



## Influence of Sampling on Estimates of Clustering and Recent Transmission of *Mycobacterium tuberculosis* Derived from DNA Fingerprinting Techniques

J. R. Glynn, E. Vynnycky, and P. E. M. Fine

The availability of DNA fingerprinting techniques for *Mycobacterium tuberculosis* has led to attempts to estimate the extent of recent transmission in populations, using the assumption that groups of tuberculosis patients with identical isolates ("clusters") are likely to reflect recently acquired infections. It is never possible to include all cases of tuberculosis in a given population in a study, and the proportion of isolates found to be clustered will depend on the completeness of the sampling. Using stochastic simulation models based on real and hypothetical populations, the authors demonstrate the influence of incomplete sampling on the estimates of clustering obtained. The results show that as the sampling fraction increases, the proportion of isolates identified as clustered also increases and the variance of the estimated proportion clustered decreases. Cluster size is also important: the underestimation of clustering for any given sampling fraction is greater, and the variability in the results obtained is larger, for populations with small clusters than for those with the same number of individuals arranged in large clusters. A considerable amount of caution should be used in interpreting the results of studies on clustering of *M. tuberculosis* isolates, particularly when sampling fractions are small. *Am J Epidemiol* 1999;149:366–71.

bias (epidemiology); epidemiologic methods; epidemiology, molecular; study design; tuberculosis

Clustering of *Mycobacterium tuberculosis* strains using DNA fingerprinting techniques has been used to estimate the proportion of tuberculosis cases likely to be attributable to recent transmission of infection in a population (1, 2). People with isolates considered identical to those of others in the study population, by whatever typing method is used, are said to be "clustered," and those with unique isolates are said to be "nonclustered." The degree of clustering is thought to be related to the extent of recent transmission of *M. tuberculosis*, and it has been calculated as the proportion of all tuberculosis patients who are included in clusters, with or without a correction factor to take into account the possible presence of an index case within each cluster.

Several studies have now estimated clustering in different populations over different time periods (1–5). The reliability of such estimates depends on case ascertainment and inclusion: although contact tracing will overestimate the proportion clustered, incomplete ascertainment will underestimate clustering, and the extent of

underestimation will depend on the cluster size distribution in the population. We have explored the effect of different levels of case ascertainment and inclusion on estimates of clustering, using a stochastic simulation model applied to published data from real populations and to hypothetical populations with clusters of different sizes.

### MATERIALS AND METHODS

Random samples of different sizes were obtained from predefined populations using a random number function. For each sampling fraction, 1,000 samples were obtained from each population, and the mean values and 95 percent confidence intervals for the observed proportions of tuberculosis cases clustered were calculated. Without a correction factor (the "n" method), the proportion clustered is given by the number of patients in clusters divided by the total number of individuals. With the correction factor method (the "n – 1" method), it is assumed that a cluster of size  $n$  contains  $n - 1$  individuals with recent infection and one individual with an old infection (1). All simulations were conducted using both the "n" and "n – 1" methods of estimating recent transmission.

We do not know the true distribution of cases in any population by "strain," so we have used published data from two populations with different reported cluster pat-

Received for publication September 18, 1997, and accepted for publication June 15, 1998.

From the Infectious Disease Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, England, United Kingdom. (Reprint requests to Dr. Judith R. Glynn at this address).

terms, though similar overall clustered proportions, as our starting populations, and have further explored the effect of cluster size in a series of artificial populations.

Our first example uses the cluster distribution published for San Francisco, California, for 1991–1992 (1). In this population, after exclusion of patients with a single band on the fingerprint and those whose culture result was thought to reflect contamination, 191 (40.4 percent) of 473 patients with available fingerprints were in clusters, with cluster sizes ranging from two patients to 30 patients (table 1). There were 44 clusters, giving a cluster proportion, as estimated by the “ $n - 1$ ” method, of 31.1 percent  $((191 - 44)/473)$ . The 473 patients included are themselves a sample: over the time period of the study, 688 cases of tuberculosis were reported to the authorities (“notified cases”), and 585 of those were culture-confirmed.

The second example is based on the cluster distribution observed recently in a high incidence area of Cape Town, South Africa (3). Fingerprints were available from 90 percent of culture-proven tuberculosis patients identified in the community over a period of 18 months. In this study, 49 patients with restriction fragment length polymorphism patterns with less than five bands were excluded, leaving 44.2 percent (126/285) clustered in 40 clusters, or 30.2 percent  $((126 - 40)/285)$  by the “ $n - 1$ ” method (table 1). The number of non-culture-confirmed cases identified over this period in the Cape Town study is not stated in the published article (3), and the numbers of cases of tuberculosis that were undiagnosed or unreported in these study areas are also unknown.

