Supporting information for Murray (January 29, 2002) *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.022618299.

Model Structure

We describe here a general model of tuberculosis (TB) transmission in which the identity of specific circulating strains of *Mycobacterium tuberculosis* is known. Each individual in the model is characterized by age category, household, neighborhood, HIV infection, and latent TB infection. These variables were assigned randomly to individuals based on their overall distribution in the population. The model is run in time steps of 1 week, during which a series of events may occur. Each of these events occurs as the stochastic realization of specified probability; where it is not otherwise stated, individuals at risk for an event were assigned random numbers between 0 and 1, and those with a number equal to or less than the "rate" specified for a process were "successes" for this process, whereas the remainder were "failures." The distribution of these outcomes therefore followed a binomial distribution with success probability equal to the specified rate.

(*i*) A person with latent TB may reactivate at an age-specific and HIV-dependant rate and enter the pool of cases of active TB. Infected people develop either pulmonary or extrapulmonary disease depending on age and HIV status.

(*ii*) TB may be transmitted from an individual with pulmonary TB to a susceptible or partially susceptible person in the population. Infection may occur within a household, within a neighborhood, or between neighborhoods. The transmission probabilities are different for each of these possibilities and are bounded by estimates derived from the literature on the probability of TB infection after exposure in different social settings. Within these bounds, transmission probabilities are varied to obtain an approximate fit to the incidence of disease in each of nine populations modeled. Each person in the population is assigned a transmissibility factor that remains fixed throughout the simulation. Thus, the factors that impact the likelihood that a transmission event will occur are modeled as host characteristics and are not transmitted from one individual to another with the infecting strain. These host-specific transmission factors are drawn randomly from a widely dispersed beta distribution. This distribution was chosen to represent the full range of clinical presentations of TB and their impact on transmission. Thus, individuals assigned a low transmissibility factor represent those with a low probability of transmission such as smear-negative pulmonary cases and those with limited social mixing. Those with a high transmissibility factor correspond to cases with a very high probability of transmission such as those with laryngeal TB and/or those with a large number of social contacts. The probability of transmission to a latently infected individual is modified also by the age-specific immunity afforded by previous infection with TB to the individual at risk for infection.

(*iii*) People infected during the time step may develop primary active TB at age- and HIV-dependent rates and enter the pool of cases of active disease. Those who do not develop primary disease enter the pool of people at risk for primary disease. Those not removed from the pool remain at risk for 2 years.

(*iv*) People in the pool of cases of primary disease may leave this pool and enter or reenter the state of latent TB infection.

(*v*) Those with active disease are assigned a duration of disease, during which they remain in that pool. Individual disease durations are drawn from a normal distribution.

The model was run for 208 time steps (4 years), during which time the population was considered closed with no births, deaths, or migration. One case was forced to reactivate during the first time step. To calibrate the model and ensure that its assumptions were consistent with observed data, cluster distributions and measures of incidence were generated for nine different settings for which estimates of TB incidence are available. These included seven countries and two U.S.-based prisons. We fit the model to the estimated incidence of TB in the nine settings modeled by allowing the parameters of the beta distribution to vary to give transmission probabilities within the bounds specified. At the end of each simulation, final output was generated including the annual incidence of infection, annual incidence of all active disease, annual incidence of pulmonary disease, and prevalence of infectious cases. The number of cases of a specific strain of *M. tuberculosis* was counted, and a subsequent count was made of the number of a cluster of a specific size. Summary statistics describing these cluster distributions were reported including mean and maximum cluster size and the proportion of unique isolates in the data set. Simulations for each setting were repeated 20 times, and Monte Carlo means and 95% confidence intervals were calculated for each simulation setting.

