Epidemiol. Infect. (2001), **126**, 43–62. Printed in the United Kingdom

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

E. VYNNYCKY[1]*, N. NAGELKERKE[2], M. W. BORGDORFF[2], D. VAN SOOLINGEN[3], J. D. A. VAN EMBDEN[3] AND P. E. M. FINE[1]

[1] *Infectious Disease Epidemiology Unit, Department of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT*
[2] *Royal Netherlands Tuberculosis Association, PO Box 146, 2501 CC The Hague, The Netherlands*
[3] *National Institute of Public Health and the Environment, PO Box 1, 3720 BA Bilthoven, The Netherlands*

(*Accepted 16 October 2000*)

SUMMARY

Though it is recognized that the extent of 'clustering' of isolates from tuberculosis cases in a given population is related to the amount of disease attributable to recent transmission, the relationship between the two statistics is poorly understood. Given age-dependent risks of disease and the fact that a long study (e.g. spanning several years) is more likely to identify transmission-linked cases than a shorter study, both measures, and thus the relationship between them, probably depend strongly on the ages of the cases ascertained and study duration. The contribution of these factors is explored in this paper using an age-structured model which describes the introduction and transmission of *M. tuberculosis* strains with different DNA fingerprint patterns in The Netherlands during this century, assuming that the number of individuals contacted by each case varies between cases and that DNA fingerprint patterns change over time through random mutations, as observed in several studies.

Model predictions of clustering in different age groups and over different time periods between 1993 and 1997 compare well against those observed. According to the model, the proportion of young cases with onset in a given time period who were 'clustered' underestimated the proportion of disease attributable to recent transmission in this age group (by up to 25% in males); for older individuals, clustering overestimated this proportion. These under- and overestimates decreased and increased respectively as the time period over which the cases were ascertained increased. These results have important implications for the interpretation of estimates of the proportion of disease attributable to recent transmission, based on 'clustering' statistics, as are being derived from studies of the molecular epidemiology of tuberculosis in many populations.

INTRODUCTION

Since the development of DNA fingerprinting techniques for typing strains of *M. tuberculosis* [1], many studies have used the levels of 'clustering' of isolates from tuberculosis cases to estimate the proportion of disease attributable to recent transmission [2–6]. Though several studies have identified very high levels of clustering (e.g. as high as 40% [2, 6]), and have argued that the proportion of disease attributable to recent transmission was therefore higher than was expected, the relationship between the two measures is still poorly understood.

One of the most important factors determining the extent to which the clustering observed in a population reflects recent transmission is the speed at which DNA fingerprint patterns change over time (the 'molecular

* Author for correspondence.

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

clock' speed). A fast molecular clock implies that only cases separated by very short serial intervals (time intervals between successive cases [7]) are likely to share bacilli with the same DNA fingerprint pattern (and hence to be 'clustered'). A slow molecular clock implies that even cases involved in chains of transmission spanning many years may still appear clustered. As a result, the relationship between the clustering observed in a population and the proportion of disease attributable to recent transmission will depend on the length of time over which cases are ascertained. It will also depend on the age of cases, given that disease among young individuals is more likely to be attributable to recent transmission than that among the elderly, who have had many more years of life during which to become infected [8, 9].

In this paper, we examine the relationship between the clustering observed at different ages in a population and the proportion of disease attributable to recent transmission using a model of the transmission dynamics of *M. tuberculosis* applied to data from The Netherlands, where isolates from all tuberculosis patients with onset since 1993 have been routinely DNA fingerprinted [10].

METHODS

The model applied here stems from work on the transmission dynamics of *M. tuberculosis* [8, 11] by Sutherland et al. [12], and assumes that individuals experience age-dependent risks of developing primary, endogenous and exogenous disease. Since immigrants probably experience different (and unknown) infection and disease risks from the indigenous population, model predictions are restricted to the Dutch native-born population. The analyses are further restricted to respiratory ('pulmonary') forms of tuberculosis, since these are far more likely to lead to transmission than are extrapulmonary forms. Given its small contribution to the tuberculosis situation in The Netherlands [13], the effects of HIV are excluded. We first describe the general epidemiological assumptions in the model and then describe how it distinguishes between cases according to the DNA fingerprint pattern of the strain causing the disease episode in order to calculate clustering statistics.

Epidemiological assumptions in the model

The model's structure is shown in Figure 1. Individuals are assumed to be born uninfected. Infected individuals are divided into those who have not yet developed primary disease (defined by convention as disease within 5 years of initial infection [14] ($I(a, t)$) and those in the 'latent' class who are at risk of endogenous reactivation and/or of reinfection followed by exogenous disease (see definitions in caption to Fig. 1). The infection and reinfection risks are assumed to be identical and depend on calendar year, but reinfection is less likely to lead to disease than is initial infection, due to some immunity induced by the prior infection [8]. We also assume that individuals cannot be reinfected whilst *at risk* of developing either the first primary episode or exogenous disease. As several studies have found a higher prevalence of tuberculin sensitivity (and by inference, higher infection risks) among adult males than for females (e.g. by up to 10% [15]), we explore the sensitivity of model predictions to the assumption that the annual (re)infection risk for females aged over 15 years was 10% lower than for males. For simplicity, the infection risk was not otherwise assumed to depend on age.

The risks of developing disease depend on age and sex (Fig. 2*a*(i)) and were estimated by fitting model predictions to notification rates observed in England and Wales since 1953 [8, personal observation]. Those analyses found no gender differences in the risks of developing disease among children and of developing the first primary episode among adults, and lower (adult) risks of endogenous and exogenous disease among females than for males (Table 1). The risks of developing either a first primary episode or exogenous disease depend also on the time since infection and reinfection respectively (Fig. 2*a*(ii)). The probability that a disease episode is infectious (sputum smear/culture-positive) is age-dependent (Fig. 2*a*(iii)) [8]. Table 1 summarizes the parameters and the assumptions used in the model.

Simulating the diversity of strains

Analyses of serial isolates from individuals with active tuberculous disease suggest that the half-life of DNA fingerprint patterns based on IS*6110* RFLP (which has been used for the DNA fingerprinting carried out to date in The Netherlands) is 2–5 years [16, 17]. Assuming a similar molecular clock speed for IS*6110* RFLP patterns of strains involved in 'latent' infection (currently unknown), this relatively short half-life implies that most of the fingerprint patterns of the strains currently causing disease are different from

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

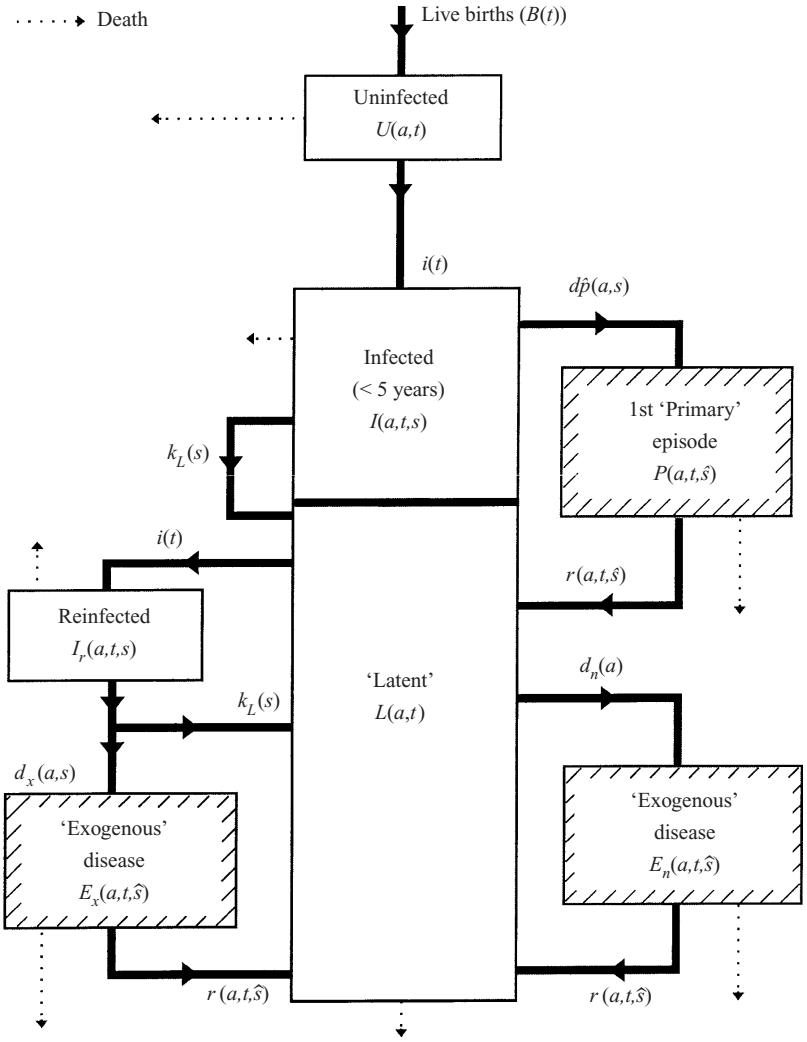

Fig. 1. Schematic diagram of the model. Primary disease is defined as disease within 5 years of initial infection [14]; exogenous disease is here defined as the first disease episode within 5 years of the most recent reinfection. Endogenous disease includes disease occurring more than 5 years after the most recent (re)infection event, and *second or subsequent disease episodes* occurring less than 5 years after the most recent (re)infection event.

those which caused disease many years ago. Similarly, it implies that the cluster distributions seen among tuberculosis cases today depend only very loosely on those which existed, e.g. 50 years ago. Following this reasoning, to derive clustering estimates for The Netherlands for the period 1993–7, the model was designed 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 were infected with a strain whose DNA fingerprint pattern had first appeared since then* and (*b*) no assumptions would be required about the distributions of strains which existed before 1950. The general steps in the calculations are outlined briefly below.