Additional theoretical starting populations were created to investigate the influence of cluster size on the effect of different sampling fractions. Hypothetical pop-

ulations of 100 individuals were arranged as pairs, triplets (totaling 99 individuals), fours, or fives, giving a “true” proportion clustered of 100 percent. Formulae for calculating the expected proportions clustered for different sampling fractions taken from populations with different underlying cluster structures are given in the Appendix.

## RESULTS

Using the San Francisco data and the “ $n$ ” method, the mean proportion of cases observed to be in clusters was 28 percent with 30 percent ascertainment, 34 percent with 50 percent ascertainment, and 38 percent with 80 percent ascertainment. Compared with the 40 percent reported, these reflect proportionate reductions of 30 percent, 17 percent, and 5 percent, respectively (figure 1, table 2). The proportionate underestimation of clustering was greater with the “ $n - 1$ ” method than with the “ $n$ ” method (figure 2, table 2). With increasing ascertainment, as the mean estimates of clustering increase, the range of likely results decreases. For example, using the “ $n$ ” method (figure 1), with 30 percent ascertainment, 95 percent of the estimates of clustering were between 21 percent and 36 percent, and with 50 percent ascertainment, the equivalent range was 28 percent to 39 percent.

With the Cape Town data, the underestimation of clustering produced by sampling was more dramatic than that seen with the San Francisco data (table 2). Using the hypothetical populations of 100 individuals arranged as pairs, triplets, or sets of four or five, we demonstrated that the influence of sampling depends on cluster size (figure 3). For example, in these popu-

**TABLE 1. Numbers and sizes of clusters of tuberculosis patients seen in San Francisco, California,\* and Cape Town, South Africa,† as defined using DNA fingerprint patterns**

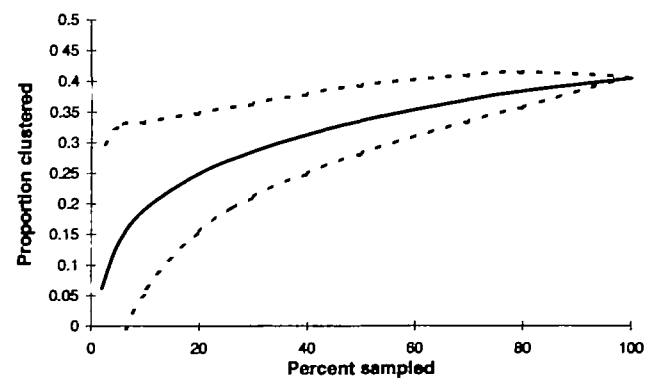
Cluster size (no. of patients)	No. of clusters	
	San Francisco‡	Cape Town§
2	20	22
3	13	8
4	4	4
5	2	2
6		1
7		1
8	1	1
10	1	
11		1
15	1	
23	1	
30	1	

\* Data were obtained from Small et al. (1).

† Data were obtained from Warren et al. (3).

‡ Isolates with single bands were excluded.

§ Isolates with fewer than five bands were excluded.



**FIGURE 1.** Simulation model estimates of the effect of sampling on the calculated proportion of tuberculosis cases clustered in a San Francisco data set (1991–1992), using the “ $n$ ” method. The mean values (—) and 95 percent confidence intervals (---) from 1,000 simulations are shown. Data were obtained from Small et al. (*N Engl J Med* 1994;330:1703–9).

**TABLE 2. Simulation model-derived estimates of the percentage by which the extent of tuberculosis patient clustering is underestimated due to random sampling in populations with different underlying cluster distributions\***

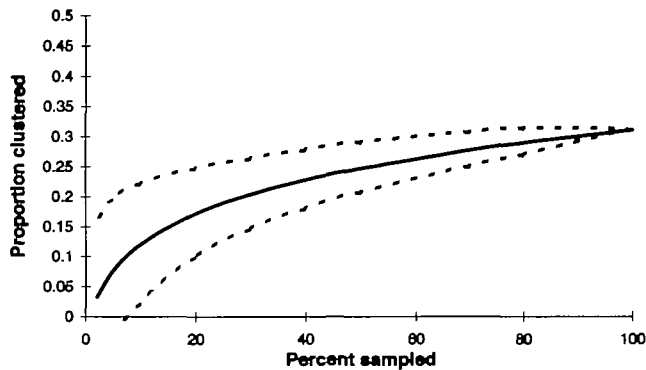
% sampled	San Francisco, California‡	Cape Town, South Africa§	Arrangement of clusters†			
			Pairs	Triplets	Fours	Fives
80	5 (7)	8 (10)	20 (20)	4 (12)	1 (8)	0.2 (6)
50	17 (20)	25 (31)	50 (50)	25 (38)	13 (30)	7 (24)
30	30 (35)	43 (51)	70 (70)	50 (60)	36 (53)	25 (45)

\* Results are presented for the "n" method, with results for the "n - 1" method shown in parentheses.