Variables

Variables characterizing individuals in the modeled populations were based on demographic and health indicators for the nine regions chosen for analysis (Table 2). These regions were selected to capture both a wide range of disease burden as well as different levels of HIV infection and demographic features. Because many of the parameters that describe TB epidemics are age-dependent, individuals were characterized by three age categories derived from country-specific age structures projected by the United Nations Development Program for the year 2000 (1). These categories included people under 11, 11-20, and over 20 years old to accommodate age-specific differences in rates of reactivation and primary disease as well as the proportion of cases of disease that are infectious. For most scenarios, the age-specific prevalence of latent TB in the population was based on projections of the annual risk of infection and the trends in the annual risk of infection for each area reported by the World Health Organization (WHO) global surveillance and monitoring project (2). The age-specific prevalence of latent infection was back-calculated from the estimated annual risk of infection and the trend in the annual risk of infection. Annual risks were established for each calendar year during which the current population had been alive, and the cumulative incidence of infection in each age group by the year 2000 was calculated. Assuming no differential mortality among the TB infected, the prevalence of latent infection is equivalent to the cumulative incidence of infection in each age cohort. The prevalence of latent infection in the three age categories and the overall population then was based on the relative contribution of

that age group to the age categories and total population. For the two prison populations, the prevalence of latent infection with TB was obtained from direct observation of purified protein derivative (PPD, tuberculin skin test) positivity (T. M. Hammet, P. Harmon, and W. Rhodes, unpublished data). Country-specific estimates of HIV sero-prevalence from 1997 were taken from those published by the Joint United Nations Programme on HIV/AIDS and WHO (3). For simplicity, HIV and TB were considered to be acquired independently, and thus the risk of coinfection is just the product of the risk of each infection in each age group. HIV prevalence within prisons was obtained also through direct observation (T. M. Hammet, P. Harmon, and W. Rhodes, unpublished data).

The effects of various control measures were incorporated into the models by estimating the use and effectiveness of bacillus Calmette--Guérin (BCG) vaccination, chemoprophylaxis among household contacts, and the duration of infectiousness for each scenario. Levels of BCG coverage among children were taken from country estimates published by the Expanded Programme for Immunization/WHO (4). Estimates of coverage among adults generally were not available and were based on reports of BCG scars in adults in areas with similar socioeconomic profiles (5). BCG efficacy was estimated at 50% (6). Good data on the use of chemoprophylaxis among household contacts is not available; nonetheless, the current debates on chemoprophylaxis among the HIV-infected in developing countries make clear that preventive therapy rarely is used outside Europe and the U.S. (7). Levels of household chemoprophylaxis used for the U.S. and Europe were based on speculation. Approximations of the mean duration of infectiousness for each country was obtained by dividing the prevalence of smear-positive cases by the incidence of smear-positive cases for each country as estimated by the WHO Global Surveillance and Monitoring Project (2).

Parameters

Estimates of the parameters summarized in Table 3 were taken from the literature and selected to correspond to those used in standard differential equation models of TB epidemics (8, 9). We defined primary TB to be a disease that occurred within the first 2 years after infection. Rates of primary and reactivation disease as well as the fraction of extrapulmonary disease in immunocompetent hosts were specified for three age categories. Although the immunity to a subsequent infection conferred by previous infection with TB has not been established in epidemiologic studies, the age-specific estimates used in this model were derived from studies that fitted epidemic models to incidence data (8). In the absence of reasonable data, we chose age-independent rates of primary disease, extrapulmonary disease, and reactivation disease in the HIV-infected. Importantly, as noted by Dye *et al.* (10), the parameter estimates available in the literature are based on analyses of the natural history of TB in developed countries; the presentation of disease may depend on a variety of factors such as genetic profile of the human host, concurrent infection, and previous exposure to other infectious diseases. Because these factors may be markedly different in the less-developed areas in which the burden of TB is high, these estimates should be considered highly provisional until better data are obtained.

Table 4 summarizes the results of the validation exercise by comparing consensus estimates to modeled incidence and modeled statistics on the annual risk of infection, mean and maximum cluster size, and proportion of isolates found to be unique.

