Methodological Problems in the Molecular Epidemiology of Tuberculosis

Megan Murray1,2 and David Alland3

In systematic studies of the molecular epidemiology of tuberculosis, DNA fingerprinting is used to estimate the fraction of incident cases attributable to recent transmission of Mycobacterium tuberculosis rather than reactivation disease and to identify risk factors for recent transmission. This approach is based on the premise that tuberculosis cases that share a DNA fingerprint are epidemiologically related while cases in which fingerprints are unique are due to remote infection that has reactivated. In this paper, the authors review the objectives and design of molecular epidemiologic studies of tuberculosis, describe current analytical approaches, and consider the impact of these different approaches on study results. Using data from a previously published investigation of the epidemiology of tuberculosis conducted from 1990 to 1993 among tuberculosis patients in New York City, New York, the authors show how selecting different measures of disease frequency, comparison groups, and sampling strategies may impact the results and interpretability of the study. They demonstrate ways to conduct sensitivity analyses of estimated results and suggest strategies that may improve the usefulness of this approach to studying tuberculosis. Am J Epidemiol 2002;155:565–71

The molecular epidemiology of infectious disease uses molecular markers to track the transmission of specific strains of infectious organisms. This information is often used to describe the distribution of these strains in human populations and to evaluate host- and parasite-specific risk factors for disease spread. In the past, efforts to type strains of Mycobacterium tuberculosis in human hosts were hampered by the lack of a strain-specific immune response and by an apparent lack of genetic polymorphism in the organism (1). However, during the past decade, a number of polymorphic sites have been identified in repetitive sequences in the genome (2). The most widely used polymorphic marker is the transposable element IS6110, which varies in both copy number and location in the genome. On the basis of observations in the laboratory as well as among patients during outbreaks, this highly diverse fingerprint is generally considered stable enough to be used as a marker of epidemiologic links among tuberculosis cases (3). Over the past decade, many clinical investigators have used IS6110 to document unsuspected transmission links (4, 5), nosocomial spread of the organism (6), spread of drug resistance (7, 8), and occurrence of reinfection in both immunocompetent and immunocompromised patients (9–12).

Increasingly, molecular epidemiologists have gone beyond traditional outbreak analyses and have carried out systematic studies designed to address specific epidemiologic questions or test hypotheses. By using this “population-based” approach, researchers have applied IS6110 typing to quantify the relative contributions of recent and remote infection to the burden of tuberculosis disease in communities, to identify risk factors for disease spread, and to establish the relative frequency of reinfection. This paper reviews the objectives and design of molecular epidemiologic studies of tuberculosis to assess their effectiveness in providing useful epidemiologic measures and to suggest strategies that may make these studies more informative.

We use a study of the molecular epidemiology of tuberculosis in New York to illustrate these points (13). This paper reports on one of the first “population”-based studies to use a rigorous epidemiologic approach to data collection and analysis, and it has served as a prototype for subsequent work in this field. We provide a commentary, suggest possible methodological problems, and outline potential solutions (13).

