A Sign of Superspreading in Tuberculosis: Highly Skewed Distribution of Genotypic Cluster Sizes

Rolf J.F. Ypma, Hester Korthals Altes, Dick van Soolingen, Jacco Wallinga, W. Marijn van Ballegooijen

1 Likelihood derivation

1.1 Final size distribution

We consider a branching process with a certain distribution for the number of offspring $X$. It is well-known that this process dies out with probability 1 if $E(X) < 1$. Let $\Pi(s) = \sum_{x=0}^{\infty} f(x)s^x$ be the probability generating function of $X$, we can then write $\Pi^y(s) = (\sum_{x=0}^{\infty} f(x)s^x)^y = \sum_{x=0}^{\infty} g(x)s^x$ for some function $g(x)$. Let $Y$ denote the total number of individuals generated by the branching process, including the starting individuals. Becker [1] showed that the probability that $Y = y$, when the process started with $n$ individuals, is given by the coefficient of $s^y$ in $\Pi^y(s)$ multiplied by $\frac{n}{y}$:

$$P(Y = y | n) = \frac{n}{y} g(y - n)$$

This result is useful when $\Pi(s)$ has a mathematically tangible form. This is the case for the so-called power series distribution, given by

$$P(X = x) = \frac{a(x) \theta^x}{A(\theta)}$$

with $\sum a(x) \theta^x = A(\theta)$; here $\Pi(s) = \frac{A(\theta s)}{A(\theta)}$. The negative binomial distribution has two parameters, $p$ and $k$, and is a special case of this distribution; $\theta = p$, $a_k(x) = \binom{x+k-1}{k-1}$ and $A_k(\theta) = (1 - p)^{-k}$. Let $\Pi_{NB}(s)$ denote its generating function, we then have

$$\Pi_{NB}^y(s) = \left( \frac{A_k(sp)}{A_k(p)} \right)^y = \frac{(1 - p)^{ky}}{(1 - sp)^{ky}} = (1 - p)^{ky} A_k(sp) = (1 - p)^{ky} \sum_{x \geq 0} \binom{x + yk - 1}{yk - 1} p^x s^x$$

where for the last equality we used $A_k(sp) = \sum a_k(x)(sp)^x$. Taking the coefficient of $s^{y-n}$ and multiplying by $\frac{n}{y}$ gives, for $y \geq n$,

$$P(Y = y | n) = \frac{n}{y} \left( \frac{y - n + yk - 1}{yk - 1} \right) p^{y-n} (1 - p)^{ky}$$

$$= \frac{n}{y} \left( \frac{y + yk - n - 1}{yk - 1} \right) p^{y-n} R^{ky} (1 + R) (k+1) y - n$$

with $R = E(X) = \frac{kp}{1-p}$. Obviously, $P(Y = y | n) = 0$ for $y < n$. Note that this distribution is only proper if $R \leq 1$. However it still holds for larger $R$ if we define $P(Y = \infty | n) = 1 - \sum_{y \in \mathbb{N}} P(Y = y | n)$ [2].
1.2 Mutation

It is possible that mutation (e.g. a transposition) occurs during infection in the Netherlands, resulting in offspring belonging to a different genotypic cluster. We model this by saying that for each infection event a mutation occurs with probability $p_m$. Let $X_{cl}$ denote the number of offspring of a certain individual having the same DNA fingerprint as that individual. Since each of the offspring has a different type with probability $p_m$, having $x$ offspring of the same type means either having $x$ offspring and no mutations, or $x+1$ offspring and one mutation, or $x+2$ offspring and two mutations, etc. This is the convolution of a negative binomial distribution with a binomial distribution

$$P(X_{cl} = x) = \sum_{i \geq x} P(X = i) \binom{i}{x} p_m^{i-x}(1-p_m)^x$$

which is again a negative binomial distribution with parameters $R(1-p_m)$ and $k$ (Lemma 1).