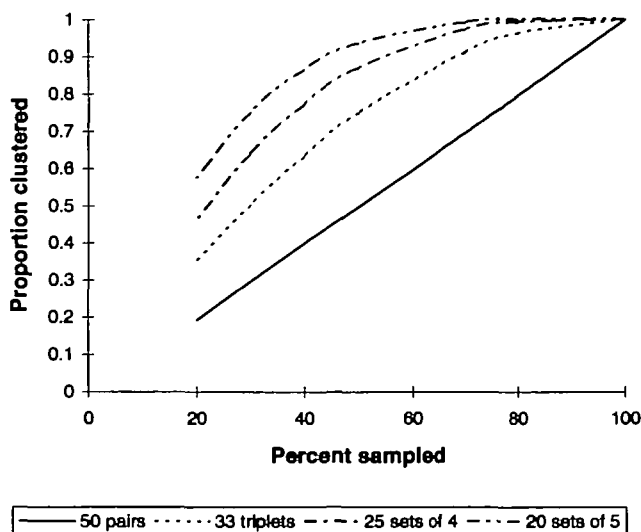
† Hypothetical populations.

‡ Data were obtained from Small et al. (1).

§ Data were obtained from Warren et al. (3).



**FIGURE 2.** Simulation model estimates of the effect of sampling on the calculated proportion of tuberculosis cases clustered in a San Francisco data set (1991–1992), using the "n - 1" method. The mean values (—) and 95 percent confidence intervals (---) from 1,000 simulations are shown. Data were obtained from Small et al. (*N Engl J Med* 1994;330:1703–9).



**FIGURE 3.** Simulation model estimates of the influence of sampling proportion on the calculated mean proportion of tuberculosis cases clustered, using the "n" method, in theoretical populations with different cluster patterns.

lations, 50 percent ascertainment gives an average of 50 percent clustering when the individuals are arranged as 50 pairs, compared with 75 percent when they are arranged as 33 triplets, 87 percent when arranged as 25 groups of four, and 93 percent when arranged as 20 groups of five. Moreover, for a given sampling fraction, the coefficient of variation (the standard deviation expressed as a proportion of the mean) decreased with increasing cluster size (results not shown). The relative underestimation in clustering was again greater using the "n - 1" method than using the "n" method, except for the population of pairs, for which the relative underestimation was the same for the two methods (table 2). For a given population cluster structure and sampling proportion, the total number of individuals in a population made little difference to the expected proportion clustered (see Appendix).

## DISCUSSION

Ascertainment of cases of tuberculosis in a population will never be 100 percent complete, and of those diagnosed, only a certain proportion will have cultures available for DNA fingerprinting. Estimates of clustering are therefore always based on a sample of all cases in a given community. These samples will not be random. People with tuberculosis in some sections of the community will be more likely to have their disease ascertained than others, and active contact tracing may be used. Both of these factors will tend to *increase* the clustering seen. However, in this paper, we have attempted to quantify the influence of *random* sampling, which will tend to *decrease* the estimate of clustering.

For a given population, as the sampling fraction increases, the (average) proportion found to be clustered also increases. Using the "n - 1" method of estimating clustering, the underestimation due to incomplete sampling is more marked than that with the "n" method (except in the artificial situation of populations

existing entirely of pairs of individuals, as shown in table 2). The extent of the underestimation of clustering depends on the cluster structure of the population: the influence of incomplete sampling is weaker in populations with large clusters than in those with small clusters. The presence of a few very large clusters has a strong influence on the results: the underestimation of clustering produced by incomplete sampling in the San Francisco data was less than that produced in a population consisting entirely of groups of five, using the " $n - 1$ " method (table 2).

As the sampling proportion increases, the variance of results obtained decreases. The proportion clustered measured in a population in which the sampling *proportion* is small is therefore difficult to interpret, even if the *number* of patients included is large. In general, the total number of patients included will not affect the expected proportion found to be clustered for any given sampling fraction, except when the numbers are very small (see appendix table).