Design of molecular epidemiologic studies

A major goal of population-based molecular epidemiologic studies of tuberculosis has been to use molecular methods to distinguish between reactivation tuberculosis and recent transmission. Another important objective has been to estimate measures of association for risk factors for recent transmission of tuberculosis. In principle, the purpose of reporting these associations is to identify people at high
risk of being infected so that control measures can be targeted to this vulnerable population. Studies in which molecular methods have been used to classify isolates as recently transmitted or reactivated have generally followed a standard approach. Cases of tuberculosis are recruited from a defined area over an extended period of time. Sampling strategies have ranged from complete ascertainment of all available cases listed in a hospital- (13), city- (14), or country- (15) based tuberculosis registry to convenience sampling of a small proportion of the available cases (16–20). Enrolled cases submit isolates of \( M. \) \( \text{tuberculosis} \) that are cultured and fingerprinted by using IS\( 6110 \) or some set of markers. Once all cases have been typed, fingerprints are compared and each isolate is classified as clustered (i.e., sharing a fingerprint with another isolate in the study sample) or unique. Most researchers interpret clusters to be epidemiologically linked chains of recently transmitted disease and unique isolates to be cases of reactivation disease resulting from “remote” tuberculosis infection. This interpretation clearly depends on the heterogeneity of the molecular markers used for typing. For example, multiple studies have suggested that resolution of IS\( 6110 \) may be impaired when copy numbers are low (21–24). Furthermore, since there is no “gold standard” for epidemiologic relatedness, the sensitivity and specificity of the markers have yet to be determined. Investigators have also assessed risk factors for recent disease by comparing the distribution of an exposure in clustered cases to that in unclustered cases (14, 15, 17). Such studies typically report the proportions of unique cases. Young age, US birth, Hispanic ethnicity, and risk factors for clustering contrast the odds of an exposure among clustered cases with the odds of that exposure among unique cases. Since the investigators enrolled cases who presented to a specific hospital rather than monitoring all cases of tuberculosis that occurred in a given population, the size of the study population from which the cases were drawn was not known, and the overall incidence of tuberculosis in the Bronx could not be estimated. If some factor, such as HIV, led to an increase in the incidence of both primary disease and reactivation disease, the relative proportion of clustered cases might have remained unchanged and not have reflected the rise in recent transmission. As a measure of disease frequency, the proportions of reactivated and recently transmitted disease may therefore be an inadequate measure with which to estimate the extent of recently transmitted disease or to assess the performance of control measures over time.

### Choice of the “n” versus “n minus one” method

Two different methods for counting the number of clustered cases in a study population have been proposed, the “n” and the “n minus one” methods (26). Alland et al. (13) used the “n” method, which simply sums the number of cases in clusters assuming that each cluster of size “n” contains \( n \) recently transmitted cases.

Formally, we assume that the set of all isolates is comprised of \( n_k \) clusters of size \( k \) for \( k = 1, 2, \ldots, k_{\text{max}} \). When an isolate is unique, \( k \) is equal to one and we refer to the isolate as belonging to a “cluster” of size one but as not being “clustered” to simplify explication. Each isolate can be designated as the \( i \)th subject \( i = 1, \ldots, k \) from the cluster, \( j = 1, \ldots, n_k \) of size \( k \). Then, the number of clustered cases given by the “n” method is

\[
\sum_{k=2}^{k_{\text{max}}} kn_k.
\]

The proportion of clustered cases in the entire study population is this quantity divided by the total number of cases.

\[
\frac{\sum_{k=2}^{k_{\text{max}}} kn_k}{\sum_{k=1}^{k_{\text{max}}} kn_k}.
\]

The “n minus one” method is based on the assumption that one case per cluster was due to reactivation and that this “index” infectious case gave rise to the other cases in the cluster either by infecting them directly or by infecting a secondary case who then infected other members of the cluster. The “n minus one” method calculates the number of recently transmitted cases by summing within clusters after reducing each cluster size by one. The number of clustered cases given by the “n minus one” method is

\[
\sum_{k=2}^{k_{\text{max}}} n_k(k - 1).
\]
Similarly, the proportion of clustered cases is
\[ \frac{\sum_{k=2}^{k_{\text{max}}} n_k (k - 1)}{\sum_{k=1}^{k_{\text{max}}} n_k}. \] (4)

The relative merits of using the “n” versus “n minus one” approach depend on the specific questions being addressed in each particular study. The “n minus one” method identifies the proportions of cases due to reactivation and to primary disease. The estimated proportion of reactivated cases apportions the burden to tuberculosis disease between those distantly and those recently infected. On the other hand, the “n” method compares the number of people involved in transmission chains with the number of people not involved in active transmission chains. Thus, the number of reactivated cases counted by using the “n” method is simply the number of reactivated cases who do not cause active disease in any other people that becomes apparent during the study period. Factors that may determine whether a reactivated case then causes disease include time to diagnosis, clinical type of disease, availability and effectiveness of chemoprophylaxis among contacts, and number of social contacts a reactivated case might have (27–30). Choosing one or the other measure depends on what information one is trying to convey.

