## **Calibration methods**

Values of model parameters are inferred from fitting the dynamic model simultaneously to country-level annual disease incidence and mortality estimates (2000-2014) as well as disease incidence estimates from mining communities in 2008 from the Thibela TB study. For each data measurement , where *j* and *t* represent the type of data and time, respectively, we assume errors are normally distributed. The likelihood of the dynamic model input parameter vector , can then be defined as

or more conveniently in terms of log-likelihood (up to an additive constant) as

Here represents the variance associated with the uncertainty in the data and *s* represents the random number seed used to initialize the dynamic model such that represents one statistically correct trajectory of the dynamic model with input parameter .

As we seek to give the mining incidence data comparable importance in our calibration to country-level data and simultaneously account for mining incidence data and overall country time trends, we assign a calibration weight to each type of data in the calibration. Assuming data points are independent and incorporating the calibration weights, we arrive at the overall log-likelihood function, which can be expressed up to an additive constant as:

We choose weights for country-level incidence, country-level mortality, and mining incidence as 2/15, 2/15, and 1, respectively. Noting we have 15 time points for country level mortality and incidence from 2000-2014 and one measurement of mining incidence at 2008, we divide the data into two periods (2000-2007 and 2008-2014) and give country-level incidence and mortality and mining incidence over each period equal contributions in terms of geometric-mean likelihood over the time periods.

The calibration procedure is carried out in an iterative manner using Incremental Mixture Importance Sampling (IMIS) using the likelihood function. We conducted seven iterations, with each iteration consisting of 200 sample points in parameter space for a total of 1400 points. In the first iteration 200 points in parameter space were chosen independently for each parameter via Latin Hypercube Sampling from prior distributions on the parameters for relative mine transmission rate and susceptibility to re-infection. We assume the prior distribution for susceptibility to re-infection to be uniform, Susc\_p ~ U (0, 1). Similarly we assume the relative mine transmission rate prior distribution to be U (1, 5), representing transmission ranging from that of one to five times the overall country-level rate. We denote as the prior probability density and as all of the sample points up to and including the *j*th iteration. We compute the importance weights for the first iteration of the algorithm as:

Points for the next iteration are chosen by constructing a multivariate normal distribution centered at the value which maximizes with a weighted covariance matrix constructed from the other points in the sample.

Denoting this multivariate normal distribution as we then choose 200 new points from this distribution and compute their likelihood under the dynamic model. We then construct the importance weights for all of the points up to and including the second generation as:

Where *q* is the density of a mixture distribution consisting of the prior distribution and the multivariate normal distribution , and the mixture is weighted by the number of points in each iteration in proportion to the total. In this case since we have an equal number of points in each iteration

Points in subsequent iterations are chosen in the same manner. For the *j*th generation we define a multivariate normal distribution centered at the input parameter with the maximum importance weight with a weighted covariance matrix constructed from the 200 points closest in terms of Mahalanobis distance with respect to the prior distribution and new importance weights are then computed with . In the seventh and final iteration, we computed the sample posterior distribution of by sampling without replacement from the importance weights corresponding to each .

## **Computing population attributable fraction of incidence**

We compute the mean incidence measured over a time window attributable to transmission from a given group occurring over the time window . Here we allow for the transmission time window to extend further into the past as compared to the incidence window to account for the delay from infection to disease.

Given a point in our epidemiological input parameter space and an input random number seed for the stochastic model *s*, we denote to be the cumulative disease incidence, I.e., total number of cases, of the model over the time period . Similarly, for the counterfactual model where we have artificially removed transmission from group *G* beginning at time we denote the corresponding cumulative incidence to be. For the stochastic simulation it follows that the attributable fraction is computed as

We then estimate the mean attributable fraction as the sample mean over a large set of independent identically distributed stochastic simulations defined by the set of input random number seeds

= <

and compute the associated confidence intervals. Analogously we can also estimate the fraction of incidence from recent transmission, i.e, , incidence from any transmission occurring over the time window , attributable to a group *G* as

,

where *All* denotes all groups in the model. Table 3 provides estimates and confidence limits of incidence fractions attributable to mining, and , for the estimated most likely input parameters. Here we define years corresponding to a five-year incidence window beginning in 2014 and recent transmission as occurring after 2012. The sample mean and confidence limits are based on N= 200 stochastic simulations.