The number of individuals of each age in each of the epidemiological categories for 1950 was calculated using the model based on the equations described in [8]. From 1950, each of these age-sex classes was subdivided to describe those who had and had not been (re)infected since 1950 separately. Those who had been (re)infected since 1950 were subdivided further according to the time of infection. The transmission dynamics were tracked simultaneously for all individuals using the equations in Appendix A (using Forward Euler differencing [18]), with time steps of 6 months and 1 year for calendar year and age respectively.

In each time interval, a proportion of infected individuals was assumed to develop disease, and a proportion of these disease episodes was associated

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

Table 1. *Summary of parameter values used in the model*

Variable	Definition	Assumption
$i(t)$	Infection and reinfection rates at time t	20% until 1880, declining by 2% annually until 1911, by 5·4% annually until 1940 and 11·8% annually thereafter [15, 21]
$d_p(a, s)$	Risk of developing the first primary episode at time s after infection at age a	Depends on age and time since first infection (Fig. 2i, ii); assumed to be identical for males and females. Cumulative risks within 5 years of initial infection: 4·06%, 8·98% and 13·8% for 0–10 year olds, 15 year olds and individuals aged over 20 years respectively [8]
$d_x(a, s)$	Risk of developing exogenous disease at time s after reinfection at age a	Relationship between age and time since *reinfection*, identical to that between $d_p(a, s)$ and time since *first* infection (Fig. 2*a*(i, ii)). Cumulative risks within 5 years of reinfection for males: 6·89%, 7·57% and 8·25% for 0-10 year olds, 15 year olds and individuals aged over 20 years respectively [8]; 6·89%, 4·17% and 0·01% for the same age groups, respectively for females (personal observation)
$d_n(a)$	Annual risk of developing endogenous disease at age a	Males: $9{\cdot}82 \times 10^{-8}$%, 0·0150%, and 0·0299% for 0-10 year olds, 15 year olds and individuals aged over 20 years respectively [8]. Females: $9{\cdot}82 \times 10^{-8}$%, 0·0025% and 0·0048% for the same age groups, respectively (personal observation)
$d_+(a)$	Proportion of total disease incidence among cases aged a assumed to be infectious	10% for 0-10 year olds, increasing linearly to 65% for 20-year-olds and increasing linearly to 85% for 90-year-olds (Fig. 2*a*(iii)).
$k_L(s)$	Rate at which individuals who have been infected or reinfected for time s without developing disease move into the 'latent' class	Transition occurs exactly 5 years after infection/reinfection, i.e. $k_L(s) = 0$ if $0 < s < 5$ and ∞ for $s = 5$ years
$r(a, t, \hat{s})$	Recovery rate for cases of age a at time t at time $\hat{s}$ after disease onset	Individuals are diseased for 2 years unless they die in the meantime (see below)
$m_+(t, \hat{s})$	Case-fatality of infectious pulmonary cases at time t and time $\hat{s}$ since disease onset	Case fatality in second year after disease onset is 65% of that in first year. Overall case-fatality: 50% until 1950, declining to 30% and 25% by 1953 and 1956, respectively, and constant until 1976. Identical to mortality in general population thereafter [8]
$m_g(a, t)$	Mortality rate of non-infectious and non-diseased individuals in the general population of age a at time t	Identical to all-cause mortality (after subtracting deaths among infectious cases, estimated in the model). Annual age-specific all-cause mortality rates from 1892 obtained from the Dutch Centre for Population Statistics; data until 1892 obtained by back-extrapolation

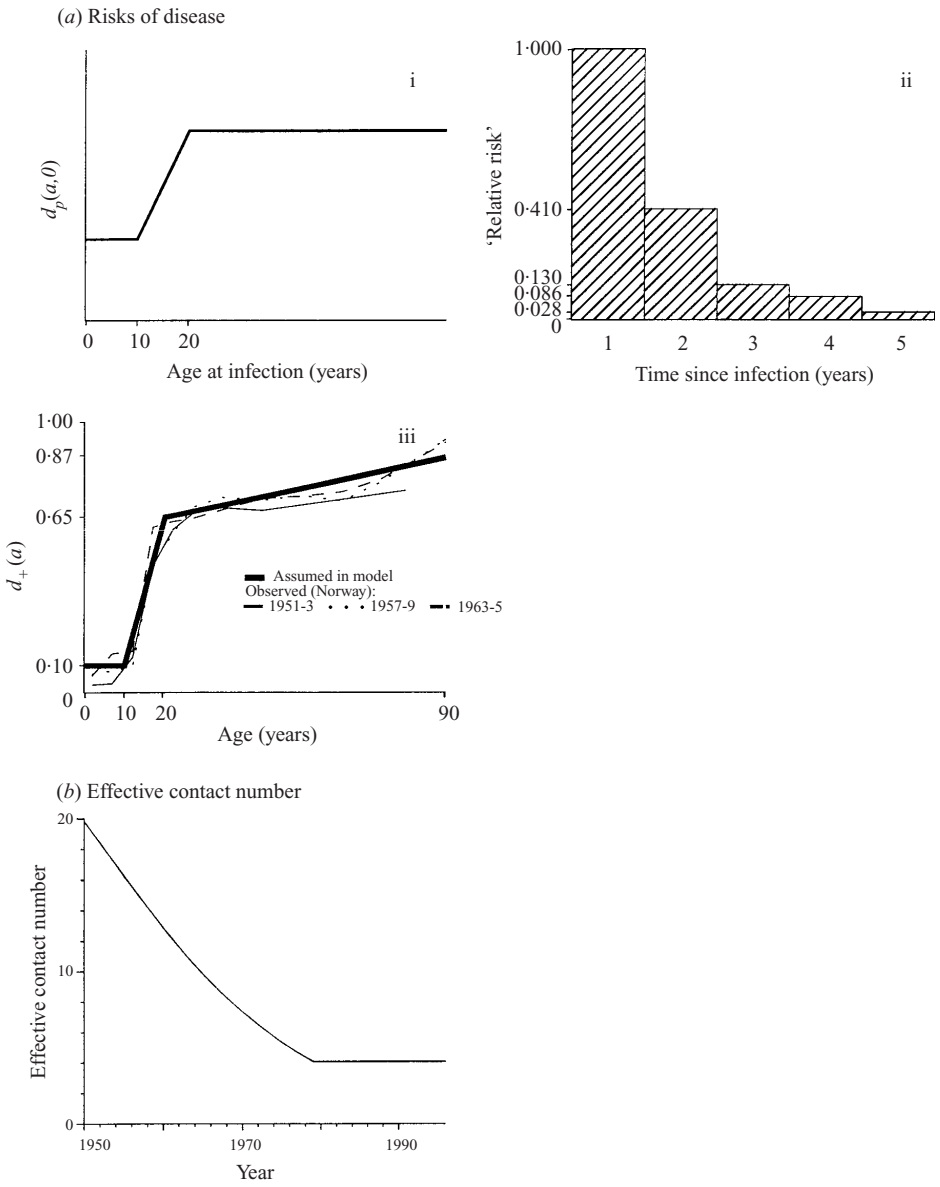

Fig. 2. Summary of the main assumptions in the model relating to (*a*) the risks of developing disease. (i) General relationship between the risk of developing the first primary episode (during the first year after infection) and age at infection. An identical relationship is assumed to hold between the risk of exogenous disease and the age at reinfection and between the risk of endogenous disease and the current age of individuals. See Table 1 for the magnitude of the disease risks. (ii) 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. Relationship derived using data from the UK MRC BCG trial during the 1950s [37]. (iii) Proportion of respiratory disease incidence manifested as sputum-positive (i.e. infectious). (Data source: the late Dr K. Styblo (TSRU) and Dr K. Bjartveit (Norwegian National Health Screening Service).) (*b*) The effective contact number. Values for the period 1950–79 are derived using the ratio between the annual risk of infection and the prevalence of infectious cases estimated by the model; the effective contact number is assumed to remain unchanged after 1979.

with a strain whose DNA fingerprint pattern differed from that with which the individuals were originally infected. This latter proportion depended on the time since infection, and each of the new DNA fingerprint patterns was assigned a unique identity number. Each infectious case with onset at a given time was assumed to contact a different number of individuals (see below) and the frequency distributions of the number of individuals contacted by each case was used to derive the total number of individuals who were newly infected/reinfected at this time. The corresponding equations were then applied to this number to determine the *total* number of individuals who developed disease at a later time T among those who

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

had been infected or reinfected at time t. The DNA fingerprint patterns of the strains in these diseased individuals were then determined using the frequency distribution of the number of individuals contacted by each case at time t. These calculations are described further below.

Modelling the infection process

Deriving the incidence of infection/reinfection in each time interval

An effective contact is here defined, as by Frost [19], as one sufficient to lead to infection if the contacted individual has never been infected. For simplicity, it was assumed that all effective contacts occurred immediately after onset of (infectious pulmonary) disease in the source case. This is a reasonable assumption for developed countries in recent years, e.g. the time interval between onset of symptoms and diagnosis for transmission-linked cases with onset between 1993 and 1996 in The Netherlands was generally less than 16 weeks [20]. The number of individuals effectively contacted by each case (defined here as the 'effective contact number') was assumed to follow a Negative Binomial distribution, defined by a time-dependent mean and variance as follows.