**Odds ratios for risk factors for recent transmission**

Alland et al. (13) compared the prevalence of risk factors among clustered and unique cases and measured the relative risk of identifying the exposure in each group. The design of this study resembled that of a case-control study in that study participants were enrolled on the basis of their outcome status. Thus, these measures were presented as odds ratios, where

Odds ratio of risk factor \( (E) = \frac{\text{Probability}(E|\text{clustered disease})}{1 - \text{probability}(E|\text{clustered disease})} \times \frac{\text{Probability}(E|\text{unique disease})}{1 - \text{probability}(E|\text{unique disease})}.

Comparison of the characteristics of cases of recently transmitted disease to cases of reactivated disease does not identify factors that put people at risk of contracting primary tuberculosis disease. Rather, it contrasts the characteristics of those with primary disease to a group of people who have already been infected with tuberculosis and then reactivated. It is not surprising then that a number of studies have found that younger age was a risk factor for recent tuberculosis (13). If all first infections were acquired at a given age, people with primary disease would always be younger than those with reactivated disease, since one cannot have reactivated disease without having previously been infected. Similarly, since HIV infection increases both the risk of primary disease and the risk of reactivation disease (31, 32), comparing these two groups may not reveal the increased risk of primary tuberculosis infection experienced by those with HIV in contrast to those without HIV. In this instance, the unadjusted odds ratio, estimated to be 2.7, almost certainly underestimates the additional risk of tuberculosis for people with HIV.

To serve the public health goal of identifying groups at high risk, a more appropriate measure of association would compare the odds of a risk factor for those with recently transmitted disease with that for a control group that had not become infected with tuberculosis during the same period. Thus, the odds ratio would be

Odds ratio of risk factor \( (E) = \frac{\text{Probability}(E|\text{clustered disease})}{1 - \text{probability}(E|\text{clustered disease})} \times \frac{\text{Probability}(E|\text{no disease})}{1 - \text{probability}(E|\text{no disease})}.

In the absence of such a control group, we can still estimate the prevalence of risk factors in the population that gave rise to the controls. The proportion of clustered cases positive for HIV was 0.67, so the odds of HIV in the exposed group was 2.03. By using an estimate of age-adjusted HIV prevalence of 0.079 (range: 0.034–0.12) (33–35) in the Bronx in 1994 for the hypothetical control group in the previous formula, we obtain an odds ratio of 23.6 (range: 14–400), much higher than the odds ratio of 2.7 calculated by using the unique cases as a control group.

**Bias in estimating the proportions of unique and clustered isolates**

Of the 130 cases Alland et al. (13) recruited over a 3-year period, approximately 20 percent met exclusion criteria; accordingly, their isolates were not studied. If these excluded cases had a DNA fingerprint identical to that of one of the isolates classified as unique, that isolate would have been misclassified. Similarly, if a case included in the Bronx study had recently acquired tuberculosis from a source who had not been diagnosed at Montefiore Hospital, that case may also have been misclassified as unique.

Although small samples usually render an estimate imprecise but not necessarily biased, the situation is different when the outcome is a measure of clustering. Several recent studies have shown that methods used to estimate the proportion of clustered and unique cases in a sample may be biased when the sample does not include all clustered cases in the population from which the sample was drawn (36, 37). Glynn et al. (36) reported a series of simulations of random sampling from previously published series of \( M. tuberculosis \) isolates. They found that estimates of the proportion of clustered cases frequently underestimated the true proportion and that this bias was a function of both the sampling fraction and the underlying cluster distribution.

In molecular studies, the process of data collection can be simulated by sampling some fraction of isolates from the complete set of isolates. The complete data set consists of a...
list of observations describing all cases of *M. tuberculosis* from a closed community over an extended period of time. After each strain type is compared with the others in the complete data set, a cluster can be assigned to each isolate. As above, an isolate can be designated as the *i*th subject *j* = 1, ..., *k* from the cluster, *j* = 1, ..., *n* of size *k*. We assume that each subject in the true set of isolates is sampled independently by using a common sampling fraction *p*.

