Annual *Mycobacterium tuberculosis* Infection Risk and Interpretation of Clustering Statistics

Emilia Vynnycky,* Martien W. Borgdorff,† Dick van Soolingen,‡ and Paul E.M. Fine*

Several recent studies have used proportions of tuberculosis cases sharing identical DNA fingerprint patterns (i.e., isolate clustering) to estimate the extent of disease attributable to recent transmission. Using a model of introduction and transmission of strains with different DNA fingerprint patterns, we show that the properties and interpretation of clustering statistics may differ substantially between settings. For some unindustrialized countries, where the annual risk for infection has changed little over time, 70% to 80% of all age groups may be clustered during a 3-year period, which underestimates the proportion of disease attributable to recent transmission. In contrast, for a typical industrialized setting (the Netherlands), clustering declines with increasing age (from 75% to 15% among young and old patients, respectively) and underestimates the extent of recent transmission only for young patients. We conclude that, in some settings, clustering is an unreliable indicator of the extent of recent transmission.

Studies are increasingly using levels of clustering of isolates from tuberculosis cases (proportion sharing identical DNA fingerprint patterns) to estimate the extent of disease attributable to recent transmission. To date, few studies have been conducted in unindustrialized countries, where the impact of tuberculosis and the proportion of disease attributable to recent transmission are greatest. Whether or not the properties and interpretation of clustering statistics in such settings are similar to those in industrialized populations is unclear.

Studies in industrialized countries have found relatively low overall levels of clustering (e.g., 30% to 40% during a 3-year period [1–3]) but much higher levels among younger versus older patients. This age differential probably reflects past trends in the annual risk for infection, which was high in the early 20th century (e.g., >2% per year before 1940 in the Netherlands [4]) and is currently very low. Thus, a large proportion of disease in older patients is attributable to reactivation of infections acquired many years ago, and, given the short half-life of DNA fingerprint patterns (5), only a small proportion of old patients share identical isolates with other patients**.** In some unindustrialized countries, on the other hand, where the annual risk for infection may not have changed much over time, the age differential in clustering might be small, given that a large proportion of disease even among older persons may be attributable to recent (re)infection. Understanding the effect of the magnitude of the annual risk for tuberculous infection on clustering frequency helps determine how molecular epidemiologic data can be best applied to estimate the extent of ongoing transmission of *Mycobacterium tuberculosis*, and hence to identify optimal control strategies.

We explored how the magnitude and trend in the annual risk for infection influence the age-specific proportion of clustered cases and its relationship to the extent of disease attributable to recent transmission. We use a model of the transmission dynamics of *M. tuberculosis* previously calibrated to data from the Netherlands (6), where isolates from all tuberculosis cases with onset since 1993 have been routinely DNA fingerprinted (1). We describe the general epidemiologic assumptions in the model and how it distinguishes between cases according to the DNA fingerprint pattern of the strain causing the disease episode, which is needed to calculate clustering statistics.

Methods

Our analysis is based on a model developed recently to interpret data on clustering of DNA fingerprint patterns in the Netherlands (6). Equations describing the model's formulation are provided in the Appendix.

Epidemiologic Assumptions in the Model

The model's structure, parameters, and assumptions have been published (6). Persons are assumed to be born uninfected. Infected persons are divided into those in whom primary disease has not yet developed (defined by convention as disease within 5 years of initial infection [7]), and those in the "latent" class, who are at risk for endogenous reactivation or for reinfection, which can be followed by exogenous disease. Exogenous disease is here defined as the first disease episode within 5 years of the most recent reinfection; endogenous disease includes disease occurring >5 years after the most recent (re)infection event, and second or subsequent disease episodes occurring <5 years after the most recent (re)infection event. (These definitions differ slightly from those of Sutherland et al. (8) to include the assumption that once persons have recovered from disease during the first 5 years after initial infection

*London School of Hygiene & Tropical Medicine, London, England; †Royal Netherlands Tuberculosis Association, The Hague, the Netherlands; and ‡National Institute of Public Health and the Environment, Bilthoven, the Netherlands

or reinfection, their risk of developing disease becomes the same as that of developing disease through reactivation, until they are newly reinfected.)

The infection and reinfection risks are assumed to be identical, but reinfection is less likely to lead to disease than is initial infection, due to some immunity induced by the prior infection (9). We explored the implications of four assumptions for the magnitude (and trend) in the annual risk for infection, namely, that the risk for infection 1) declined over time, as estimated for the Netherlands (from approximately 2% in 1940 to approximately 5/10,000 by 1979 [4,10]); 2) remained unchanged over time at a very low level (0.1%); 3) remained unchanged at 1%; or 4) remained unchanged at 3%. Infection risks of 1% have been found in several populations (e.g., Malawi [11]). Infection risks of 3% are uncommon today but have been reported in parts of South Africa (12). For simplicity, we assumed that persons cannot be reinfected during the period between initial infection (or reinfection) and onset of the first primary episode (or exogenous disease).