The mean total number of individuals effectively contacted by each case with onset at a given time t, during his/her infectious period, was calculated as the ratio between the annual risk of infection (which declined from approximately 2% in 1940 by 11·8% each year until 1979 [21, 22]), and the prevalence of infectious cases estimated for The Netherlands using the model and equations in [8] (see also [23]). In the absence of hard data, the average effective contact number was assumed to have remained unchanged since 1979 (Fig. 2*b*), and the variance of the effective contact number was taken to be the value which led to *overall* levels of clustering which compared best against those observed (see below). Figure 7*a* (Appendix) summarizes the frequency distributions of the number of individuals effectively contacted by each case implied by different values of the variance: mean ratio.

Contact patterns between individuals

The implications of different degrees of preferential mixing between individuals were explored, namely (*a*) random mixing (*b*) 'assortative' mixing, assuming that individuals who contact many individuals (e.g. those with a high-risk lifestyle) mix preferentially with similar individuals and (*c*) sex-specific assortative mixing, assuming that males contact more individuals than do females and that they mix preferentially with males. To implement these assumptions, the model kept track of the identity number of the DNA fingerprint of the strain with which the cases who contacted the different numbers of individuals were infected, for later calculations. For the assortative mixing assumptions, the number of individuals contracted by a given case depended both on the number of individuals contacted by the case who infected him/her and on the year of infection of that case, since 8 effective contacts for a case in 1990, when the average effective contact number was low, implies a 'higher' risk lifestyle than does 8 effective contacts in 1950, when the average effective contact number was relatively high. Further details of how the mixing assumptions were implemented are provided in Appendix B.

Calculating the distribution of strains among cases at a given time

It was assumed that all reactivations (which generally involve individuals infected for more than 5 years) of infections *acquired before 1950* were with unique strains and that the strain isolated from individuals who had been reinfected more than once was from the most recent (re)infection event. The DNA fingerprint pattern of the strain causing disease among each of the $C(T, t)$ cases 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 case of that individual (identified using the algorithm in Appendix C) had been infected, *unless* it had since changed through random mutations. The proportion of cases who had been infected at time t for whom the DNA fingerprint was assumed to have changed was given by $(1-e^{-0\cdot21661(T-t)})$ which describes a half-life of 3·2 years for DNA fingerprint patterns, as found in a recent study [17]. The implications of half-lives of 2, 5 and 10 years for DNA fingerprint patterns were also explored. The clustering by sex and age for cases with onset in different time periods (e.g. 1993, 1993–4, 1993–5, 1993–6 and 1993–7) was calculated using the age and sex distribution of the cases with onset in that time period (see Appendix D). These were compared against the age-specific proportion of disease attributed by the

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

model to recent transmission, defined here as the proportion of cases experiencing either primary or exogenous disease (i.e. involving disease within 5 years of the most recent (re)infection). These analyses assumed implicitly that clustered cases were involved – at some level – in the same chain of transmission, and not, e.g. as a result of preferential insertion of IS*6110* into some location in the genome.

Comparisons between observed and predicted levels of clustering in The Netherlands

Restriction fragment-length polymorphism (RFLP) analyses were used to determine the 'DNA fingerprint' of all 5122 isolates in The Netherlands in the period January 1993 to December 1997 [10]. Information on patient characteristics was retrieved from The Netherlands Tuberculosis Register maintained by KNCV (NTR/KNCV). As the NTR/KNCV does not record names, postal code and date of birth were used to link the NTR/KNCV database to laboratory results, yielding matches on 4357/5122 (85%) patients. Matching was not associated with age or sex. Patients with extrapulmonary tuberculosis only and those who were known to be HIV-positive were excluded from the data, as the complications of HIV and extrapulmonary tuberculosis were excluded from the model. The overall clustering observed within 1, 2, 3, 4 and 5 year time windows starting from 1993, 1994, 1995 and 1996 after excluding cases clustered with immigrants ('mixed clusters') during the period 1993–7 was compared against model predictions to determine the optimal variance:mean ratio for the effective contact number for the native Dutch population in The Netherlands. Model predictions of the clustering observed among male and female cases by age for the periods 1993, 1993–4, 1993–6, and 1993–7 were then compared against that observed after excluding mixed clusters. The characteristics of Dutch cases who were in mixed clusters (and therefore excluded from the data in these analyses) are described elsewhere [10, 24].

DNA fingerprinting was carried out as described in [24] using the IS*6110* insertion sequence as a probe [25]. Because the differentiation of *M. tuberculosis* strains carrying few copies of IS*6110* is poor [26–28], and subtyping of strains with a high ($\geqslant 5$) IS*6110* copy number does not typically lead to further subdivision of clusters, all strains carrying fewer than 5 IS*6110* copies ($n = 433$) were subtyped with the polymorphic GC-rich sequence (PGRS) probe [27, 28]. This further subtyping ensured that clustered strains with a low IS*6110* copy number were likely to have been involved (at some level) in the same chain of transmission, given that epidemiologic linkage between clustered strains with a low IS*6110* copy number can be rare but is likely if the clustering is also defined, e.g. by PGRS [29, 30].

Computer-assisted analysis of the IS*6110* DNA fingerprints was done with Gelcompar software, version 3.1b for Windows (Applied Maths, Kortrijk, Belgium) [26, 31]. Clusters was defined as groups of patients having isolates with identical DNA fingerprints patterns.

RESULTS

Estimates of clustering by age for different time windows

Figure 3*a* summarizes the clustering observed among male tuberculosis cases with onset during the time periods 1993, 1993–4, 1993–6 and 1993–7 in The Netherlands, after excluding 'mixed' clusters. Considering males with onset during 1993, more clustering was observed among young than old cases, e.g. 50–75% of 5-24 year olds were clustered as compared with less than 20% of individuals aged over 55 years; no 35-44 year olds were clustered. The decline in clustering with age became more regular as the size of the time window increased: considering the period 1993–7, the clustering decreased steadily from 100% for 0-4 year olds to about 50% and 20% for 25-34 year olds and those aged over 65 years respectively. The age-specific patterns among female cases (Fig. 3*b*) in each time window generally resembled those among males, except that the clustering was consistently low (25–45%) for females aged 15–64 years. For reference, the clustering observed among all cases (i.e. including mixed clusters) is also shown – this was generally higher (particularly for adults) than that observed after excluding mixed clusters, though the age-specific patterns (e.g. decreases in clustering with increasing age) were similar.

Figure 4*a* contrasts model predictions of clustering among males in different time periods against the observed data. These are based on a variance:mean ratio of 20 for the distribution of the effective contact number, which led to *overall* clustering estimates which compared best against those observed (Fig. 7*b*).

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

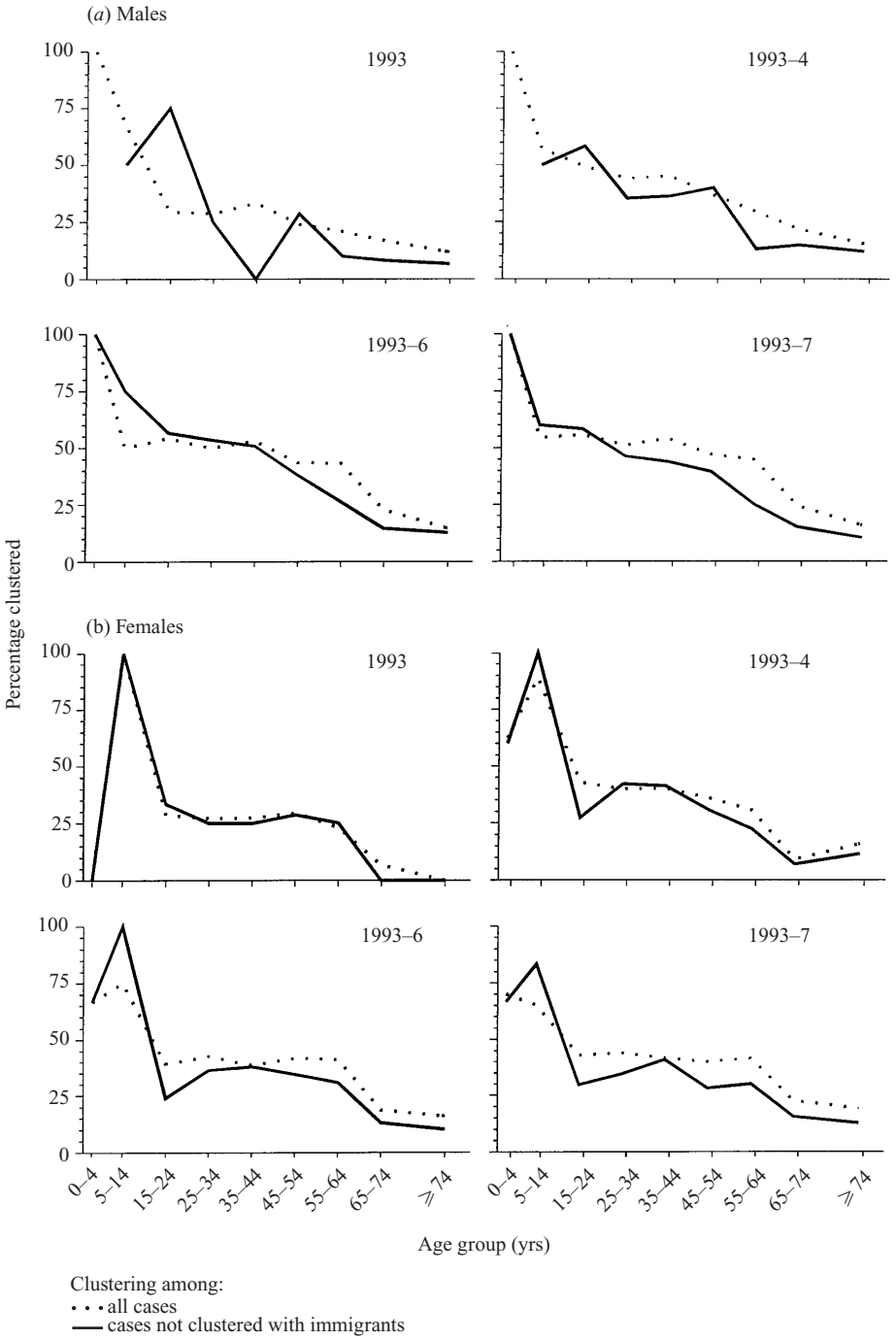

Fig. 3. Summary of the clustering observed in different age groups in The Netherlands among (*a*) male, and (*b*) female cases with onset during the time periods 1993, 1993–4, 1993–6 and 1993–7, after excluding mixed clusters. (solid line). The dotted lines show the clustering observed among all cases, including mixed clusters.