Let *I*<sub>ijk</sub> be the indicator of whether an isolate has been sampled. Under our assumptions, the *I*<sub>ijk</sub> are independent and identically distributed Bernoulli (*p*) random variables. The total number of subjects sampled is

\[
N = \sum_{k=1}^{\text{max}} \sum_{j=1}^{n_k} I_{ijk}.
\]  

(5)

Therefore, the expected value of the number of isolates is

\[
E(N) = \sum_{k=1}^{\text{max}} n_k p.
\]  

(6)

Let *U*<sub>jk</sub> = one if the number of isolates sampled from the *j*th cluster of size *k* is precisely one and *U*<sub>jk</sub> = zero otherwise. Then, the total number of unique isolates is

\[
U = \sum_{k=1}^{\text{max}} \sum_{j=1}^{n_k} U_{ijk}.
\]  

(7)

Now *U*<sub>jk</sub> is a Bernoulli random variable with success probability *kp*(1 − *p*)<sup>*k−1</sup> equal to the probability of choosing exactly one member from the *j*th cluster of size *k*. Hence,

\[
E(U) = \sum_{k=1}^{\text{max}} n_k p(1 - p)^{k-1}.
\]  

(8)

The approximate expected value of the proportion of unique isolates is then given by

\[
\frac{E(U)}{E(N)} = \frac{\sum_{k=1}^{\text{max}} n_k \sum_{k=2}^{\text{max}} n_k p(1 - p)^{k-1}}{\sum_{k=1}^{\text{max}} n_k \sum_{k=1}^{\text{max}} n_k kp}.
\]  

(9)

Thus, under the assumption of random sampling, the bias in the number of unique isolates is always positive and decreases to zero as the sampling fraction, *p*, approaches one. If sampling is not random, clustering may be overestimated or underestimated depending on whether the sampling strategy is more likely to select cases with primary or reactivation disease. Since most cases are identified through hospital or clinic attendance, the predominant type of case will be determined by the target population of the hospital or clinic in question and by the typical patterns of patient and physician referral within that setting. Although expression 9 allows us to conclude that sampling will lead to overestimation of the fraction of cases due to reactivation, it does not allow us to correct for it directly because the number of clusters of size *k* is not observed in the sampled data.

We can use similar calculations to explore the possible extent of bias in the estimate of clustered cases in the Bronx. To do so, we need to posit a sampling fraction and a distribution of cluster size in the “true,” but unknown set of cases. We can estimate the sampling fraction by approximating the number of tuberculosis cases expected to arise in the hypothetical population from which the Montefiore Hospital cases arose. Assuming that residents of the Bronx mix mostly with each other, we can use the estimated Bronx population of 1,196,500 (38) and the estimated annual tuberculosis incidence in New York City of 30 cases/100,000 in 1993 (39) to estimate that 1,076 cases accrued during the 3-year study period. The hypothetical population cluster distributions were constructed by using a two-step process. When sampled cluster sizes in the New York City data were greater than one, “true” cluster sizes were chosen as the product of the sampled cluster size and the inverse sampling fraction. For sampled cluster sizes of one, the true cluster sizes were sampled from a distribution whose support was the interval 1 to 1/*p* rounded down to the next integer. This distribution was constructed as follows. First, an initial distribution was taken to be the empirical distribution of cluster sizes between 1 and 1/*p* based on the output of the tuberculosis transmission microsimulation model (40); iterative adjustment of the initial distribution was done by hand so the distribution of sampled cluster sizes in the simulated data would more closely approximate the empirical distribution of cluster sizes observed in the New York City data. Note that the “true,” but unknown cluster distribution is constrained but not identified by knowing the sampled distribution and sampling fraction so several different “true” distributions could be hypothesized. By using a computer to randomly sample 104 of the isolates from a hypothetical cluster distribution of 1,076, we can rederive the cluster distribution observed in the Bronx and illustrate the bias that would be observed in the proportion of clustered cases if the proposed hypothetical cluster distribution were true. Table 1 illustrates the “true” and observed estimates of the propor-

<table>
<thead>
<tr>
<th>N&lt;sup&gt;a&lt;/sup&gt;</th>
<th>SF&lt;sup&gt;†&lt;/sup&gt;</th>
<th>Proportion of cases due to reactivation</th>
<th>Odds ratio for HIV&lt;sup‡&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>104</td>
<td>1.00</td>
<td>0.6</td>
<td>0.70</td>
</tr>
<tr>
<td>300</td>
<td>0.35</td>
<td>0.43</td>
<td>0.58</td>
</tr>
<tr>
<td>1,076</td>
<td>0.10</td>
<td>0.35</td>
<td>0.52</td>
</tr>
<tr>
<td>2,000</td>
<td>0.05</td>
<td>0.21</td>
<td>0.47</td>
</tr>
</tbody>
</table>

<sup>a</sup> N, number of cases in the hypothetical complete cluster distribution.