The risks of developing disease depend on age and sex (Figure 1A; [6]); they are based on previous analyses, in which we fitted predictions of disease incidence to observed notifications in the U.K. (9). The risks of developing either a first primary episode or disease following exogenous reinfection also depend on the time since infection and reinfection, respectively (Figure 1B). The probability that a disease episode is infectious (sputum smear/culture-positive) is age dependent (Figure 1C) (9). The demography of the population described in the model is assumed to be that for the Netherlands. Analyses are restricted to respiratory (pulmonary) forms of tuberculosis, since these are far more likely than extrapulmonary forms to lead to transmission. Although additional factors such as immigration and HIV can influence the extent of clustering in complicated ways (14), these factors are not considered here, where the focus is upon the effect of the magnitude and trend in the annual risk for infection on clustering.

Derivation of Clustering Statistics

Recent studies suggest that the half-life of DNA fingerprint patterns based on IS*6110* restriction fragment length polymorphism (RFLP, which has been used for the DNA fingerprinting conducted to date in most studies) is 2–5 years (5,15). If the molecular clock speed for IS*6110* RFLP patterns of strains involved in latent infection (currently unknown) were to be similar, this relatively short half-life implies that most of the fingerprint patterns of the strains causing disease today differ from those that caused disease many years ago. Similarly, this short half-life implies that the *M. tuberculosis* fingerprint types and cluster distributions in tuberculosis cases today depend only loosely upon those that existed 50 years ago. Based on this assumption, to derive clustering estimates for a given population for recent years, we designed the model to simulate the introduction and subsequent transmission of strains with new DNA fingerprint patterns from a sufficiently distant time in the past (taken to be 1950), so that a) all cases with onset in recent years involved a strain whose DNA fingerprint pattern had first appeared since then and b) no assumptions would be required about the distributions of strains that existed before 1950. The general steps in the calculations are outlined briefly below.

The numbers of persons of each age in each of the epidemiologic categories for 1950 were calculated by using the model, based on described equations (9). From 1950, each of these age-sex classes was stratified to distinguish between those who had, versus those who had not, been (re)infected since 1950. Those who had been (re)infected since 1950 were subdivided further according to the time of infection or reinfection. The transmission dynamics were tracked simultaneously for all persons with the equations described in the Appendix and elsewhere (6), by using time steps of 6 months and 1 year for calendar year and age, respectively.

In each interval, disease was assumed to develop in a proportion of infected persons, and a proportion of these disease episodes was attributed to a strain for which the DNA finger-

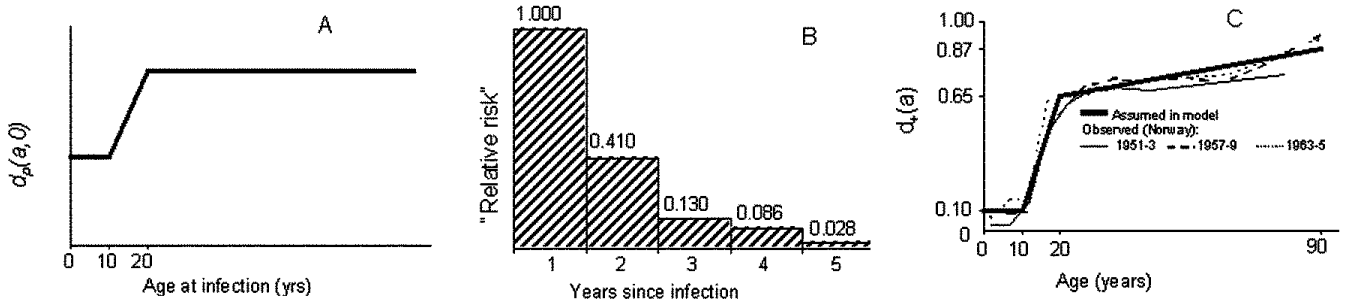

Figure 1. Summary of the main assumptions in the model relating to the risks of developing disease. A) General relationship between the risk of developing the first primary episode (during the first year after infection) and age at infection. An identical relationship is assumed to hold between the risk of exogenous disease and the age at reinfection and between the risk for endogenous disease and the current age of persons. B) Risk of developing the first primary episode (or exogenous disease) in each year following initial infection (or reinfection), relative to that experienced in the first year after infection. The relationship was derived by using data on the interval between tuberculin conversion and disease among persons who were tuberculin negative at intake during the U.K. Medical Research Council BCG trial during the 1950s (13). C) Proportion of respiratory disease incidence manifested as sputum-positive (i.e., infectious) (pers. comm. K. Styblo, Tuberculosis Surveillance Research Unit, and K. Bjartveit, Norwegian National Health Screening Service).

print pattern differed from that of the strain with which the persons were originally infected. This latter proportion depended on the time since infection (see below), and each of the new DNA fingerprint patterns was assigned a unique identity number. Each infectious patient with onset at a given time was assumed to contact a different number of persons (see Appendix and Figure 2). The frequency distributions of the number of persons contacted by each patient were used to derive the total number of persons who were newly (re)infected at this time. The corresponding equations were then applied to this number to determine the total number of persons in whom disease developed at a later time, *T,* among those who had been infected at time *t*. The DNA fingerprint patterns of the strains in these diseased persons were then determined by using the frequency distribution of the number of persons contacted by each case-patient at time *t*. These calculations are described further in the Appendix.