In general, for each time period and mixing assumption, model predictions compared well against the observed clustering among males (Fig. 4*a*). For males with onset in 1993, the predicted clustering declined steadily with increasing age e.g. from 40% for 0-4 year olds, to 10% for those aged over 55 years, assuming no gender differences in either the infection risks or in mixing patterns. For males aged under 55 years, the clustering based on this assumption was slightly lower than that predicted assuming that males contact more individuals than do females. For each mixing assumption, the clustering predicted in each age group increased as the size of the time window increased (e.g. to 70% and 20% for 0-4 year olds and

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

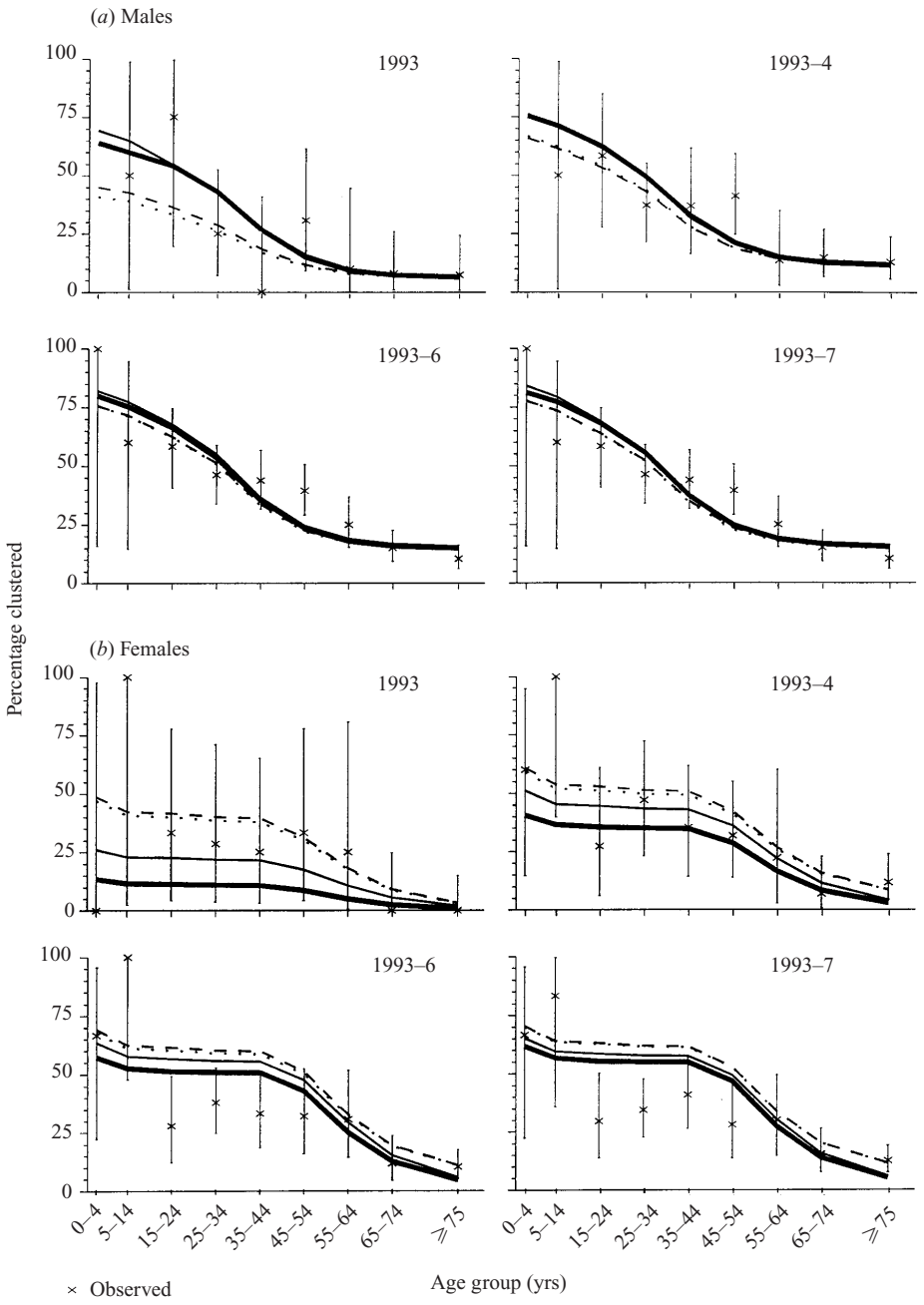

Fig. 4. Comparison between the observed and model predictions of clustering in different age groups between 1993 and 1997 in The Netherlands among (*a*) males and (*b*) females, assuming that there were no gender differences in mixing patterns (assuming random (···) or assortative (– – –) mixing) or that males were more likely to have many contacts than females, mixed preferentially with males, and experienced infection risks which were either identical (——) or 10% higher (——) than those of females. The observed data show the clustering seen after excluding mixed clusters with 95% (exact) confidence intervals. The best overall fit to the data resulted from the assumption that females contact fewer individuals than males and that their infection risks were identical (e.g. the total sum of squares of the difference between the observed and predicted age-specific clustering was 26713 *vs.* 27573, 27492 and 30297 for the other mixing assumptions).

those aged over 55 years respectively, assuming that there were no gender differences in mixing patterns). Differences between the clustering predicted assuming the various mixing patterns also decreased as the size of the time window increased.

Similar patterns were observed among females in each time window (Fig. 4*b*), except that the clustering declined less steeply between the ages 15 and 44 years than it did for males. Model predictions compared reasonably against the observed data for the time

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

periods 1993 and 1993–4 for all the mixing assumptions used; for the periods 1993–6 and 1993–7, the fit among 15-44 year olds was poor for all of the assumptions. Given that the best overall fit to the age-specific data resulted from the assumption that females contact fewer individuals than do males and that their infection risks were identical (see caption to Fig. 4) the remaining analyses are based on this assumption. This assumption also led to cluster distributions which compared well against those observed (Fig. 8 in the Appendix).

Estimates of the relationship between disease attributable to recent transmission and clustering

Figure 5 shows that the relationship between the *predicted* clustering and the *predicted* proportion of disease attributable to recent transmission (within 5 years of (re)infection) depends on both the age of the cases and the time period considered. Proportions of isolates *observed* to be clustered are also shown. For males with onset in 1993 (Fig. 5*a*), the clustering predicted among younger individuals greatly underestimated the predicted proportion of disease attributable to recent transmission (e.g. 60–70% of 0-14 year olds were clustered, whereas almost all disease was attributed to recent transmission); for cases aged over 55 years the two statistics were very similar. As the size of the time window used increased, the extent to which the clustering predicted underestimated recent transmission *decreased* for younger individuals (e.g. 70–80% of 0-14 year olds were clustered in the period 1993–6). For individuals aged over 55 years, the extent to which clustering overestimated recent transmission *increased* with the size of the time period used. For females, on the other hand, clustering *underestimated* the proportion of disease attributed to recent transmission, irrespective of the age group and time period considered (Fig. 5*b*); these underestimates were smallest for the elderly and *decreased* as the size of the time window used *increased*.

Figure 6 shows the implications of time windows of longer than 5 years for the relationship between model *predictions* of clustering among males and the proportion of disease attributable to recent transmission (held at the 1993 level) and the sensitivity to the molecular clock speed. For each molecular clock speed, the increase in clustering with the size of the time window was minimal for time periods longer than 4 years. Considering model predictions based on a half-life of 3·2 years for DNA fingerprint patterns, the clustering *in any given time period* underestimated the proportion of disease attributed to recent transmission for individuals aged under 45 years. Clustering based on time periods of longer than 5 years compared well against this proportion only for 45-54 year olds; for those aged over 55 years, model predictions of clustering tended to overestimate the proportion of disease attributable to recent transmission.

For all age groups, the predicted clustering within each time window *decreased* as the assumed half-life of DNA fingerprint patterns *decreased* (e.g. from 85% to 65% for 0-14 year olds during the period 1993–6, assuming a half-life of 10 years and 2 years respectively). The implications of the clock speed were smallest for the elderly e.g. the clustering predicted for individuals aged over 75 years was 20% and 15% for the period 1993–6 assuming half-lives of 10 and 2 years respectively. For young individuals, predictions of clustering based on a 10 year half-life compared most closely against the proportion of disease attributed to recent transmission; for older individuals, the extent to which clustering *overestimated* recent transmission *increased* with the half-life of the DNA fingerprint pattern.

DISCUSSION

There has been much discussion in recent years that the availability of DNA fingerprinting techniques for defining strains of *M. tuberculosis* should help to answer one of the most important questions in the epidemiology of tuberculosis, namely the proportion of disease which is attributable to recent transmission. To date, many studies have inferred this proportion from the clustering of DNA fingerprint patterns, with or without assuming the presence of an index case in each cluster [2, 3, 24]. The analyses presented here illustrate that the relationship between clustering and the proportion of disease attributable to recent transmission is not straightforward.