<sup>†</sup> SF, sampling fraction based on (104/N).

<sup‡</sup> Odds ratio for human immunodeficiency virus (HIV) = (odds of HIV[clustered disease])/odds of HIV[unique disease].
tion of unique cases for a hypothetical cluster distribution of 1,076 cases as well as for smaller and larger distributions.

The bias obtained by using the “n minus one” method can be calculated similarly. This method removes one case per cluster from the count of “clustered” cases (groups of two or more isolates) and classifies it as a unique or reactivated “source” case, that is, a reactivated case that gives rise to further cases of tuberculosis. We assume that there is one source case for each cluster greater than size one and that the number of reactivated cases in the “true” population of isolates is equal to the number of unique isolates plus the “source” cases.

After sampling, a cluster present in the true data set can meet one of three fates. It can be counted as unique if exactly one isolate from the true cluster is included in the sampled set. It may not be observed at all if none of the isolates in the cluster is sampled. Finally, it may be counted as a cluster in the sampled data set if more than one of the isolates is sampled. Given a frequency distribution of cluster size, we can estimate the probabilities that a cluster of size \( k \) will meet one of these fates, sum these probabilities across the distribution of cluster sizes, and thereby estimate the bias by using the “n minus one” method given a true cluster distribution and a known sampling fraction (refer to the Appendix). Table 1 compares the estimates of reactivation when the “n minus one” method is used for a range of sampling fractions.

Bias in the odds ratios for risk factors for clustering

In a univariate analysis of the risk factors associated with clustering, Alland et al. (13) found that acquired immuno-deficiency syndrome was not significantly associated with recently transmitted tuberculosis. In addition to the bias that occurs in estimating the proportion of cases due to recent transmission, risk factors for recent transmission may be underestimated when sampling is incomplete. This bias results from the misclassification of clustered cases as unique and the subsequent bias toward the null hypothesis of no effect of the exposure. Let \( N \) be the number of cases in the complete sample and \( N^* \) be the number of sampled cases. Similarly, let \( U \) and \( CL \) be the number of unique and clustered cases, respectively, and \( U^* \) and \( CL^* \) be the number of cases classified as unique and clustered, respectively, in the sample. We designate subscripts 0 and 1 to refer to unexposed and exposed cases, respectively, so that \( U_1^* \), for example, refers to an exposed unclustered case in the sampled data set. Let \( r_U \) and \( r_{CL} \) be the true prevalences of an exposure in the unique cases and the clustered cases, respectively, and \( r_U^* \) and \( r_{CL}^* \) be the observed prevalences in the unique cases and the clustered cases, respectively. Since a portion of the observed unique cases is truly clustered but misclassified as unique, the cases classified as exposed uniques will consist of the truly exposed, \( U_1 / r_U \), and the exposed clustered cases misclassified as unique. Because the expected number of truly clustered cases in the sample is \( CL(N^*/N) \), the approximate number of exposed cases misclassified as unique is \( (CL(N^*/N) - CL^*)r_{CL} \). Once the extent of bias in estimating the number of clustered cases has been calculated, the impact of this misclassification on the observed odds ratios can be estimated by reestimating the odds ratios when misclassified cases are moved from the category of unique to that of clustered cases (41). By using this method, we can calculate the “true” prevalences of exposure in the unique and clustered cases that would have given the odds ratios observed in the Bronx study (13) had the isolates in that study been sampled from the hypothetical cluster distributions constructed above. Table 1 shows that these odds ratios can be much more extreme than those found in the sample, even when the comparison group consists of the reactivated cases.

Discussion

Current study design and analysis of the molecular epidemiology of tuberculosis do not consistently yield interpretable and comparable results, especially when small sampling fractions have been used. Previous reviews have suggested ways in which study design may be improved to facilitate comparability (26). In addition to these recommendations, the following four points should be considered in the design and analysis of these studies:

1. The use of numbers, rather than proportions, of unique and clustered cases of tuberculosis allows estimation of the incidence of recent transmission if the base population from which the cases are recruited is known. When this population cannot be identified, for example, with convenience samples, measuring the proportion of clustered and unique cases may not yield interpretable indicators of the burden of tuberculosis.