Determinants of Cluster Distribution

We assessed the impact of determinants on the frequency distribution of cluster size by varying levels of specific variables over reasonable ranges while holding the other factors constant. These factors were varied in the moderate-burden setting of Algeria. We chose the following factors for analysis: age structure of the population, prevalence of latent disease, duration of infectiousness, BCG coverage, use of chemoprophylaxis, and HIV prevalence. Prevalence of latent infection was calculated by applying a range of annual risks of infection from .004 to .02 and a declining trend in the annual risk of .04 to the age structure of Algeria to estimate age-specific prevalence of infection. The prevalence reported in Table 2 is the overall prevalence of infection obtained by summing over age groups. Summary statistics for the cluster distributions were reported for each level of these factors, again based on Monte Carlo means and ranges for 20 simulation runs.

Limitations of the Model

The model described in this study was designed to generate cluster distributions and to identify factors that impact these distributions. Many of the parameters used in this and other models of TB transmission have been based on limited observations in developed countries and may not characterize the behavior of TB in other settings adequately. Similarly, often there was little information available on some of the variables in the models, and the estimates used may vary substantially from their true values. In particular, the results of the models are highly sensitive to the parameters that describe the natural history of TB in the HIV-infected and to the overall and age-specific prevalences of latent TB infection. Because tuberculin skin test (TST) surveys traditionally have been used to estimate the annual risk of infection and have been limited to children, estimates of overall prevalence are necessarily based on diverse methods of imputation that have not been validated by comparison to the results of TST data in adults (2). The finding that 89% of cases are clustered in Afghanistan, for example, seems inconsistent with most molecular epidemiologic studies. This discrepancy may reflect the fact that the prevalence of latent infection would have to be higher than that reported to attain the observed incidence of disease without overestimating rates of ongoing transmission.

In addition to these potential problems, the simulations presented here were based on a population of 40,000 people. Because many, if not most, “real-world” mixing populations are much larger than this, we would expect larger clusters and altered dynamics in real communities.

Previous studies have noted that the estimates of clustering in a community will depend on the molecular clock of the marker used, the duration of the study, and rates of

immigration and emigration of cases (11–13). The model is not constructed to capture the behavior of specific molecular markers or to address issues such as the effect of immigration, emigration, or sampling strategies on estimation of cluster size. We therefore have not modeled genetic change in the molecular marker over time, assuming instead that the generic marker is stable over the 4-year simulation but that it is adequately polymorphic and that all latent strains are different. Cluster distributions in real populations therefore will differ from those generated by this model because the prevalent latent cases would not all have been caused by different strains but also would be grouped into clusters defined by the original source of their infection. We thus ignore the possibility that identical strains may reactivate during the time period of the study. Although this is the assumption made by most molecular epidemiologic studies, this simplification is reasonable only if the latently infected strains are diverse. Furthermore, the model statistics do not take into account any clusters that might already have been circulating at some arbitrary start-up time and persisted into the time period modeled. Although this simplification may alter the proportions of clustered cases, it should not impact the observed effect of specific determinants of cluster distribution.

We also simplified the model by assuming that many of the variables were distributed independently within the population. Although it is clear that factors such as HIV and TB infection often are jointly distributed, country-specific data on these associations are not available, and we therefore lack the information with which to structure a more complex but realistic model. Finally, this model was run for a 4-year period and thus captures only the short-term effects of changing specific determinants on cluster distribution. Long-term TB dynamics are determined largely by changes in the prevalence of latent TB infection, i.e., in the number of potential “source” cases for transmission. By looking only at short-term effects, we essentially treat the class of latently infected individuals as fixed and do not consider the ultimate effect of changing variables on the pool of source cases. Changes in HIV prevalence, intervention practices, and demography, however, may occur much more quickly than significant changes in the pool of latently infected people. Further, because many of the factors that determine the prevalence of latent infection have changed sporadically over the lifetimes of people in the population, it is not clear how relevant the long-term impact of these factors may be. By using consensus estimates of latent infection in populations as the initial conditions in this model, we circumvent the problem of arriving at a set of parameters that correctly describes how these conditions came to be. Such estimates are unlikely to be precise, especially in areas of rapid demographic and social change. We believe this approach may provide insight into the interpretation of cluster distributions that could not be obtained by assessing the impact of these variables over longer periods. Although the limitations of the model may limit its usefulness in describing “true” cluster distributions in real population, these analyses provide perspective on how transmission dynamics affect the distribution of clusters and thus the data from molecular epidemiologic studies of *M. tuberculosis*.

