**Estimating recent tuberculosis transmission from tuberculosis cluster distributions**

People (so far): me - Ellen Brooks-Pollock, Maeve Lalor, Jennifer Davidson, Hester Korthals Altes, Andy Pollock, Leon Danon.

**Abstract**

**Background**

In many low incidence countries, a high proportion of tuberculosis (TB) cases are born abroad. However it is not known what proportion of foreign-born persons were exposed abroad, or since moving to the low incidence country. We aimed to estimate the proportion of TB cases in two low incidence countries that were due to recent infection.

**Methods**

We developed a mathematical model of TB cluster generation involving recent transmission and importation of infection. We calibrated the model to available MIRU-VNTR data for TB cases in England and the Netherlands. We used the model to estimate the average reproduction number, the importance of super-spreaders and the proportion of cases that were due to recent transmission in each setting.

**Results**

For both England and the Netherlands, we found that the number of secondary cases per TB case was captured by a Poisson-lognormal distribution, in which 15% (8%, 27%) of cases were responsible for the majority of recent infections. The within-country reproduction numbers in England and the Netherlands were 0.41 (0.30, 0.60) and 0.24 (0.14, 0.48) respectively. The percentage of imported cases was estimated at 73% (64%, 79%) in England, compared to 81% (69%, 89%) in the Netherlands.

**Conclusions**

Despite many differences between the TB epidemics in England and the Netherlands, we have developed a general model to capture TB cluster generation. Our analysis suggests that transmission in England and the Netherlands is similar; the major difference is in the arrival of new cases. Our parameter inference is based on VNTR data; future whole genome sequencing data can be used to re-estimate essential transmission parameters.

**Introduction**

Tuberculosis (TB) is a chronic infectious disease and a major global public health threat. In 2015, more than 8 million people were newly diagnosed with TB and nearly 2 million people died from TB worldwide. In many low incidence countries, a high proportion of cases occur in persons born abroad, and control measures such as migrant screening have been introduced to reduce cross-border transmission [ref Rob Aldridge]. However, it is often not known if foreign-born were exposed to TB before or after arriving in the low incidence country, which limits the potential impact of such interventions [ref Rob Aldridge].

Over the past 20 years, genotyping has informed our knowledge of how TB evolved, spread around the world and survives within hosts. However, unlike for some other infections, genotyping TB strains cannot definitively identify who-infected-whom. This is because associated cases are often infected with genetically identical strains, masking the direction of transmission. Instead, DNA fingerprinting is often used to rule out transmission, for instance between household members infected with different strains2. In low incidence countries, non-matching strains are used to estimate the fraction of cases that are notdue to recent transmission3. However, as matching strains do not necessarily indicate recent transmission1, alternative measures of recent transmission are required.

TB clusters are defined as multiple cases infected with an identical genotype. Large clusters are assumed to signify sustained recent transmission, however their significance depends on the epidemiological setting. Ypma et al. analysed the distribution of TB transmission cluster sizes in the Netherlands: they found that the highly skewed distribution of cluster sizes was indicative of super-spreading individuals, i.e. individuals that generate many more secondary cases than average.

In this paper, we develop the method of Ypma et al. to be able to estimate the amount of recent transmission from TB cluster data from England. In the process of developing the model,

**Methods**

*Data extraction and processing*

*GB data*

The analysis was conducted using the Enhanced Tuberculosis Surveillance (ETS) system. We extracted year (2010-2015), country of residence (England, Scotland, Wales, Northern Ireland), Public Health England centre of residence (England only: East Midlands, East of England, London, North East, North West, South East, South West, West Midlands, Yorkshire and the Humber), country of birth, disease type (pulmonary or extra-pulmonary), strain type (24 loci mycobacterial interspersed repetitive unit-variable-number tandem repeat (MIRU-VNTR) type), cluster name (assigned by PHE naming tool based on strain type) and whether a case was categorised as clustered (yes/no) for all TB cases diagnosed between 2010 and 2015.