Estimating the Effect of the Annual Risk for Infection on Clustering as an Indicator of Recent Transmission

Our model was used to calculate the age-specific proportion of disease attributable to primary and exogenous disease from 1993 to 1997 for the Netherlands and for settings in which the annual risk for infection is assumed to have remained unchanged over time at 0.1%, 1%, and 3%. Primary and exogenous disease involve disease occurring during the first 5 years after the most recent (re)infection event, although the majority of persons in whom primary or exogenous (reinfection) disease develops acquire the disease within 2–3 years (Figure 1B). The clustering by sex and age for cases with onset in different periods between 1993 and 1997 for the Netherlands, and for settings in which the annual risk for infection is assumed to have remained unchanged over time at 0.1%, 1%, and 3%,was also calculated by using the age and sex distribution of the cases with onset in that period (see equations in [6]). For simplicity, we present age-specific levels of clustering for male patients only. Model predictions for male patients generally compared better against the observed data in the Netherlands than did those for female patients (6).

The predictive values of clustering for the identification of recent transmission were calculated as follows. The positive predictive value of clustering for identifying recent transmission in different age groups in different periods was calculated as the proportion of case-patients who were in a cluster in a given period who had been infected or reinfected <5 years before disease onset. The negative predictive value of clustering for identifying recent transmission in different age groups was calculated as the proportion of case-patients who were not in a cluster in a given period who had been infected or last reinfected >5 years before disease onset.

Results

Model Predictions of the Extent of Clustering and Disease Attributable to Recent Transmission

As shown in Figure 3, very different age patterns in the proportion of disease attributable to recent transmission were predicted for the Netherlands and for settings in which the annual risk for infection has remained unchanged over time. In the Netherlands, the proportion of disease attributed to recent infection decreased dramatically with age, e.g., from 100% in the young to approximately 50% and 10% for 45- to 54-year-old patients and persons >65 years of age, respectively. The proportion of disease attributed to recent reinfection was very low for all age groups (<3%). For constant infection risk settings, the predicted proportion of disease attributable to recent

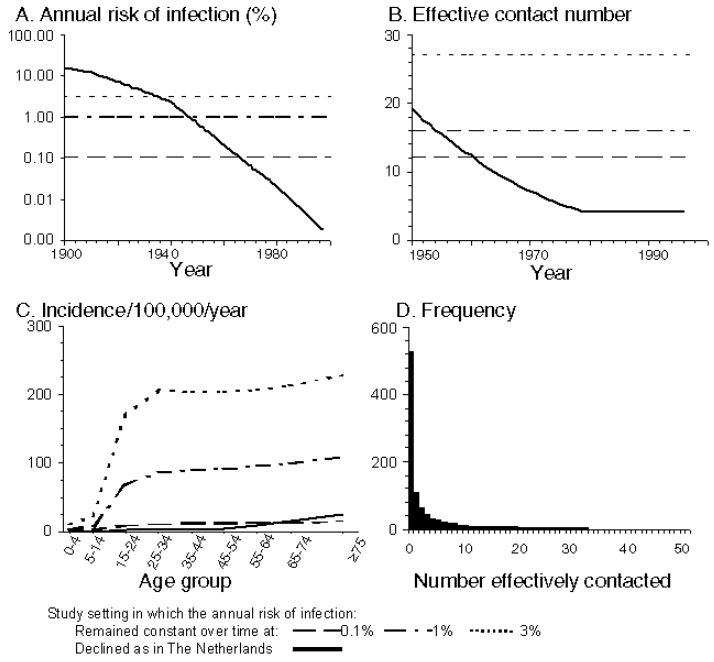

Figure 2. Summary of the assumptions defining contact between persons in the model. A, B, and C show the annual risk for infection, estimates of the average effective contact number in the model, and the average age-specific annual incidence of infectious disease per 100,000 population respectively in the various settings. For settings in which the annual risk for infection has not changed over time, the effective contact number is obtained from the ratio between the annual risk for infection and the incidence of infectious cases predicted in the model. The values for the effective contact number in the Netherlands are identical to those calculated in reference 6. D shows the frequency distribution of the assumed number of persons effectively contacted by each infectious case-patient, if the population were to comprise 1,000 infectious cases and the average effective contact number was approximately 4, as assumed for the Netherlands for recent years. This (negative binomial) distribution (defined by a variance 20 times the mean) led to observed cluster distributions that best compared against those observed in the Netherlands (6). Contact between persons is assumed to be assortative (so that, for example, those with a high-risk lifestyle, mix preferentially with similar persons) and, for simplicity, independent of age and sex.