Another aspect of sampling is the use of different time periods during which cases are recruited. These periods have ranged from 1 month (4) to several years (5) in published studies of tuberculosis. When very short time intervals are used, it is unlikely that observed clusters will contain successive cases in a chain of transmission. Because of the long incubation period of tuberculosis, clustered patients identified within a short time period are more likely to have been infected from a common source than from a member of the cluster. Under these circumstances, cluster size is expected to be small, the " $n - 1$ " method becomes inappropriate, and the proportion of cases observed to be clustered is likely to be small. This has been shown empirically in the Netherlands, where the percentage of cases clustered in a 3-month period was 20 percent, but the cumulative percentage clustered rose to >40 percent over 2.5 years (5). With an increase in study duration, an increase in observed clustering is to be expected, but the interpretation of clustering as reflecting recent transmission changes. The cumulative percentage clustered should ultimately reach a plateau for every population, and the level of that plateau is of considerable interest, as it must depend on the stability (molecular clock) of the marker used, the extent of *M. tuberculosis* transmission, the amount of migration, the types of tuberculosis included (infectious and non-infectious; primary, endogenous, and exogenous), and, for any empirical study, the sampling fraction of isolates.

Many studies exclude patients with extrapulmonary tuberculosis, who are generally considered noninfectious. It is not immediately clear what effect this has on the observed cluster patterns. On the one hand,

noninfectious cases are less likely to be clustered than infectious cases, since they cannot contribute as index cases; but on the other hand, the inclusion of noninfectious cases will increase the average number of cases attributable to each infectious case, thereby increasing observed clustering and cluster size. In studies carried out over short time periods, the latter effect of increasing the amount of clustering is likely to be the more important, since the role of index cases is negligible. In longer term studies, the net effect of including extrapulmonary cases is harder to predict. It may not be constant from place to place, since it will depend on the preexisting cluster structure of pulmonary cases and on the groups of people who contract tuberculosis in the community. The distribution of patients by age, ethnic group, human immunodeficiency virus infection status, and other factors that can be expected to influence both 1) the proportions of patients with different types of tuberculosis and 2) the risk of being infected with *M. tuberculosis* and developing active tuberculosis would be expected to influence the effect of including or excluding extrapulmonary cases. The different incubation periods of the different types of disease can also be expected to influence the effect of including noninfectious cases, depending on the time windows used. Similar arguments apply to the interpretation of studies including all pulmonary cases and those including only patients who are smear-positive.

In any real population, one of the biggest influences on the proportion clustered will be the amount of migration, since immigrants infected in other populations are likely to carry different strains, and emigrants may leave after transmitting their strain to others within the population and might not themselves be included in the study. If we exclude this influence, and the influences of time windows and sampling, by supposing that we could examine clustering by comparing all strains in a closed population, with full ascertainment of tuberculosis over a long period of time, what proportion would we expect to see clustered? It would not be 100 percent, since there will always be some cases who acquire an infection or develop disease with a newly mutated strain of *M. tuberculosis* and who fail to transmit that strain to anyone who subsequently manifests disease. The prevalence of such people in the population must depend on the molecular clock of the marker system, the transmission dynamics of tuberculosis in the population, and the types of tuberculosis included.

In the San Francisco study (1), 69 percent of diagnosed and notified cases of tuberculosis were included in the published analysis. It is likely that some cases remained undiagnosed, and that some diagnosed cases

were not reported. If 90 percent of all tuberculosis cases in the community were notified cases, we can estimate that the proportion of cases included in the study was 62 percent. If the true distribution of cluster sizes was similar to that in the 473 patients included, then a 62 percent sample may have underestimated the proportion clustered by approximately 10 percent (or 12 percent using the “ $n - 1$ ” method). In the Cape Town study (3), 77 percent of culture-proven cases were included, but the total number of patients was not stated. Using the number of smear-positive cases, and assuming that approximately 50 percent of cases are smear-positive, the proportion included is calculated to have been 57 percent, or 51 percent if we allow for 90 percent case finding. If the true cluster structure was similar to that of the 285 cases included, then the proportion clustered may have been underestimated by as much as 25 percent (or 31 percent with the “ $n - 1$ ” method). Of course, we do not know the underlying cluster structures, and cases included will not have been randomly selected.

Interpretation of cluster results requires a considerable amount of caution. We need to know as much as possible about who was included, how their cases were ascertained and over what time period, and what the sampling proportion was; and we need to appreciate the fact that low sampling proportions not only underestimate clustering but also produce estimates with wide confidence intervals.

#### ACKNOWLEDGMENTS

J. R. G. was supported by the British Department for International Development. E. V. was supported by the British Medical Research Council.

The authors thank Basia Zaba for help in designing the model and Ian White for help with the Appendix.