Of the 47,161 cases diagnosed between 2010 and 2015, 23,646 cases were associated with a 23+ loci VNTR profile that could be used for cluster identification; these cases were extracted for further analysis. Cluster size was defined as the number of cases with an identical VNTR profile, where clusters of size 1 were cases with a unique 24 loci VNTR profile. Cases with a single missing locus that matched 23 loci of another cluster were considered part of that cluster1.

Cluster size was binned logarithmically to retain the distribution shape while minimising noise due to low numbers of large clusters2. The association between cluster size and the proportion of pulmonary cases and cluster size and country of birth was characterised in the process of model development.

*NL data*

In order to compare the GB model with the original analysis, we extracted obtained data on TB cases with a VNTR type from the Netherlands Tuberculosis Register. VNTR typing has been systematically conducted in the Netherlands since 2004. As for the GB data, we extracted year of diagnosis (2004-2015), province of residence, country of birth, disease type (pulmonary or extra-pulmonary), strain type (24 loci mycobacterial interspersed repetitive unit-variable-number tandem repeat (MIRU-VNTR) type). 8,448 cases diagnosed between 2004 and 2015 were associated with a 23+ loci VNTR profile that could be used for cluster identification. When directly comparing with Great Britain we used Netherlands data from 2010 to 2015.

*Models for cluster size distribution*

We developed a mortal branching process model3,4 with importation of infection to describe the process by which TB clusters are generated in the UK. The central premise behind the model is that each case diagnosed must either have acquired infection abroad or before the observation period (imported infection) or have been infected in the UK (transmitted infection), in a similar structure to household transmission models5.

The first case in a cluster is assumed to have been imported as there are no other cases with matching VNTR profile. We assume that while a case is infectious it generates  secondary cases, where  is drawn from a probability distribution , and that  further cases infected with an identical VNTR profile are imported, where  is the result of a Bernoulli process. Each of the generated and imported cases could generate secondary cases; this process is repeated until no new cases are created. The total size of a cluster is the sum of all the cases:



We assume that clusters are finite so that so that the average number of secondary cases per case is less than one.

*Distribution of secondary cases per individual*

However for many diseases, it has been shown that a small number of individuals generate a high number of secondary cases4,6,7. Previously, the number of secondary cases per TB case has been described using a negative binomial distribution, which arises when the expected number of secondary cases per individual, , follows a Gamma distribution with dispersion parameter  and scaling parameter ,  and the average number of secondary cases per individual is given by . The distribution of TB cluster sizes in the Netherlands was shown to be well described by a negative binomial distribution with dispersion parameter  and 4. The small dispersion parameter illustrated the importance of super-spreaders in the Netherlands.

However, initial analysis revealed that a negative binomial model was not able to capture the observed distribution of cluster sizes: the best-fit model underestimated the occurrence of the large clusters and the frequency of unmatched cases.

We found that a Poisson-lognormal distribution was able to capture the entire distribution of TB cluster sizes. A Poisson-lognormal distribution is frequently used in ecological literature as an alternative to a negative binomial to describe species abundances for communities with many rare species. It arises when the logarithm of the expected number of secondary cases per individual, , follows a normal distribution with mean  and variance , . In a lognormal distribution, the average number of secondary cases per individual is given by .

*Fraction of imported cases*

During the time that an individual is infected, they generate an average  secondary cases and another case is imported with probability. Therefore, the fraction of imported cases not infected in Great Britain is given by As imported cases can generate secondary cases, eliminating importation would reduce the number of TB cases by more than the fraction of cases imported. To estimate the potential reduction in TB cases with reduced imported infection, we simulated the model with and without importation (), assuming that the initial case still occurred. We calculated the reduction in cases with and without importation.

*Model fitting*

In contrast to previous approaches3,4 that have used exact likelihood methods for fitting cluster size models to data, we used *Approximate Bayesian Computation (ABC)*7,8. In ABC, the likelihood is approximated by distance metrics based on summary statistics derived from the data and a realisation of the model, therefore can naturally incorporate the impact of sampling and importation. We used the Majoram MCMC search algorithm implemented in the R package EasyABC9 with the distribution of TB cluster sizes as the summary statistic.