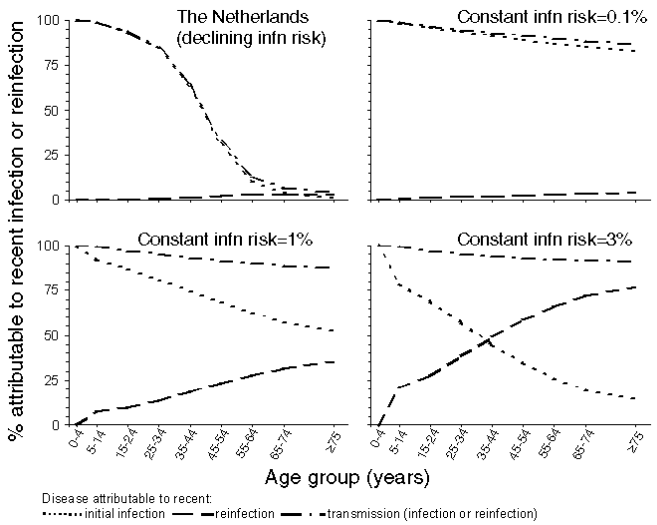

Figure 3. Model predictions of the proportion of disease attributable to primary and exogenous disease during the period 1993–1997 in the Netherlands and settings in which the annual risk for infection has remained unchanged over time at 0.1%, 1%, and 3% per year.

transmission (i.e., recent infection or reinfection) was very similar, falling from 100% in the young to 85%, 88%, and 90% in the oldest age groups for the 0.1%, 1%, and 3% infection risk scenarios, respectively. On the other hand, large differences between settings were predicted in the proportion of disease attributed to initial infection or to reinfection. In all instances, the proportion attributed to reinfection was zero in the youngest age groups, but this proportion increased with age to 3%, 35%, and 80% for the 0.1%, 1%, and 3% annual infection risk assumptions, respectively. The proportion attributed to recent initial infection in these settings decreased from 100% in the young to 80%, 50%, and 15%, respectively, in old patients.

As shown in Figure 4A, for each setting, the overall clustering (i.e., that seen among all age groups) was predicted to increase with study duration, e.g., from 15% for the Netherlands for a 1-year period to approximately 25% for a 5-year period. The clustering predicted for all the constant infection risk scenarios was similar in magnitude for each study period and increased from 60% to 70% for a 1-year period to 75% to 85% for a 5-year period. Since the overall clustering was not predicted to increase much for study periods of more than 3 years, clustering is defined using a 3-year period in the remainder of these analyses (represented by 1993–1995). As shown in Figure 4B, the clustering predicted for each age group was similar for each of the settings in which the annual risk for infection remained unchanged over time, and declined only slightly with age, e.g., from 83% for the youngest age group to approximately 75% for the oldest age category. In contrast, for the Netherlands, the clustering was predicted to decrease dramatically with age, from approximately 75% among young case-patients to approximately 15% in very old patients. This prediction is consistent with observed data (Figure 4B).

Reliability of Clustering as a Measure of the Extent of Recent Transmission

For settings in which the annual risk for infection remained unchanged over time at 0.1%, 1%, and 3%, the predicted clustering in each age group underestimated the proportion who had been recently infected or reinfected (Figure 5). In settings with an annual risk for infection of 0.1%, at least 90% of cases in each age group were predicted to have been recently (re)infected, whereas the proportion clustered decreased from about 85% in the youngest age group to approximately 70% for the oldest persons. For the Netherlands (described elsewhere [6]), clustering underestimated the proportion of disease attributable to recent transmission in the young (by up to 43%) and overestimated that for older patients (by up to 50%).

The positive and negative predictive values of being in a cluster, as an indicator of recent transmission, depended both on age and the study setting (Figure 6). For settings with a high annual risk for infection that had remained unchanged over time, model predictions suggested that most patients clustered in each age group were likely to have been recently (re)infected, corresponding to a positive predictive value of clustering for recent transmission of almost 100% in each age

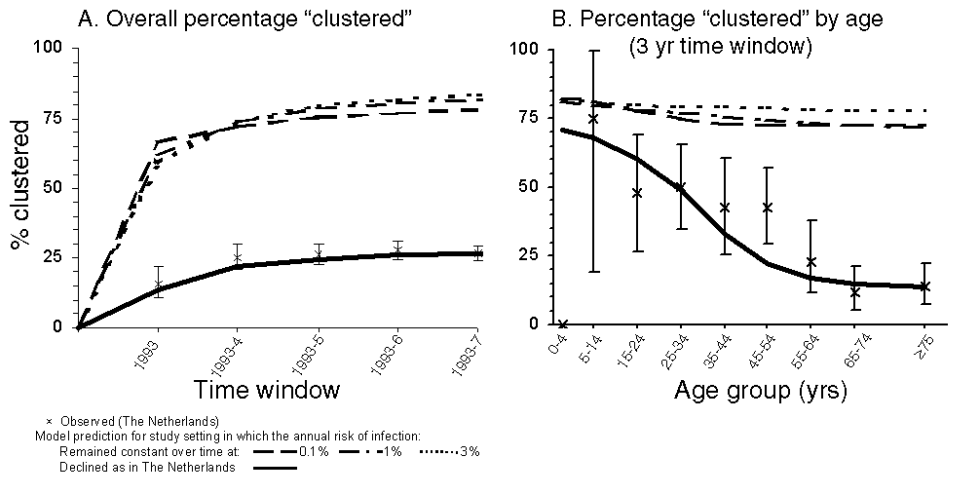

Figure 4. Model predictions of A) the overall percentage of cases clustered during different time periods from 1993 to 1997 and B) the age-specific percentage of (male) cases clustered during the period 1993–1995 in the Netherlands and in settings in which the annual risk for infection has remained unchanged over time at 0.1%, 1%, and 3%. The clustering observed in the Netherlands, after excluding clusters involving immigrants, is also shown.

