## **Calibration methods**

Values of model parameters were 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 were normally distributed. The likelihood of the dynamic model input parameter vector , was then 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 .

We sought 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; therefore, we assigned a calibration weight to each type of data in the calibration. Assuming data points were independent and incorporating the calibration weights, we arrived at the overall log-likelihood function, which could be expressed up to an additive constant as:

We chose weights for country-level incidence, country-level mortality, and mining incidence as 2/15, 2/15, and 1, respectively. Because we had 15 time points for country-level mortality and incidence from 2000-2014 and one measurement of mining incidence at 2008, we divided the data into two periods (2000-2007 and 2008-2014) and gave 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, where each iteration consisted 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 assumed the prior distribution for susceptibility to re-infection to be uniform, Susc\_p ~ U (0, 1). Similarly we assumed the relative mine transmission rate prior distribution to be U (1, 5), representing transmission ranging from one to five times the overall country-level rate. We denoted as the prior probability density and as all of the sample points up to and including the *j*th iteration. We computed the importance weights for the first iteration of the algorithm as:

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

We denoted this multivariate normal distribution as then chose 200 new points from this distribution and computed their likelihood under the dynamic model. We then constructed 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 was weighted by the number of points in each iteration in proportion to the total. In this case since we had an equal number of points in each iteration

Points in subsequent iterations were chosen in the same manner. For the *j*th generation we defined 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 . New importance weights were 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 .

## **Calibration results**

The estimated joint posterior probability density for the calibrated model parameters, and is shown in Figure S2A. Marginal densities for the parameters are shown in Figure S2B and S2C. The distributions are unimodal and strongly peaked, with a maximum *a posteri* value located at and . Because protection from future disease in previously exposed individuals in the model works through prevention of re-infection, the estimate of protection from re-infection was similar to estimates of BCG protection against active disease of 0.58 (CI95: 0.35-1.01) (Mantgani et al. 2014).

Posterior predictive distributions for disease incidence and mortality are shown at the country level (Figures S3A and S3B) and specifically in the mining population (Figure S3C and S3D). Note that country level values were computed as a population-weighted average of the four groups in the model. The distributions recapitulated the time trends in both mortality and incidence and were within the bounds of the WHO estimates indicated by the error bars. Similarly predicted incidence in the mining community recapitulated the 2008 Thibela TB estimates. In addition, annual incidence in the early 2000s ranging from ~2 000 to 6 000 per 100 000 was consistent with other mining study populations from this period. Mining community TB disease mortality was predicted to be ~1% per annum during this period which was also consistent with mine records showing overall mortality of ~1% and 4% sent home due to illness.

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

We computed the mean incidence measured over a time window attributable to transmission from a given group occurring over the time window . Here we allowed 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 denoted to be the cumulative disease incidence in terms of cases of the model over the time period . Similarly, for the counterfactual model where we had artificially removed transmission from group *G* beginning at time we denoted the corresponding cumulative incidence to be. For the stochastic simulation it followed that the attributable fraction could be computed as

We then estimated 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 computed the associated confidence intervals. Analogously we also estimated 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 defined years corresponding to a five-year incidence window beginning in 2014 and recent transmission as occurring after 2012. The sample mean and confidence limits were based on N=200 stochastic simulations.