Our conclusions are based on a model of the transmission dynamics of *M. tuberculosis* which, whilst incorporating realistic assumptions relating to the epidemiology of tuberculosis (e.g. age-dependent risks of developing 'primary', 'endogenous' or 'exogenous' disease), includes several simplifications. The most important simplification in this context may be our assumption that the rate of change of DNA

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

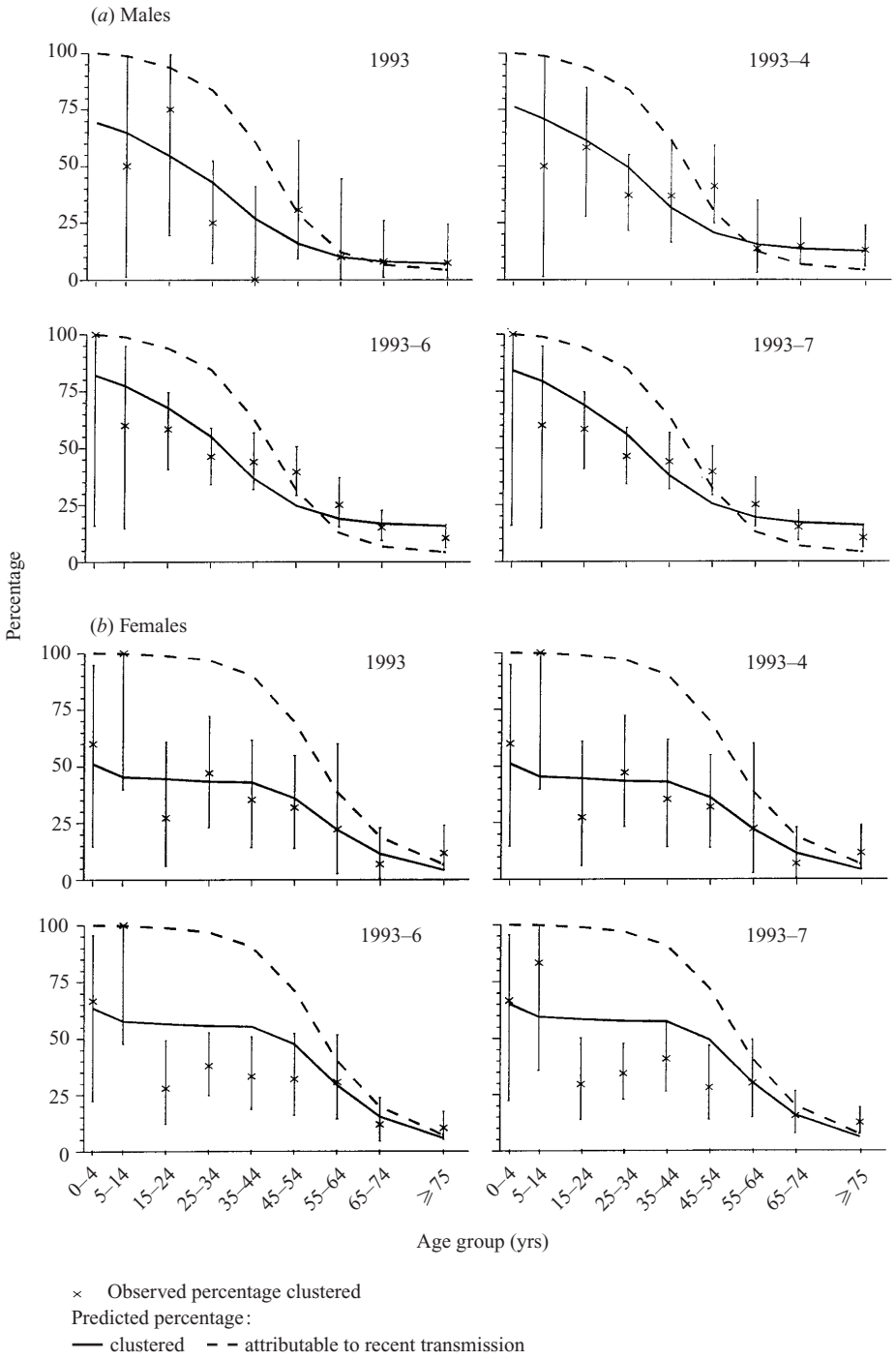

Fig. 5. Comparison between model predictions of clustering between 1993 and 1997 and the proportion of disease attributable to recent transmission (defined as disease within 5 years of (re)infection) in different age groups in The Netherlands among (*a*) males and (*b*) females. Model predictions assume that males and females faced identical (re)infection risks, but males were more likely to have many contacts than females and mixed preferentially with males. The observed data show the clustering seen after excluding mixed clusters, with 95% (exact) confidence intervals.

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

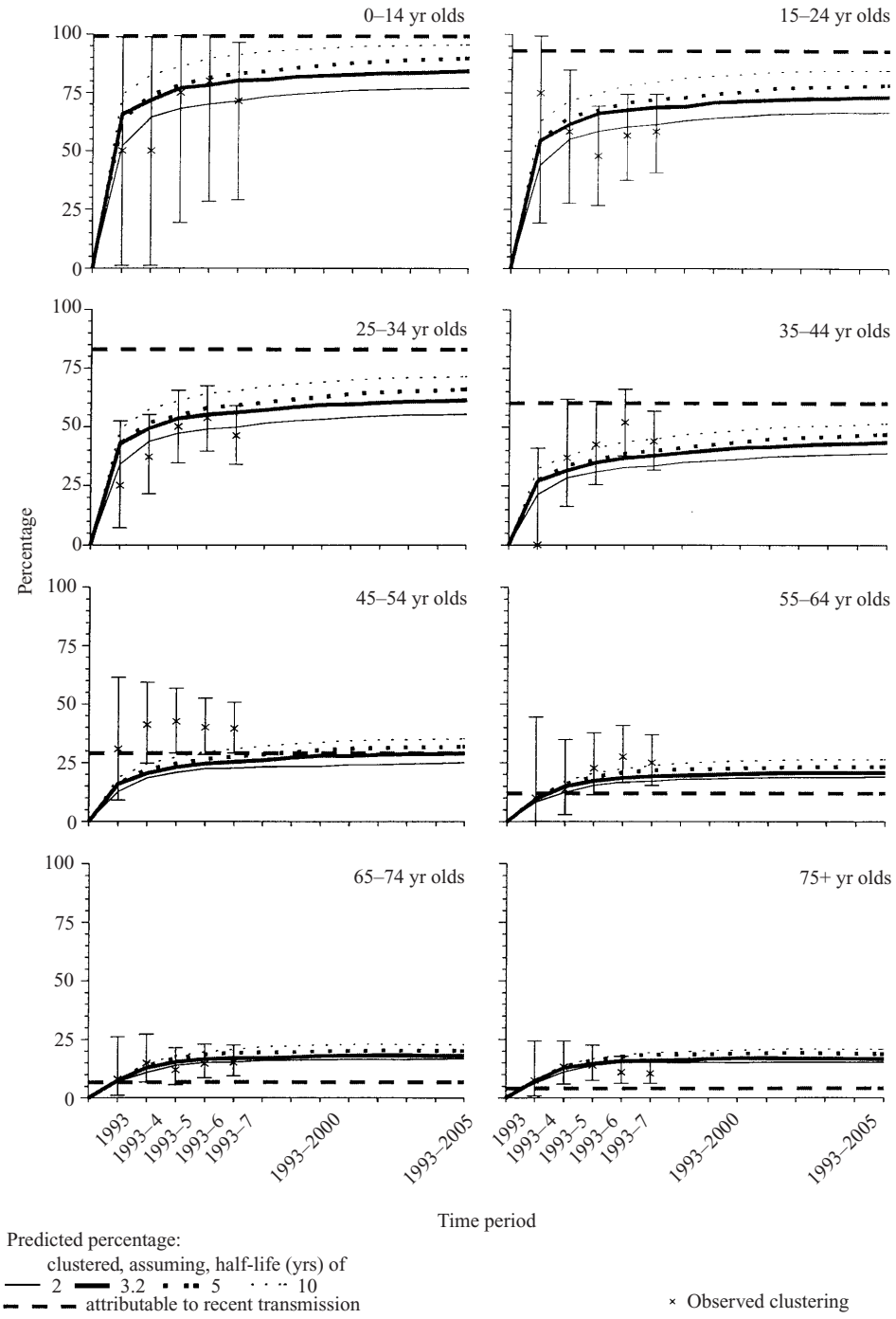

Fig. 6. Comparison between model predictions of clustering for the period 1993–2005, derived assuming different rates of change in DNA fingerprint patterns, and the proportion of disease attributable to recent transmission held at the 1993 level, in different age groups in The Netherlands among males. Model predictions assume that males and females faced identical (re)infection risks, but males were more likely to have many contacts than females and mixed preferentially with males. The observed data show the clustering seen after excluding mixed clusters, with 95% (exact) confidence intervals.

fingerprint patterns is identical for strains involved in both active disease and 'latent' infection. In addition, mixing patterns between individuals and infection risks are not assumed to be age-dependent. We discuss the implications of this assumption below. Another obvious simplification is that model predictions have been calibrated to the data observed after excluding clusters comprising immigrants. However, exclusion of mixed clusters did not affect the general pattern in the age-specific clustering (decreasing clustering with

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

age, Fig. 3) observed in The Netherlands and so would not have influenced our conclusions.

The relationship between clustering and age

That the proportion of tuberculosis disease attributable to recent transmission should be higher for young individuals than for the elderly (Fig. 5) is intuitively reasonable. Most young cases must have been infected recently [9], given that they have had relatively few years of life in which to have become infected. In contrast, most elderly cases alive today in The Netherlands were probably infected early in life, given the high risk of infection experienced in the past [15, 21], the relatively low risk of (re)infection in recent years, and the low risks of developing disease after reinfection. Though our conclusions are based on data from a low incidence country, the same logic should apply everywhere, though the age-differential should be least in populations with a high risk of tuberculous infection.