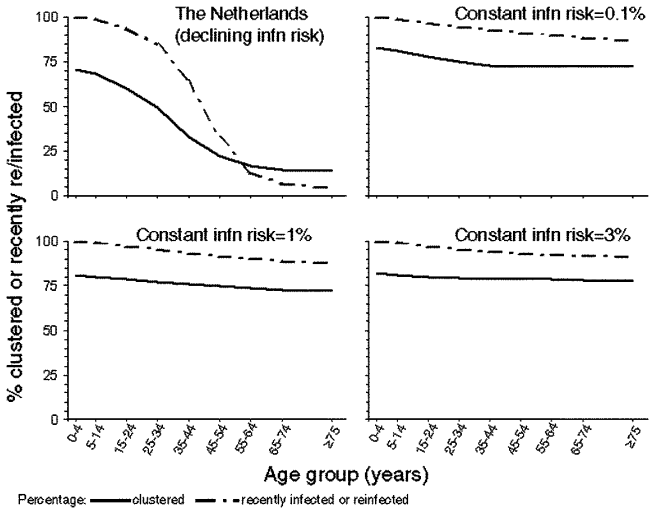

Figure 5. Comparison between model predictions of the clustering in different age groups and the proportion of disease attributable to recent infection or reinfection in the Netherlands and in settings in which the annual risk for infection has remained unchanged over time at 0.1%, 1%, and 3%.

group (Figure 6A). The positive predictive value was estimated to decrease with age in the Netherlands from 100% in the very young to about 20% for the oldest patients.

When unclustered cases were considered, the proportion of clinical case-patients who were estimated to have been infected >5 years previously was low (<5%) for young patients and increased with age for all settings, approaching 100% for patients >55 years of age in the Netherlands (Figure 6B). Almost all case-patients of ages >55 years who were not in a cluster in the Netherlands were therefore estimated to have been infected >5 years previously and thus owed their disease to reactivation of latent foci. Of the adult case-patients who were not in a cluster in the other settings, the proportion who had been infected >5 years previously was <45%, 35%, and 20% if the annual risk for infection was 0.1%, 1%, and 3%, respectively.

Discussion

The availability of DNA fingerprinting techniques has led to a large number of studies that measure clustering of isolates from tuberculosis cases (1–3,16). Most of these studies have been conducted in industrialized settings and have found relatively low levels of clustering (30% to 40%) and decreases in clustering with age. Our analyses indicate that those findings have been influenced strongly by the large secular decline in the annual risk for infection that occurred in industrialized settings during the 20th century and that very different findings are expected in settings where the annual risk for infection has changed little over time. The clustering predicted is high (>60% for 2-year periods) in such settings, similar for all age groups, and may nevertheless still underestimate the extent of disease that is due to recent transmission.

Our conclusions are based on a model of the transmission dynamics of *M. tuberculosis* that includes several simplifications. The most obvious is our assumption that the risks for disease, given infection in settings in which the infection risk is high, are the same as those estimated for industrialized populations. HIV influences these risks (17,18), although its effect on clustering is not yet understood (14). Another simplification is our assumption that the half-life of DNA fingerprint patterns is identical for strains involved in active disease and in latent infection. If latent infections are associated with a slow rate of genetic change of the bacilli, our assumption would have led to an underestimate of clustering but would not have affected our conclusions for settings in which the annual risk for infection has remained unchanged over time, where only a small proportion of disease is attributed to reactivation of a latent infection (Figure 3). The effect of this assumption on clustering estimates for the Netherlands is discussed elsewhere (6).

Our finding that the overall amount of clustering in populations with a low (constant) annual infection risk should be similar to that observed in populations with a high (constant) infection risk may appear paradoxical. Our finding follows from the fact that in such populations any decline in the pro-

