalence data

Census data provided as PDF tables were extracted into CSV files and merged with shape file data obtained from The Humanitarian Data Exchange, administered by United Nations Office for the Coordination of Humanitarian Affairs (https://data.humdata.org/dataset/pakistan-administrative-level-0-1-2-and-3-boundary-polygons-lines-and-central-places), which had more complete linkage than the provided shapefiles. This merge resulted in 131 districts. We excluded 18 districts located in the North East of the country (in the provinces of Pakistan-administered Kashmir, India-administered Kashmir, and Gilgit Baltistan). To permit linkage between datasets, we aggregated census data across six administrative areas that comprise Karachi City (Karachi Central, Karachi East, Karachi South, Karachi West, Korangi, Malir). We calculated each districts male:female ratio for 2017, and the total population growth between 1998 and 2017 for inclusion as covariates potentially correlated with TB prevalence. This data was merged with the provided TB prevalence survey data.

DHS data

Demographic and Health Surveys (DHS) data from the 2017-2018 survey of Pakistan was applied for and obtained. Accounting for survey design, the fraction of the demography of each district was computed by age and sex category using the household recode dataset. In addition, district-level survey averages of individual and household-level covariates with a subjective *a priori* potential of correlating with TB were computed. These included: mean household size, DHS wealth score (SES), indoor smoke, smoking, body mass index (BMI), weight-for-age Z-score (WAZ), prevalence of vaccination, prevalence of BCG vaccination, prevalence of chronic cough, self-reported tuberculosis (TB) prevalence, distance to nearest healthcare facility, and awareness of TB.

Notification data

Age-and sex-category specific bacteriologically confirmed TB notifications were available for years from 2009- 2012. The age-and sex-stratified per-capita notification rates were calculated by dividing the age-sex stratified bacteriologically confirmed case counts by the corresponding age-sex stratified population denominators generated as a product of proportion of age-sex stratified population (from census) and total district population for a given year (provided with notifications). We were unable to link 23 districts in notification data which were subsequently.

Modelling of prevalence data

Model types and implementation

We considered a variety of Bayesian hierarchical regression models for prevalence. These included informative priors on regression coefficients (analogous to frequentist ridge regression) that help avoid overfitting. All models were fitted using Markov chain Monte Carlo approaches using Stan. Some models included conditional autoregressive (CAR) priors to capture spatial effects: i.e. the propensity of neighbouring districts to be more similar to each other. Our notion of ‘neighbour’ was based on a network, with each district joined to its geographically-nearest four other districts. Missing data were typically replaced with means.

id	*data likelihood*	*covariates*	*structure*	*priors*	*LOO***
rm0	Binomial by cluster	none	MV normal regression	normal for α, normal for α	logS = 0.0179905; MSE = 1.22-05; RSE = 1.4445483
rm1	Binomial by cluster	age x sex	MV normal regression	normal for α, normal for β	logS = 0.0189252; MSE = 7.90e-06; RSE = 2.2372568
rm2	Binomial by cluster for age x sex	none	MV normal regression	normal for α, LJK for sex/age covariance,	logS = 0.0183224; MSE = 7.90e-06; RSE = 1.4451301
rm3	Binomial by cluster for age x sex	Census covariates (pop. growth, m:f ratio)	MV normal regression	LJK for sex/age covariance, normal for α, normal for β	logS = 0.0185076; MSE = 8.20e-06; RSE = 1.4451703
rm4	Binomial by cluster for age x sex	Census covariates (pop. growth, m:f ratio); DHS covariates with $\|r\| > 0.1$*	MV normal regression	LJK for sex/age covariance, normal for α, normal for β	logS = 0.01855328; MSE = 7.90e-06; RSE=1.495038
car0	Binomial by cluster for age x sex	none	MV normal regression with spatial RE	MV Leroux CAR for spatial REs; LKJ for sex/age covariance	logS = 0.0172895; MSE = 3.10-06; RSE = 0.7416439
car1	Binomial by cluster for age x sex	DHS covariates with $\|r\| > 0.1$*	MV normal regression with spatial RE	MV Leroux CAR for spatial REs; LKJ for sex/age covariance; normal for β	logS = 0.0169187; MSE = 2.60-06; RSE = 0.6204400
car2	Binomial by cluster for age x sex	Per capita sex x age notifications	MV normal regression with spatial RE	MV Leroux CAR for spatial REs; LKJ for sex/age covariance; normal for β	logS = 0.0169627; MSE = 2.30-06; RSE = 0.5674514
car3	Binomial by cluster for age x sex	Above covariates combined	MV normal regression with spatial RE	MV Leroux CAR for spatial REs; LKJ for sex/age covariance; normal for β	logS = 0.016953; MSE = 2.00-06; RSE = 0.4841486

Abbreviations: β = regression coefficients; CAR = conditional autoregressive; LKJ = Lewandowski-Kurowicka-Joe; LogS = log-score; LOO = leave-one-out; MSE = mean squared error (for TB in district samples); MV = multivariate; RE = random effects; RSE = relative squared error.

* I.e. with cluster-level correlation: SES, HHS, indoor smoke, BMI, WAZ, vaccination coverage, prevalence of cough, distance to health facility. ** Smaller is better.

Model evaluation and selection

In order to evaluate the predictive performance of the models, we used a log-scoring rule (i.e. the expected value of $-\log(Prob(observation))$ when predicting the probability of sampled individuals in a district having TB based data only in other districts), as well as the required MSE and RSE. We used of the approximate method of Vehtari et al for computation. Visual inspection (see plots/) also informed the final choice of model (highlighted).

Calculation of district TB prevalence

The highlighted model was above was combined with demographic data to give age/sex/district raw estimates (uncertainty intervals based on MCMC-approximated predictive distributions). To maintain consistency with published estimates, raw estimates central estimates and uncertainty were scaled to match. Per capita results were projected using the WHO estimated trend (and uncertainty) and updated district sizes. These were aggregated over sex/age (see CSV files).

Strengths and limitations of approach

We used models that could penalize over-fitting and assessed predictions visually and via LOO metrics. Models borrowed strength between districts and category, also helping to avoid over-fitting by shrinking towards shared means. We worked independently on models and adopted a disciplined approach to version control and documentation. Our approach enforced consistency with previously published national estimates.

However, the data are sparse and uncertainties large. We only investigated k=4 nearest neighbour notions of adjacency. Inference for the high-dimensional CAR models was suboptimal and could be improved given more time. It is possible more informative priors could stabilize the CAR models with more covariates, whose predictions where TB data was absent appeared less realistic. We felt no covariates were strongly-enough associated with TB to base projections on, and so adopted a conservative approach to this problem.