Given this relationship between age and the proportion of disease attributable to recent transmission, the finding that the clustering seen in a population in a given time interval is age-dependent makes sense, and is consistent with results from other studies [10, 32, 33]. These analyses also imply that the clustering in a given time interval is likely to underestimate recent transmission for younger individuals, and overestimate that for older individuals. The reasons for these under- and overestimates are interesting. The underestimate for younger individuals follows from the fact that some of their sources of infection would have had onset outside the study period, and thus would not be identified; as a result, the underestimate is greatest when the time window used for calculating the clustering statistics is short (Fig. 5). For older individuals, given the relatively small proportion of disease attributable to recent transmission (at least in the model) (Fig. 5), much of their overall predicted clustering is attributed to cases being *sources* of infection of other cases during the study period, and thus is likely to *overestimate* the proportion of disease attributable to recent transmission unless very short time periods are used.

Our results suggest that *contact patterns* between individuals will also influence the clustering observed in different age groups at least within relatively short time windows (Fig. 4). Age-dependent mixing patterns, such as those implied by a recent study of clusters of size 2 in The Netherlands [34], may also influence clustering statistics. Cases who contact mainly young children are also less likely to be clustered within any time window than those who contact young adults, since many years may elapse between infection in a child and onset of infectious disease [9], and the DNA fingerprint pattern of the strain with which that child is infected may well have changed by disease onset.

It is interesting that lower levels of clustering were observed for adult females than for adult males, e.g. 25–30% for 15-54 year olds even during the time period 1993–7 (Figs 3, 4). This difference may be a chance finding, but may also reflect two other factors. First, it may indicate that the proportion of disease attributable to recent infection among young adult females in The Netherlands is indeed lower than for males, either through low infection risks and/or low disease risks following recent infection. Though many studies have found a lower prevalence of tuberculin sensitivity among young adult females than for males [15], it is not known whether this reflects differences in the infection risk or in DTH response [35]. It is also unclear whether the low infection risk implied by these studies could explain the low clustering levels among females, since the clustering predicted within time windows longer than two years was similar irrespective of the (re)infection risk assumed for females (Fig. 4). It also unclear whether females face lower risks of disease after recent infection than do males. The disease risks assumed for females in these analyses were derived by fitting model predictions of the disease incidence among females (in an analogous way to that for males [8]) to notification rates in England and Wales, The Netherlands and Norway, assuming that females had the same or lower infection risks than males. In each population, the estimated risk of developing primary disease was at least as high as that for males, and those of endogenous and exogenous disease were correspondingly lower. The only other published study which has estimated the disease risks among females [12] had similar findings.

Second, the low levels of clustering among females in The Netherlands could result from relatively low transmission risks among females, either because they are less infectious than males or contact fewer individuals in the workplace or socially. In contrast, males may be likely to mix closely with (mainly) males either in occupational settings (e.g. factories, coal mines, the army), socially (e.g. in bars), or through institutional segregation (e.g. in hospitals, prisons, etc.). Following the logic described above, the low

levels of clustering among females could also occur if most of those effectively contacted by female cases were very young.

The relationship between clustering, age and study duration

There has been some discussion recently of the effect of long time windows on the *overall* levels of clustering in a population [10, 36]. The analyses presented here suggest that, as for the overall levels in a population, the clustering seen within a given age group is likely to increase with the width of the time window, but that the level reaches a plateau after a few years (Fig. 6). Given these increases, it is reasonable that the extent to which clustering under and over-estimates the proportion of disease attributable to recent transmission for young and old individuals respectively should also change over time.

Our results illustrate that the correlation between clustering and the proportion of disease attributable to recent transmission, depends greatly on the rate of change of DNA fingerprint patterns (Fig. 6). For young individuals, for example, the correlation between the two measures was closest if the half-life of DNA fingerprint patterns was 10 years, whereas for older individuals, the extent to which clustering overestimated recent transmission *increased* with the half-life of DNA fingerprint patterns. To date, the half-life for DNA fingerprint patterns (based on IS*6110* RFLP) has been estimated only for strains involved in active disease. If, as is possible, the half-life associated with 'latent' infection is considerably longer than 3 years, then clustering among the elderly could overestimate the proportion of disease attributable to recent transmission to a greater extent than that predicted in these analyses, e.g. if elderly individuals involved in the same chain of transmission many years ago were to reactivate at the same time. The relatively low risk of developing disease through reactivation [8] implies that the probability of this occurring is relatively small.

In this context it is interesting that most of the clusters involving elderly Dutch cases observed in The Netherlands to date have been relatively small. This is consistent with the low risk of reactivation, though it is recognized that it also depends on how the diversity of strains in the population in the past (e.g. the distribution of isolates with a given fingerprint pattern) changed as the annual risk of infection decreased over time, which is not yet fully understood.

On the basis of clustering of DNA fingerprints of *M. tuberculosis* strains isolated from tuberculosis cases over various time periods, several studies have concluded that approximately 30% of tuberculous disease in various developed country populations in recent years is attributable to recent transmission. As demonstrated in these analyses such conclusions may be misleading – this 'crude' proportion hides the fact that the vast majority of disease among younger individuals may be attributable to recent transmission, as compared with less than 10% of that among the elderly, and thus may not be comparable between different case series populations with different age distributions. The analyses presented here demonstrate that the extent to which clustering reflects the proportion of disease attributable to recent transmission depends also on the rate at which DNA fingerprints change over time, which is presently poorly understood, and on both the age of the cases considered and the time windows used. Given the increasing availability of DNA fingerprinting techniques, an appreciation of the strengths – and limitations – of clustering statistics is important, if they are to be used to further our understanding of the natural history of tuberculosis.

ACKNOWLEDGEMENTS

We thank the British Medical Research Council for financial support, the late Dr K. Styblo (Tuberculosis Surveillance Research Unit, The Hague) and Dr K. Bjartveit (Tuberculosis Screening Service) for supplying tuberculosis data from Norway, Dr N. Kalisvaart (TSRU) for supplying notifications from The Netherlands, and the EC Concerted Action Programme for facilitating visits.

APPENDIX A

PDEs describing the model formulation

We use the notation summarized in Table 2 to describe the transmission dynamics of *M. tuberculosis* in the model. Note that all the variables are stratified by sex; for notational convenience, we have omitted this stratification in the following description.

The equations describing the transmission dynamics are as follows:

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

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

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

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

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

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

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

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

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 rate at time t ($i(t)$) is given by $\Sigma_n nF(t,n)/N(t)$, where $N(t)$ is the total population size at time t, and $F(t,n)$ is the frequency distribution of the number of individuals contacted by the cases who had onset at time t determined for different values of the variance:mean ratio in the distribution of the effective contact number (Fig. 7). The total number of infectious cases at time t is given by the total number of individuals experiencing their first primary episode, endogenous and exogenous disease, summed over all possible ages a and times of infection T:

$$\Sigma_a\Sigma_T\{P_T(a,t,0)+E_{n_T}(a,t,0)+E_{x_T}(a,t,0)\}.$$

APPENDIX B

Implementing the assumptions relating to contact patterns between individuals

The overall frequency distribution of the number (n) of individuals contacted by cases with onset at time T ($F(T,n)$) was first subdivided into separate frequency distributions ($f(T,t,n)$) for the cases who had been (re)infected at the same time t. These frequency distributions were calculated so that they followed the overall frequency distribution of the number of individuals contacted at time T as closely as possible (each of them a Negative Binomial with the same mean and variance), and the sum of these distributions was the same as the overall distribution, as follows:

$$F(T,n)=\sum_t f(T,t,n) \tag{8}$$

To implement the random mixing assumption, cases who had been infected at the same time t were then assigned at random to contact the different number of individuals specified by the corresponding distribution $f(T,t,n)$.

To implement the assortative mixing assumption, without assuming any gender differences, cases with onset at a given time T who had been infected at the same time t were first ranked in *decreasing order of the number of individuals contacted by the case who contacted him/her*. Those who had the highest rank were then assigned *to contact the highest number of individuals*, as defined by the frequency distribution $f(T,t,n)$ above.

To implement the assumption that males contact more individuals than do females, this process was repeated separately for males and females. Thus, only *male* cases were first ranked in decreasing order of the number of individuals contacted by the case who had contacted them, and were assigned *to contact the highest number of individuals*, as defined by the distribution $f(T,t,n)$ above. The ranking process was then repeated for *female* cases and those who had the highest rank were assigned *to contact the highest number of individuals*, as defined by $f(T,t,n)$ after the contacts of male cases had been assigned.

APPENDIX C

Identifying the source of infection of a given case

The steps in the calculations were as follows:

(1) The frequency distribution of the number of

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

Table 2. *Summary of the variables used in the model formulation*

Variable name	Definition
$B(t)$	Number of live births at time t. Obtained from the Dutch Central Bureau for Statistics since 1892
$U(a, t)$	Number of uninfected individuals of age a at time t
$I_T(a, t, s)$	Number of individuals of age a at time t who were infected at time T and have been infected for time s ($\leqslant 5$ years) without having yet developed disease
$P_T(a, t, \hat{s})$	Number of individuals of age a first infected at time T, experiencing their first primary episode at time t, who have been diseased for time $\hat{s}$
$L_T(a, t)$	Number of individuals 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 more than 5 years) whose most recent (re)infection event occurred at time T
$I_{r_T}(a, t, s)$	Number of individuals of age a at time t, whose most recent reinfection occurred at time T, who have been reinfected for time s ($\leqslant 5$ years) and who have not yet developed exogenous disease
$E_{x_T}(a, t, \hat{s})$	Number of individuals of age a with exogenous disease at time t, who have been diseased for time $\hat{s}$, and whose most recent reinfection event occurred at time T
$E_n(a, t, \hat{s})$	Number of individuals of age a with endogenous disease at time t, whose most recent (re)infection event occurred at time T and who have been diseased for time $\hat{s}$

individuals contacted by each case at time t ($F(t, n)$) was first used to calculate the proportion of the infections and reinfections at time t which were attributable to individuals who had contacted $n = 1, 2, 3, \ldots$ individuals. This was given by $p(t, n) = F(t, n)n/T_c(t)$, where $T_c(t)$ is the total number of individuals effectively contacted at time t.