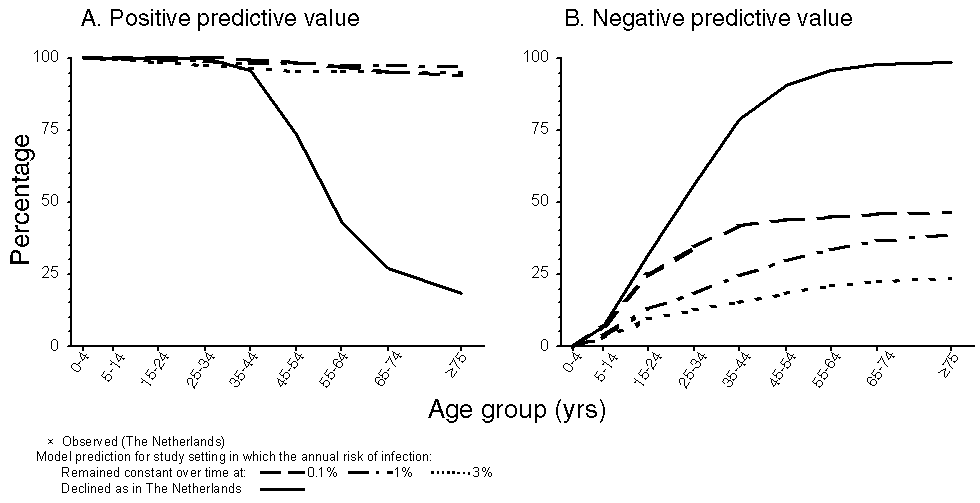

Figure 6. Summary of model predictions of the A) positive predictive values of clustering (proportion of cases who are in a cluster who have been infected or reinfected <5 years before onset) and B) negative predictive values of clustering (proportion of cases who are not in a cluster who are experiencing disease as a result of infection or reinfection acquired >5 years before onset) in different age groups in the Netherlands and in settings in which the annual risk for infection has remained unchanged over time at 0.1%, 1%, and 3%.

portion of disease attributable to recent primary infection with age is compensated by increases in the proportion attributable to recent reinfection with age (Figure 3). As a result, both the overall and age-specific predicted proportions of disease attributable to recent transmission in these populations are very similar; this finding leads to predictions that the overall and age-specific levels of clustering in these settings would also be similar.

Previous model-based analyses (6) have indicated that in industrialized settings such as the Netherlands clustering among young case-patients will underestimate the extent of disease attributable to recent transmission (because some sources of infection have onset outside the study period and because DNA fingerprint patterns can change between infection and disease onset), and clustering among old case-patients may overestimate recent transmission (because clustering among older case-patients is more likely to be attributable to their being sources of infection rather than their being recently reinfected). These analyses extend those findings and indicate that in settings in which the annual risk for infection has not changed much over time, the overall level of clustering in any given age group is likely to underestimate the extent of recent transmission (Figure 5). This underestimate follows from the fact that in these settings, most disease in all age groups is attributable to recent transmission, and some patients will have been infected or reinfected immediately before the study started and thus may not be in a cluster.

These analyses provide the first estimates of the positive and negative predictive values of clustering. Overall, these analyses highlight the fact that in settings in which the annual risk for infection has not changed greatly over time, most clustered case-patients are likely to have been recently infected or reinfected (i.e., the positive predictive value of clustering is high) (Figure 6). This finding suggests that in such settings, application of the "n-1" rule (2), which assumes that each cluster comprises an index case attributable to reactivation and the other cases result (in)directly from that case, will lead to even more unreliable estimates of the extent of recent transmission than those based on the "n" rule. Similarly, estimates of the proportion of disease attributable to reactivation will be unreliable if they are based on the proportion of patients who fail to be in a cluster in a given period.

Our analyses demonstrate that the properties and interpretation of clustering statistics depend strongly on the trend and magnitude in the annual risk for infection and thus will vary between settings. For example, in settings in which the annual risk for infection has remained unchanged at either a high or a low level, the age differential in clustering is likely to be small, in contrast with that in industrialized settings, and clustering is likely to underestimate the extent of recent transmission in all age groups. Given the growing importance of clustering studies, which, to date have been conducted in populations in which the annual risk for infection declined dramatically over time and is currently very low, these insights are important for an improved understanding of the natural history of tuberculosis.

Acknowledgments

We thank the late K. Styblo and K. Bjartveit for supplying tuberculosis data from Norway and N. Kalisvaart for supplying notification data from the Netherlands.

We thank the British Medical Research Council and the European Community Concerted Action Programme for financial support.

Dr. Vynnycky is a lecturer in infectious disease modeling at the London School of Hygiene & Tropical Medicine. Her research interests include modeling the transmission dynamics of infectious diseases, and the epidemiology and molecular epidemiology of tuberculosis.