2. The choice of the “n” versus “n minus one” method should be made on the basis of the specific epidemiologic question being addressed. The “n” method is appropriate if the investigators want to estimate the number of people involved in transmission chains, while the “n minus one” method should be used if the investigators want to ascertain the number of people with primary versus reactivated disease. Although this choice depends on the information one is trying to convey, note that both methods will yield biased results after sampling. Reporting the empirical distribution of cluster sizes would facilitate interpretation of these data.

3. Many current studies of risk factors for recent transmission compare the probability of exposure in clustered cases to the probability of exposure in unique cases. Therefore, they do not identify factors that put uninfected people at risk of primary disease. This question can be addressed by using an appropriate comparison series to determine the distribution of the exposure in the population that gave rise to the cases.

4. Sampling of a subset of cases in an epidemic followed by naive methods of analysis will lead to underestimation of the number and proportion of clustered cases. This bias may be extreme when very small sampling fractions are used, as is the case in “convenience” sampling and when clusters tend to be small. Sensitivity analyses can be performed to explore the
extent of bias that would occur if the observed sample had been drawn from a particular hypothetical cluster distribution. When these analyses show that there is the potential for significant bias, the incidence of clustered cases should be reported as a lower bound and the extent of potential bias made explicit.

Adoption of more rigorous reporting standards in studies of the molecular epidemiology of tuberculosis would improve the comparability of studies and help investigators assess the implications of their results. Given the widespread use of molecular tools in epidemiologic studies of tuberculosis and the tremendous need for a better understanding of the epidemiology of tuberculosis, further methodological advances in these areas are badly needed to make the most of the technical resources now available. Rather than relegate molecular typing in tuberculosis to the status of a tool without a research question, we need to find ways to enlist this tool to answer questions of major public health importance.

ACKNOWLEDGMENTS

Dr. Megan Murray was supported by National Institutes of Health grant k08 AI-01430-01. The authors are indebted to Dr. James Robins for his input into the analytical solutions presented in this paper. In addition, they are grateful for helpful suggestions on the manuscript from Sam Bozeman and from Drs. Marc Lipstitch, Barry Bloom, and Jean Marie Arduino.

REFERENCES

We want to find the bias in the number of “source” cases when there is one source case for each cluster greater than size 1. Let $CL_{1jk} = 1$ if the number of isolates sampled from the $j$th cluster of size $k$ is precisely 1 and $CL_{1jk} = 0$ otherwise. Similarly, let $CL_{0jk} = 1$ if the number of isolates sampled from the $j$th cluster of size $k$ is precisely zero and $CL_{0jk} = 0$ otherwise. Finally, let $CL > 1jk = 1$ if the number of isolates sampled from the $j$th cluster of size $k$ is greater than 1 and $CL > 1jk = 0$ otherwise. Let $CL(1) = \sum_{k=2}^{k_{\text{max}}} \sum_{j=1}^{n_k} CL_{1jk}$ and $CL(0) = \sum_{k=2}^{k_{\text{max}}} \sum_{j=1}^{n_k} CL_{0jk}$. Then, the expected values of $CL(1)$ and $CL(0)$ are given by $E(CL(1)) = \sum_{k=2}^{k_{\text{max}}} \sum_{j=1}^{n_k} n_k p(1 - p)^{k-1}$ and $E(CL(0)) = \sum_{k=2}^{k_{\text{max}}} n_k (1 - p)^k$. Hence, the expected bias in the estimated proportion of source cases after sampling is

$$E(CL(1)) + E(CL(0))E(N) = \frac{\sum_{k=2}^{k_{\text{max}}} \sum_{j=1}^{n_k} pkn_k}{\sum_{k=1}^{k_{\text{max}}} pkn_k}.$$  

The overall additive bias for the proportion of reactivated cases estimated by using the “$n$ minus one” method is equal to the sum of bias in the proportion of uniques obtained by using the “$n$” method and the bias in the proportion of source cases described above.