(2) $p(t, n)$ was then used to derive the *total number* of cases $C_n(T, t)$ with onset at time T who had been (re)infected by cases who contacted $n = 1, 2, 3, \ldots$ etc. individuals, as given by $p(t, n) \times C(T, t)$, where $C(T, t)$ is the *total number* of cases who had disease onset at time T who had been infected at time t. This implicitly assumes that the total number of secondary cases which resulted at time T, e.g. from 5 cases who had each contacted 15 individuals at time t was identical to the number attributable to infection by 15 cases who had each contacted 5 individuals at time t.

(3) The number of male and female cases assumed to have been infected or reinfected by cases who contacted $n = 1, 2, 3, \ldots$ etc. individuals was then determined, according to the assumed mixing pattern as follows.

(a) For the assumption that there are no gender differences in mixing patterns, the distribution of male and female cases among each of the $C_n(T, t)$ cases was set to be identical to that among all cases who had onset at time T and had been infected or reinfected at time t.

(b) For the assumption that males have more contacts than females and mix preferentially with males, the $C_n(T, t)$ cases for each possible time of infection t were first ranked in decreasing order of n (the number of individuals contacted at time t by the case(s) who had (re)infected them). If $C(T, t|m)$ male cases had onset at time T and had been (re)infected at time t, then it was assumed that the first $C(T, t|m)$ cases specified by the ranking of $C_n(T, t)$ were males and the remainder were females.

(4) The source of infection of each of the $C_n(T, t)$ cases was then determined, using the *cumulative* number of secondary cases which had resulted until then from each of the ($F(t, n)$) cases who had contacted n individuals at time t, assuming that (*a*) the *total* number of secondary cases which resulted from cases who contacted the same number of individuals at time t was identical and (*b*) each individual could lead only to an *integer* number of secondary cases.

For example, if 4 cases contacted n individuals each at time t, and 3 cases with onset at a later time T could attribute their infection to these cases, then it was assumed that 3 of the 4 (source) cases had each infected 1 of the 3 cases, and the other (source) had not infected any. If none of the 4 source cases had yet led to any secondary cases, then the first 3 of the 4 (source) cases at time t were assumed to have been the source cases. If another case with onset at time $T+1$ could attribute his/her infection to the 4 cases who had contacted n individuals at time t, then the 'last' of the 4 cases who had contacted n individuals at time t was assumed to have been the source of infection of that case.

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

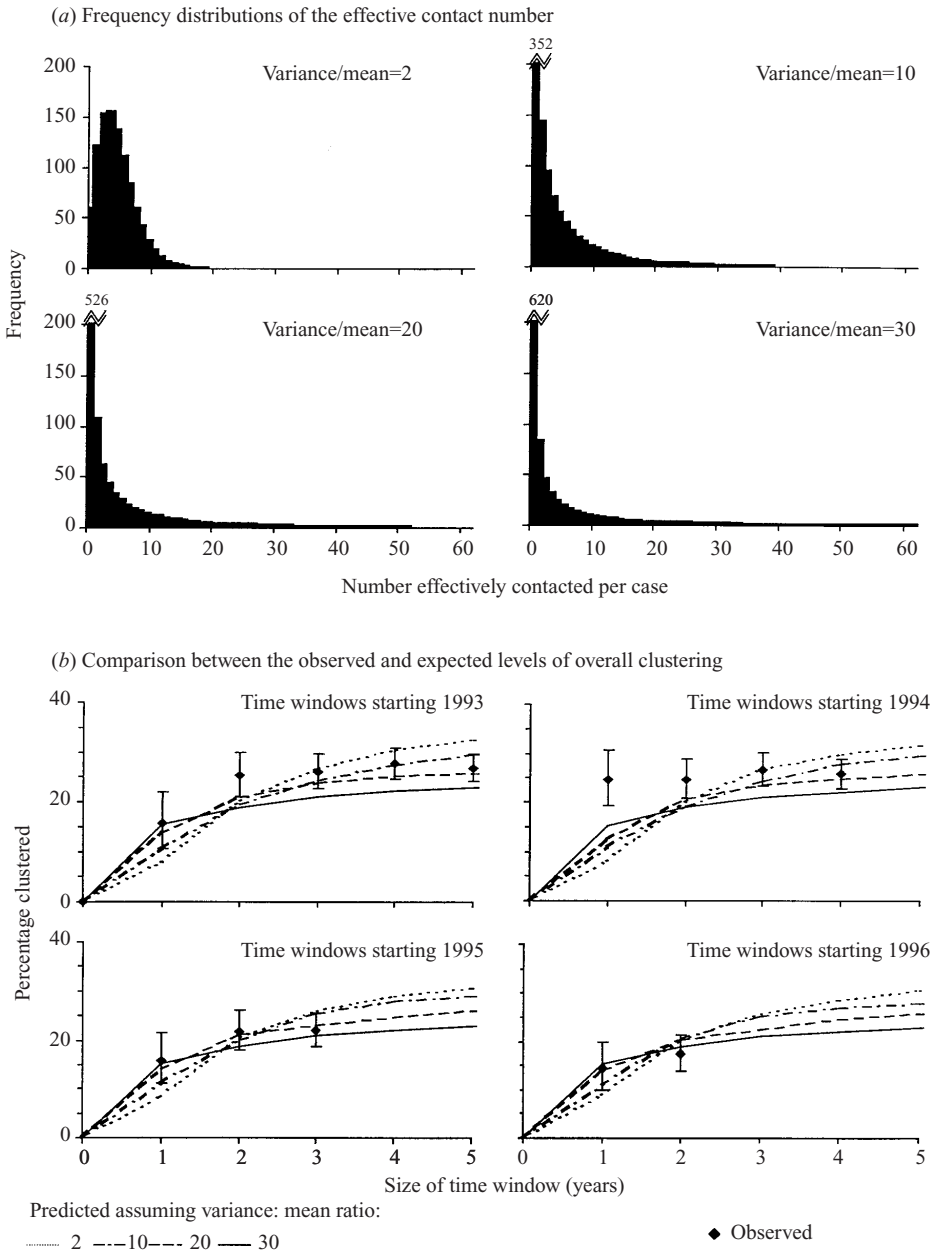

Fig. 7. (*a*) Frequency distribution of the number of individuals effectively contacted by each infectious case, in a population comprising 1000 infectious cases, assuming that it follows the Negative Binomial distribution with a variance of 2, 10, 20 and 30 times the mean (*b*) comparison between the overall clustering observed within different time windows starting from 1993, 1994, 1995 and 1996 after excluding mixed clusters and model predictions derived assuming variance : mean ratios of 2, 10, 20 and 30 for the distribution of the effective contact number. The best-fitting predictions resulted from a variance : mean ratio of 20 (sum of squares of the differences (SSq) = 225; variance : mean ratios of 2, 10 and 30 led to SSq values of 541, 356 and 293, respectively).

APPENDIX D

Calculating the clustering in different age groups

We define $p_i(T, t, a|m)$ as the proportion of male cases with onset at time T and who had been infected at time t with a strain whose DNA fingerprint pattern *had not since changed*, who were of age a.

Similarly, we define $p_d(T, t, a|m)$ as the proportion of male cases with onset at time T and who had been infected at time t with a strain whose DNA fingerprint pattern *had since changed*, who were of age a.

If there were $M_c(T)$ male cases who had onset at time T, and were clustered within the time interval $T_1 - T_2$, then the total number of male cases of *age a*