References

1. van Soolingen D, Borgdorff MW, de Haas PE, Sebek MM, Veen J, Dessens M, et al. Molecular epidemiology of tuberculosis in the Netherlands: a nationwide study from 1993 through 1997. J Infect Dis 1999;180:726–36.
2. Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, et al. The epidemiology of tuberculosis in San Francisco: a population-based study using conventional and molecular methods. N Engl J Med 1994;330:1703–9.
3. Bauer J, Yang Z, Poulsen S, Andersen AB. Results from 5 years of nationwide DNA fingerprinting of *Mycobacterium tuberculosis* complex isolates in a country with a low incidence of *M. tuberculosis* infection. J Clin Microbiol 1998;36:305–8.
4. Styblo K, Meijer J, Sutherland I. The transmission of tubercle bacilli: its trend in a human population. Bulletin of the International Union against Tuberculosis 1969;42:5–104.
5. de Boer AS, Borgdorff MW, de Haas PEW, Nagelkerke NJ, van Embden JD, van Soolingen D. Analysis of rate of change of IS*6110* RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. J Infect Dis 1999;180:1238–44.
6. Vynnycky E, Nagelkerke N, Borgdorff MW, van Soolingen D, van Embden JDA, Fine PEM. The effect of age and study duration on the relationship between 'clustering' of DNA fingerprint patterns and the proportion of tuberculosis disease attributable to recent transmission. Epidemiol Infect 2001;126:43–62.
7. Holm J. Development from tuberculous infection to tuberculous disease. The Hague, the Netherlands: Royal Dutch Tuberculosis Assocation (KNCV); 1969.
8. Sutherland I, Švandová E, Radhakrishna SE. The development of clinical tuberculosis following infection with tubercle bacilli. Tubercle 1982;63:255–68.
9. Vynnycky E, Fine PEM. The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. Epidemiol Infect 1997;119:183–201.
10. Sutherland I, Bleiker MA, Meijer J, Styblo K. The risk of infection in the Netherlands from 1967 to 1979. Tubercle 1983;64:241–53.
11. Fine PEM, Bruce J, Ponnighaus JM, Nkhosa P, Harawa A, Vynnycky E. Tuberculin sensitivity: conversions and reversions in a rural African population. Int J Tuberc Lung Dis 1999;3:962–75.
12. van Rie A, Warren R, Richardson M, Victor TC, Gie RP, Enarson DA, et al. Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment. N Engl J Med 1999;341:1174–9.
13. Sutherland I. The ten-year incidence of clinical tuberculosis following "conversion" in 2,550 individuals aged 14 to 19 years. The Hague, the Netherlands: Royal Dutch Tuberculosis Association (KNCV); 1968.
14. Glynn JR, Bauer J, de Boer AS, Borgdorff MW, Fine PEM, Godfrey-Faussett P, et al. Interpreting DNA fingerprint clusters of *Mycobacterium tuberculosis*. European Concerted Action on Molecular Epidemiology and Control of Tuberculosis. Int J Tuberc Lung Dis 1999;3:1055–60.

15. Yeh RW, Ponce de Leon A, Agasino CB, Hahn JA, Daley CL, Hopewell PC, et al. Stability of *Mycobacterium tuberculosis* DNA genotypes. J Infect Dis 1998;177:1107–11.
16. Alland D, Kalkut GE, Moss AR, McAdam RA, Hahn JA, Bosworth W, et al. Transmission of tuberculosis in New York City: an analysis by DNA fingerprinting and conventional epidemiologic methods. N Engl J Med 1994;330:1710–6.
17. Glynn JR. Resurgence of tuberculosis and the impact of HIV infection. British Medical Bulletin 1998;54:579–93.
18. Glynn JR, Warndorff DK, Fine PEM, Msiska GK, Munthali MM, Ponnighaus JM. The impact of HIV on morbidity and mortality from tuberculosis in sub-Saharan Africa: a study in rural Malawi and review of the literature. Health Transition Review 1997:75–87.
19. ten Asbroek AH, Borgdorff MW, Nagelkerke NJ, Sebek MM, Deville W, van Embden JD, et al. Estimation of serial interval and incubation period of tuberculosis using DNA fingerprinting. Int J Tuberc Lung Dis 1999;3:414–20.

Address for correspondence: E. Vynnycky, Infectious Disease Epidemiology Unit, Department of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT fax: +44 (0)20 7636 8739; e-mail: emilia.vynnycky@lshtm.ac.uk

Appendix

Summary of Equations Used in Model

We use the notation summarized in the Table and reference 6 to describe the transmission dynamics of *Mycobacterium tuberculosis*. Note that all of the variables are stratified by sex; for reporting convenience, we have omitted this stratification in the following description. The equations are as follows:

$$\frac{\partial U(a,t)}{\partial a}+\frac{\partial U(a,t)}{\partial t}=-(i(t)+m_g(a,t))U(a,t) \quad (1)$$

$$\frac{\partial I_T(a,t,s)}{\partial a}+\frac{\partial I_T(a,t,s)}{\partial t}+\frac{\partial I_T(a,t,s)}{\partial s}=-(d_p(a-s,s)+m_g(a,t)+k_L(s))I_T(a,t,s) \quad (0<s\le 5) \quad (2)$$

$$\frac{\partial P_T(a,t,\hat{s})}{\partial a}+\frac{\partial P_T(a,t,\hat{s})}{\partial t}+\frac{\partial P_T(a,t,\hat{s})}{\partial \hat{s}}=\int_0^5 d_p(a-s,s)I_T(a,t,s)ds-(m_+(t,\hat{s})d_+(a)+m_g(a,t)d_-(a)-r(a,t,\hat{s}))P_T(a,t,\hat{s}) \quad (3)$$

$$\frac{\partial L_T(a,t)}{\partial a}+\frac{\partial L_T(a,t)}{\partial t}=(I_T(a,t,5)+I_{r_T}(a,t,5))k_L(5)+r(a,t,2)(P_T(a,t,2)+E_{n_T}(a,t,2)+E_{x_T}(a,t,2))-(i(t)+d_n(a)+m_g(a,t))L_T(a,t) \quad (4)$$

$$\frac{\partial I_{r_T}(a,t,s)}{\partial a}+\frac{\partial I_{r_T}(a,t,s)}{\partial t}+\frac{\partial I_{r_T}(a,t,s)}{\partial s}=-(d_x(a-s,s)+m_g(a,t)+k_L(s))I_{r_T}(a,t,s) \quad (0<s\le 5) \quad (5)$$