From the posterior distributions, we extracted the average number of secondary cases per individual (), the degree of dispersion measured as the percentage of transmissions due to 20% of cases, and the proportion of cases that are due to recent local transmission.

The models and fitting procedure leading to the results presented here are implemented in R and available at http://github.com/ldanon/TBclustr.

**Results**

Between 2010 and 2015, there were 12,503 clusters in Great Britain: 9,802 cases were due to a unique strain and 13,844 cases were clustered. Median cluster size was 7 cases, three clusters contained over 200 cases and six clusters contained over 100 cases. In the Netherlands between 2004 and 2015, there were 4,926 clusters: 3,905 cases were unmatched and 1,021 cases were clustered. The median cluster size was 1, with three clusters over 100 cases.

*Cluster size and fraction of pulmonary cases*

The average percentage of cases in cluster with pulmonary disease was similar for Great Britain and the Netherlands. In Great Britain, on average 70% (range 42%, 98%) of cases in a cluster had pulmonary disease. There was minimal association between cluster size and the percentage of pulmonary cases in a cluster (). In the Netherlands, 70% (range 26%, 100%) of cases in a cluster had pulmonary disease ().

*Cluster size and fraction of foreign-born cases*

The relationship between cluster size and foreign-born cases was again remarkably similar in Great Britain and the Netherlands. In Great Britain, 36% (range 5%, 90%) of cases in a cluster were UK born. There was no consistent association between cluster size and the percentage of UK born cases (p-value=0.2) in a cluster (figure 1). In the Netherlands, 34% (range 0%, 67%) of cases were born in the Netherlands. Similarly, there was no relationship between cluster size and Netherlands-born cases.

*The distribution of cluster size model*

As discussed in the methods, a negative binomial model was not able to capture the observed distribution of TB cluster sizes. Instead, both cluster size distributions from Great Britain and the Netherlands were captured by a Poisson-lognormal model. The Poisson-lognormal distribution for the number of secondary cases in Great Britain had log-mean of -2.9 (95%CI -4.7, -1.5) and log-variance 2.0 (95%CI 1.2, 2.8). The estimated reproduction number is 0.41 (95%CI 0.30, 0.60).

In the Netherlands, between 2004 and 2015 the log-mean was -2.9 (95%CI -5.0, -1.6) and log-variance 1.9 (95%CI 1.0, 2.8). The estimated reproduction number over that time was 0.33 (95%CI 0.22, 0.50). Restricting the analysis to cases reported in the Netherlands between 2010 and 2015, decreased the log-mean to -3.4 (95%CI -6.7, -1.7) and slightly increased the log-variance 1.9 (95%CI 1.0, 3.4). The estimated reproduction number since 2010 was 0.25 (95%CI 0.14, 0.48).

The probability of a case being imported was 0.11 (95%CI 0.009, 0.25), leading to an estimated 21% (95%CI 2%, 41%) of cases that were imported and not infected in Great Britain. Simulating the model with and without imported infection after the initial introduction, we estimate that 34% (95%CI 0%, 61%) of cases would be eliminated without importation of infection.

*Impact of right censoring*

*Impact of number of years of data*

**Discussion**

How does this relate to distribution of social contacts? No artificial bounds on contact distribution.

What is causing the distribution? Infectious period? Infectiousness? Number of contacts?

Over time, as DNA techniques have improved, the definition of an identical strain has changed and consequently so has the definition of a cluster with more discriminatory power leading to fewer clustered cases [ref Jerker Jonsson]. Many countries have now replaced Restriction Fragment Length Polymorphism (RFLP) typing with Mycobacterial Interspersed Repetitive Units-Variable Numbers of Tandem Repeat (MIRU-VNTR), and in England MIRU-VNTR is being replaced again with whole genome sequencing (WGS).

small clusters that appear in the later years of data may only be small because they have been recorded at an earlier stage of their growth, not because the transmission potential has changed.

species abundance distribution describes the full distribution of commonness and rarity in ecological systmes.

Gamma and lognormal models are both used to model species abundance.

lognormal – communities with many rare species

gamma used for birds and fish, “centred” distribution

insect communities = both lognormal and gamma

plant communities – lognormal or unstable or artificial environments