References:

1. United Nations Population Division (1999) *World Population Prospects: the 1998 Revision* (United Nations, New York), Vol. 1.

2. Dye, C., Scheele, S., Dolin, P., Pathania, V. & Raviglione, M. C. (1999) *J. Am. Med. Asoc.* **282**, 677–686.

3. UNAIDS (1997) *Report on the Global HIV/AIDS Epidemic* (United Nations, Geneva).

4. World Health Organization Vaccine Preventable Disease Monitoring System (1999) *Country Immunization Profiles: Global Survey 1999* (United Nations, Geneva).

5. Fine, P. E. M., Ponnighaus, J. M. & Maine, N. (1989) *Bull. W. H. O.* **67,** 35–42.

6. Colditz, G. A., Brewer, T. F., Berkey, C. S., Wilson, M. E., Burdick, E., Fineberg, H. V. & Mosteller, F. (1994) *J. Am. Med. Assoc.* **27,** 698–702.

7. Porter, J. D. (1996) *J. Antimicrob. Chemother.* **37,** Suppl. B, 113–120.

8. Vynnycky, E. & Fine, P. E. (1997) *Epidemiol. Infect*. **119,** 183–201.

9. Vynnycky, E. & Fine, P. E. (1999) *Int. J. Epidemiol*. **2,** 327–334.

10. Dye, C., Garnett, G. P., Sleeman, K. & Williams, B. G. (1998) *Lancet* **352,** 1886–1891.

11. Glynn, J. R., Bauer, J., de Boer, A. S., Borgdorff, M. W., Fine, P. E., Godfrey-Faussett, P. & Vynnycky, E. (1999) *Int. J. Tuberc. Lung Dis*. **3,** 55–60.

12. Rhee, J. T., Tanaka, M. M., Behr, M. A., Agasino, C. B., Paz, E. A., Hopewell, P. C. & Small, P. M. (2000) *Int. J. Tuberc. Lung Dis.* **4,** 1111–1119.

13. Vynnycky, E., Nagelkerke, N., Borgdorff, M. W., van Soolingen, D., van Embden, J. D. & Fine, P. E. (2001) *Epidemiol. Infect*. **126,** 43–62.

14. Sutherland, I. (1968) *TSRU Progress Report* (KNCV, Hague, The Netherlands).

15. Sutherland, I., Svandova, E., Radhakrishna, S. (1982) *Tubercle* **63,** 255-268.

16. Daley, C., Small, P., Schecter, G., *et al.* (1992) *N. Engl. J. Med.* **326,** 231-325.

17. Coronado, V., Beck-Segue, C., Hutton, M., *et al.* (1993) *J. Infect. Dis.* **68,** 1052-1055.

18. Ryder, R. W., Batter, V., Kaseka, N., Behets, Sequeira, F., M'Boly, E, Kanda, M., Tshimbe, M. & Morgan, M. (2000) *AIDS Patient Care STDS* **14**, 297-304.

19. Woods, R., Maartens, G. & Lombard, C. J. J. (2000) *Acquir. Immune. Defic. Syndr*. **23**, 75-80.

20. Churchyard, G. J., Kleinschmidt, T., Corbett, L. L., Mulder, D.& DeCock, K. L. (1999) *Int. J. Tuberc. Lung Dis.* **3**, 791-798.

21. Styblo, K. (1991) *The Epidemiology of Tuberculosis* (KNCV, Royal Netherlands Tuberculosis Association, Hague, The Netherlands).