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

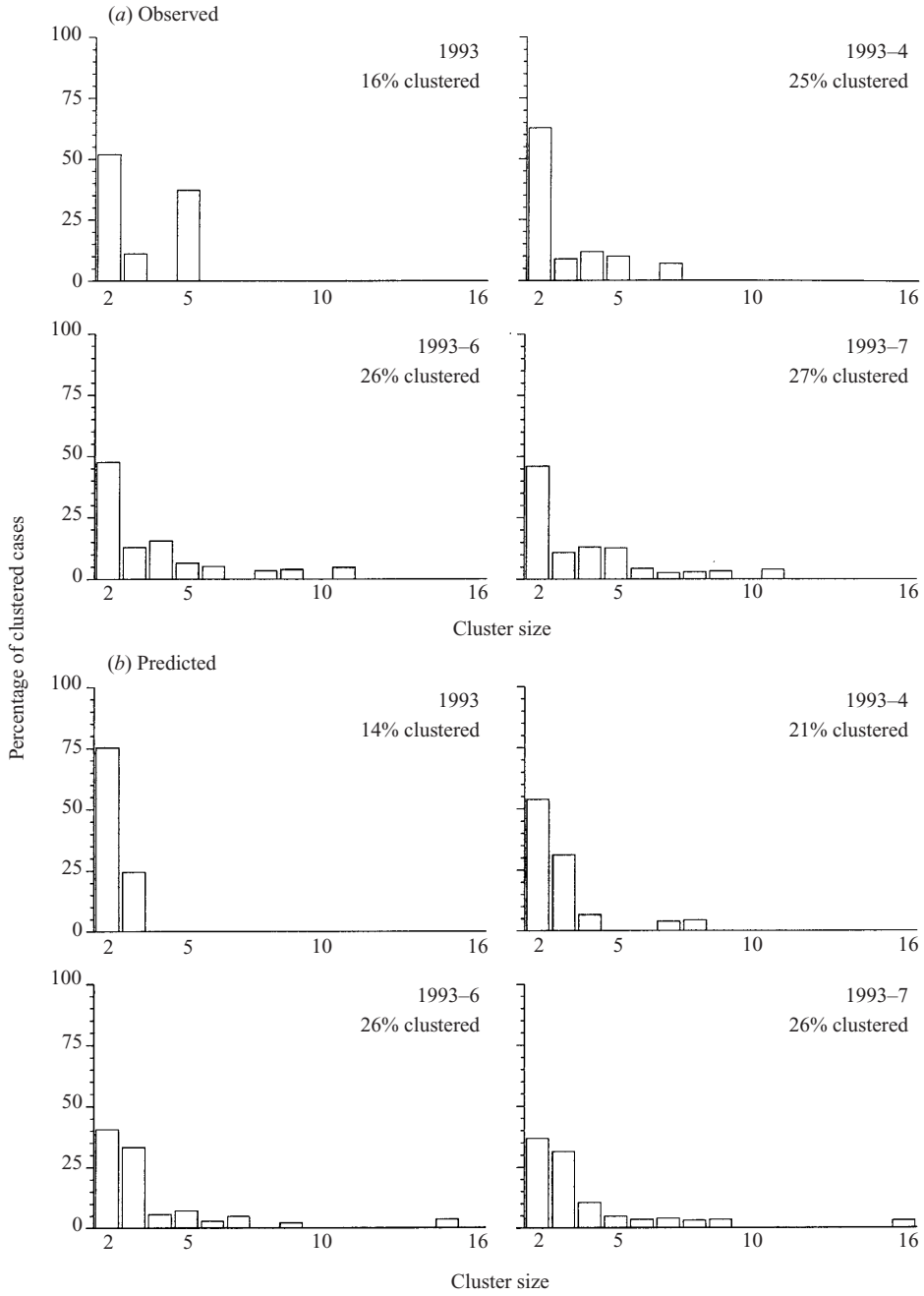

Fig. 8. Summary of cluster distributions (*a*) observed in The Netherlands between 1993 and 1997, after excluding mixed clusters and (*b*) predicted by the model.

who are clustered within the time interval $T_1 - T_2$ is given by:

$$\sum_{T=T_1}^{T_2} \sum_{c=1}^{M_c(T)} \{p_i(T, t_c, a|m) + p_d(T, t_c, a|m)\}, \quad (9)$$

where t_c is the time of infection of the cth male case who was clustered within the time interval $T_1 - T_2$.

The expressions for female cases are analogous.

REFERENCES

1. van Soolingen D, Hermans PWM, de Haas PEW, Soll DR, van Embden JDA. Occurrence and stability of insertion sequences in *M. tuberculosis* complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. J Clin Microbiol 1991; **29**: 2578–86.
2. Small PM, Hopewell PC, Singh SP, 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.

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press

3. Warren R, Haumann J, Beyers N, et al. Unexpected high strain diversity of *M. tuberculosis* in a high-incidence community. South African Med J 1996; **86**: 45–9.
4. Warren R, Richardson M, Sampson S, et al. Genotyping of *M. tuberculosis* with additional markers enhances accuracy in epidemiological studies. J Clin Microbiol 1996; **34**: 2219–24.
5. Yang ZH, de Haas PEW, Wachmann CH, van Soolingen D, van Embden JDA, Andersen ÅB. Molecular epidemiology of tuberculosis in Denmark in 1992. J Clin Microbiol 1995; **33**: 2077–88.
6. Alland D, Kalkut GE, Moss AR, 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.
7. Hope-Simpson RE. The period of transmission of certain epidemic diseases. Lancet 1948; 13 Nov.: 755–60.
8. 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.
9. Vynnycky E, Fine PEM. Lifetime risks, incubation period, and serial interval of tuberculosis. Am J Epidemiol 2000; **152**: 247–63.
10. van Soolingen D, Borgdorff MW, de Haas PEW, et al. Molecular epidemiology of tuberculosis in The Netherlands: a nationwide study from 1993 through 1997 J Infect Dis 1999; **180**: 726–36.
11. Vynnycky E, Fine PEM. The long-term dynamics of tuberculosis and other diseases with long serial intervals: implications of and for changing reproduction numbers. Epidemiol Infect 1998; **121**: 309–24.
12. Sutherland I, Švandová E, Radhakrishna SE. The development of clinical tuberculosis following infection with tubercle bacilli. Tubercle 1982; **63**: 255–68.
13. Borgdorff MW, Veen J, Kalisvaart N, Nagelkerke N. Mortality among tuberculosis patients in The Netherlands in the period 1993–5. Eur Resp J 1998; **11**: 816–20.
14. Holm J. Development from tuberculous infection to tuberculous disease. TSRU Progress Report; KNCV; The Hague, The Netherlands 1969
15. Sutherland I, Bleiker MA, Meijer J, Styblo K. The risk of infection in the Netherlands from 1967 to 1979. Tubercle 1983; **64**: 241–53.
16. Yeh RW, Ponce de Leon A, Agasino CB, et al. Stability of *Mycobacterium tuberculosis* DNA genotypes. J Infect Dis 1998; **177**: 1107–11.
17. de Boer AS, Borgdorff MW, Haas PEW, Nagelkerke NJD, van Embden JDA, van Soolingen D. Rate of change of IS*6110* genotypes of *M. tuberculosis* based on serial patient isolates and clustered strains. J Infect Dis 1999; **180**: 1238–44.
18. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in C. The art of scientific computing, 2nd edn. Cambridge: Cambridge University Press, 1992.
19. Abbey H. An examination of the Reed-Frost theory of epidemics. Human Biology 1952; **24**: 201–33.
20. ten Asbroek AHA, Borgdorff MW, Nagelkerke NJD, et al. Serial interval and incubation period of tuberculosis using DNA fingerprinting. Int J Tuberc Lung Dis 1999; **3**: 414–20.
21. Styblo K, Meijer J, Sutherland I. The transmission of tubercle bacilli: its trend in a human population. Bull Int Union Tuberc 1969; **42**: 5–104.
22. Sutherland I, Lindgren I. The protective effect of BCG vaccination, as indicated by autopsy studies. Tubercle 1979; **60**: 225–31.
23. Vynnycky E, Fine PEM. Interpreting the decline in tuberculosis during the last century – the roles of secular trends in effective contact. Int J Epidemiol 1999; **28**: 327–34.
24. Borgdorff MW, Nagelkerke N, van Soolingen D, de Haas PE, Veen J, van Embden JD. Analysis of tuberculosis transmission between nationalities in the Netherlands in the period 1993–1995 using DNA fingerprinting. Am J Epidemiol 1998; **147**: 187–95.
25. Van Embden JDA, Cave MD, Crawford JT, et al. Strain identification of *M. tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol 1993; **31**: 406–9.
26. Hermans PWM, Messadi F, Guebrexabher, et al. Analysis of the population structure of *M. tuberculosis* in Ethiopia, Tunisia and the Netherlands: usefulness of DNA typing for global tuberculosis epidemiology. J Infect Dis 1995; **171**: 1504–13.
27. Ross C, Raios K, Jackson K, et al. Molecular cloning of a highly repeated element from *Mycobacterium tuberculosis* and its use as an epidemiological tool. J Clin Microbiol 1992; **30**: 942–6.
28. Van Soolingen D, De Haas PEW, Hermans PW, Groenen PM, van Embden JD. Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of *M. tuberculosis*. J Clin Microbiol 1993; **31**: 1987–95.
29. Burman WJ, Reves RR, Hawkes AP, et al. DNA fingerprinting with two probes decreases clustering of *Mycobacterium tuberculosis*. Am J Respir Crit Care Med 1997; **155**: 1140–6.
30. Braden CR, Templeton GL, Cave MD, et al. Interpretation of restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolates from a state with a large rural population. J Infect Dis 1997; **175**: 1446–52.
31. Van Soolingen D, Qian L, De Haas PEW. Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of East Asia. J Clin Microbiol 1995; **33**: 3234–8.
32. Bauer J, Yang Z, Poulsen S, Andersen ÅB. 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.
33. Diaz R, Kremer K, de Haas PEW, et al. Molecular epidemiology of tuberculosis in Cuba outside of Havana, July 1994–June 1995: utility of spoligotyping

versus IS*6110* restriction fragment length polymorphism. Int J Tuber Lung Dis 1998; **2**: 743–50.
34. Borgdorff MW, Nagelkerke NJD, van Soolingen D, Broekmans JF. Transmission of tuberculosis between people of different ages in The Netherlands – an analysis using DNA fingerprinting. Int J Tuberc Lung Dis 1999; **3**: 202–6.
35. Fine PEM. Immunities in and to tuberculosis: implications for pathogenesis and vaccination. In: Porter JDH, McAdam KPWJ, eds. *Tuberculosis. Back to the future*. Chichester: John Wiley & Sons Ltd, 1994: 53–78.
36. Glynn JR, Vynnycky E, Fine PEM. The influence of sampling on estimates of clustering and recent transmission of *M. tuberculosis* derived from DNA fingerprinting techniques. Am J Epidemiol 1999; **149**: 366–71.
37. Sutherland I. The ten-year incidence of clinical tuberculosis following 'conversion' in 2,550 individuals aged 14 to 19 years. TSRU Progress Report; KNCV; The Hague, The Netherlands, 1968.

https://doi.org/10.1017/S095026880100499X Published online by Cambridge University Press