$$\frac{\partial E_{x_T}(a,t,\hat{s})}{\partial a}+\frac{\partial E_{x_T}(a,t,\hat{s})}{\partial t}+\frac{\partial E_{x_T}(a,t,\hat{s})}{\partial \hat{s}}=\int_0^5 d_x(a-s,s)I_{r_T}(a,t,s)ds-(m_+(t,\hat{s})d_+(a)+m_g(a,t)d_-(a)+r(a,t,\hat{s}))E_{x_T}(a,t,\hat{s}) \quad (6)$$

$$\frac{\partial E_{n_T}(a,t,\hat{s})}{\partial a}+\frac{\partial E_{n_T}(a,t,\hat{s})}{\partial t}+\frac{\partial E_{n_T}(a,t,\hat{s})}{\partial \hat{s}}=d_n(a)L_T(a,t)-r(a,t,\hat{s})E_{n_T}(a,t,\hat{s})-(m_+(t,\hat{s})d_+(a)+m_g(a,t)d_-(a)+r(a,t,\hat{s}))E_{x_T}(a,t,\hat{s}) \quad (7)$$

Boundary conditions:

$$U(0,t)=B(t)$$
$$I_T(a,T,0)=i(T)U(a,T)$$
$$I_{r_T}(a,T,0)=i(T)\sum_t L_t(a,T)$$

For notational convenience, we denote (1-$d_+(a)$) by $d_-(a)$. The infection risk at time t ($i(t)$) is given by $\sum_n nF(t,n)/N(t)$ where $N(t)$ is the total population size at time t and $F(t,n)$ is the frequency distribution of the number of persons contacted by the case-patients who had onset at time t (Figure 2). The total number of infectious cases at time t is given by the total number of persons experiencing their first primary episode, endogenous and exogenous disease, summed over all possible ages and times of infection T, i.e $\sum_a \sum_T P_T(a,t,0)+E_{n_T}(a,t,0)+E_{x_T}(a,t,0)$

Simulating Contact between Persons

For simplicity, we assumed that all effective contacts (defined as those sufficient to lead to infection by an infectious case-patient if the contacted person has never been infected) occurred immediately after onset of (infectious pulmonary) disease in the source case. This assumption is reasonable for industrialized countries in recent years (see, for example, 19) but is less realistic for some unindustrialized countries because of longer diagnostic delays in such populations. The number of persons effectively contacted by each case-patient during the infectious period (the effective contact number) was assumed to follow a negative binomial distribution, defined by a time-dependent mean and variance (Figure 2). Though assumptions about contact patterns between persons influence the predicted cluster distributions, they do not affect the overall levels of clustering (6). The data used for calibrating the model's assumptions have been described (6).

Appendix table. Definitions of state variables used in the model

Variable	Definition
$B(t)$	No. of live births at time t. Obtained from the Dutch Central Bureau for Statistics (data available from 1892 to present).
$U(a,t)$	No. of uninfected persons of age a at time t .
$I_T(a,t,\hat{s})$	No. of persons of age a at time t who were infected at time T and have been infected for time s ($\leq$5 years) without having yet developed disease.
$P_T(a,t,\hat{s})$	No. of persons of age a first infected at time T who are experiencing their first primary episode at time t, who have been diseased for time $\hat{s}$
$L_T(a,t)$	No. of persons of age a at time t in the "latent" class (comprising those who have either just recovered from their first primary episode, or who have been infected for >5 years) whose most recent (re)infection event occurred at time T.
$I_{r_T}(a,t,s)$	No. of persons of age a at time t, whose most recent reinfection occurred at time T, who have been reinfected for time s ($\leq$5 years) and in whom exogenous disease has not yet developed.
$E_{X_T}(a,t,\hat{s})$	No. of persons of age a with exogenous disease at time t, who have been diseased for time $\hat{s}$ and whose most recent reinfection occurred at time T.
$E_{n_T}(a,t,\hat{s})$	No. of persons of age a with endogenous disease at time t, who have been diseased for time $\hat{s}$ and whose most recent reinfection occurred at time T.

Calculating the Distribution of Strains among Cases at a Given Time

We assumed that all reactivations (which generally involve persons infected for >5 years) of infections acquired before 1950 were with unique strains and that the strain isolated from persons who had been reinfected was from the most recent (re)infection event. The DNA fingerprint pattern of the strain causing disease among each of the case-patients with onset at time T and whose most recent (re)infection had occurred at time t since 1950 was assumed to be identical to that with which the source of infection of that person (identified by using the algorithm described in [6]) had been infected, unless the DNA fingerprint pattern had since changed through random mutations. The proportion of case-patients who had been infected at time t for whom the DNA fingerprint pattern was assumed to have changed was given by the expression $1-e^{-0.21661(T-t)}$ which describes a half-life of 3.2 years for DNA fingerprint patterns, as found in a recent study (5). These analyses assume implicitly that clustered cases were involved, at some level, in the same chain of transmission and that clustering was not attributable, e.g., to preferential insertion of IS*6110* into any particular location in the genome.