#### REFERENCES

- 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.
- 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;330:1710-16.
- Warren R, Hauman J, Beyers N, et al. Unexpectedly high strain diversity of *Mycobacterium tuberculosis* in a high-incidence community. *S Afr Med J* 1996;86:45-9.
- Frieden TR, Woodley CL, Crawford JT, et al. The molecular epidemiology of tuberculosis in New York City: the importance of nosocomial transmission and laboratory error. *Tuber Lung Dis* 1996;77:407-13.
- van Soolingen D. Molecular epidemiology of tuberculosis in a low incidence country: a nationwide study on transmission of tuberculosis between immigrants and native population in the Netherlands. In: Use of DNA fingerprinting in the epidemiology of tuberculosis. (Doctoral thesis). Utrecht, The Netherlands: Faculty of Biology, University of Utrecht, 1996:173-95.

#### APPENDIX

##### Estimating the Clustering Expected in a Random Sample from a Population with Different Cluster Distributions

In general, the proportion clustered by the “ $n$ ” method in a given population is given by

$$1 - P_{ucl}, \quad (1)$$

where  $P_{ucl}$  is the proportion of isolates which are unclustered in that population. To show that the average proportion of isolates which are clustered in a random sample from a population is smaller than the proportion clustered in the original population, we show that the average proportion *unclustered* in a random sample is bigger than the proportion *unclustered* in the original population.

It can be shown that the expected proportion unclustered ( $P_{ucl}$ ) by the “ $n$ ” method in a random sample of size  $R$  from a population consisting of  $c + u$  strains, of which  $c$  are clustered and  $u$  are unclustered, is given by

$$P_{ucl} = \sum_{i=1}^{c+u} p_i/R, \quad (2)$$

where  $p_i$  is the probability of getting exactly one isolate of strain  $i$  in the random sample.

Equation 2 can be expressed as

$$P_{ucl} = \sum_{i=1}^c p_i/R + \sum_{i=c+1}^{c+u} p_i/R. \quad (3)$$

Note that if there are  $n_i$  cases of strain  $i$  in the population, which is of size  $N$ , then  $p_i$  is given by the hypergeometric probability:

$$\binom{n_i}{1} \binom{N - n_i}{R - 1} / \binom{N}{R}$$

i.e., (number of ways of getting exactly one isolate of strain  $i$  in a random population)  $\times$  (number of ways of choosing  $R - 1$  isolates that are not of strain  $i$ ) / (total

number of ways of choosing  $R$  isolates from the whole population).

For the unclustered strains  $i = c + 1, c + 2, \dots, c + u$ ,  $p_i$  is given by

$$\binom{N-1}{R-1} / \binom{N}{R} = \frac{(N-1)!}{(R-1)!} \times \frac{R!}{N!} = R/N.$$

Thus, substituting for  $p_i$ , we see that the second term of expression 3 is given by

$$\sum_{i=c+1}^{c+u} p_i/R = \sum_{i=c+1}^{c+u} \frac{1}{R} \frac{R}{N} = u/N,$$

where  $u/N$  is the same as the proportion of strains which are unclustered in the original population. Substituting this expression into equation 3, we see that the following holds:

$$P_{uct} = \sum_{i=1}^c p_i/R + u/N \geq u/N.$$

Thus, the average proportion of isolates which are unclustered in the random sample is larger than the proportion which is unclustered in the original population.

Expression 1 simplifies to  $(R-1)/(2K-1)$  when the population is made up of  $K$  pairs and to  $(R-1)(6K-R-2)/\{(3K-1)(3K-2)\}$  when it is composed of  $K$  triplets. It can be shown that the expressions for the proportion clustered by the "n-1" method in these two populations are given by:  $(R-1)/\{2(2K-1)\}$  (if isolates are arranged in pairs) and  $(R-1)(9K-R-4)/\{3(3K-1)(3K-2)\}$  (if isolates are arranged as triplets). The effect of the original population size on the expected percentage clustered for different sampling fractions, if the isolates are arranged as triplets, is given in the appendix table.

**APPENDIX TABLE.** Effect of original population size on estimates of the percentage of tuberculosis patients clustered by the "n" and "n-1" methods for different sampling fractions in a theoretical population arranged in triplets\*

Sampling fraction (%)	Original population size							
	15		30		99		999	
	n	n-1	n	n-1	n	n-1	n	n-1
80	96.7	58.4	96.3	58.5	96.1	58.6	96.0	58.7
60	83.5	46.9	83.7	47.5	83.9	47.8	84.0	48.0
33	50.5	26.4	53.2	28.1	54.9	29.2	55.5	29.6

\* Calculated using the expressions shown in the Appendix.