Thus the final size of each of the clusters is described by equation (1), with $R$ replaced by $R(1-p_m)$. This means it will be impossible to estimate $R$ and $p_m$ separately from the dataset. Note that, although the value of the dispersion parameter $k$ does not change, the value of the variance-to-mean ratio $1 + \frac{R}{k}$ does.

**Lemma 1.** The convolution of a negative binomial distribution with a binomial distribution is again a negative binomial distribution.

If we write $A^{(x)}(\theta)$ for the $x$-th derivative of $A(\theta)$, we have that for the negative binomial distribution

$$P(X_{cl} = x) = \sum_{i \geq x} P(X = i) \binom{i}{x} p_m^{i-x}(1-p_m)^x$$

$$= \sum_{i=x}^{\infty} \binom{i+k-1}{k-1} \left(\frac{k}{k+R}\right)^k \left(\frac{R}{R+k}\right)^i \binom{i}{x} p_m^{i-x}(1-p_m)^x$$

$$= \sum_{i=x}^{\infty} \binom{i+k-1}{k-1} \left(1-p\right)^k p_i \binom{i}{x} p_m^{i-x}(1-p_m)^x$$

$$= (1-p_m)^x p^x \sum_{i=x}^{\infty} \binom{i+k-1}{k-1} \left(\frac{p_m p}{i-x}\right)!$$

$$= (1-p_m)^x p^x \frac{(1-p)^k}{x!} A^{(x)}(p_m)$$

$$= (1-p_m)^x p^x \frac{(1-p)^k}{x!} \frac{\Gamma(x+k-1)}{\Gamma(k-1)} (1-p_m)^{k-x}$$

$$= \frac{\binom{x+k-1}{k-1} \frac{1-p}{1-p_m} \Gamma(k-1)}{\Gamma(x+k-1)} (1-p_m)^{k-x}$$

$$= \left(\frac{x+k-1}{k-1}\right) \frac{1-p}{1-p_m} \Gamma(k-1) (1-p_m)^{k-x}$$

$$= \left(\frac{x+k-1}{k-1}\right) \frac{1-p}{1-p_m} \Gamma(k-1) (1-p_m)^{k-x}$$

which is just the negative binomial distribution again, but now with mean $E(X_{cl}) = R(1-p_m)$.

1.3 Dealing with unobserved cases

We assume each individual which is not an index case is seen with probability $p_{seen}$. The probability distribution for the observed cluster size $Y_{obs}$ given $n$ index cases becomes
for $y > 0$, since clusters of size 0 are unobserved. It would be more elegant to allow for unobserved index cases as well, however this would need additional assumptions on the distribution of index cases over clusters. This is not straightforward, particularly because these additional index cases in our study are immigrants, typically from a few countries with high prevalence of tuberculosis; their DNA fingerprints are highly correlated.

1.4 Full likelihood

The likelihood of parameters $R, k, p_m$ when $a_{y,n}$ clusters of size $y$ with $n$ index cases have been fully observed, and $b_{y,n}$ censored clusters have been observed to have at least size $y$ and $n$ index cases is

$$L(R, k, p_m | a, b) = \prod_y \prod_n \left( \frac{1}{\sum_{i=y}^{\infty} P(Y_{cl} = i | n)(1 - p_{seen})^i} \sum_{i=y}^{\infty} P(Y_{cl} = i | n) \left( \begin{array}{c} i \\ y \end{array} \right) p_{seen}^y (1 - p_{seen})^{i-y} \right)^{a_{y,n}}$$

and

$$P(Y_{cl} = i | n) = \begin{cases} \frac{n}{i} \left( \begin{array}{c} i + k - n - 1 \\ ik - 1 \end{array} \right) \frac{(R(1 - p_m)^{i-n}k^k)}{(k + R(1 - p_m))^{k+1}n-n} & \text{if } n \leq i \\ 0 & \text{if } n > i \end{cases}$$

2 Sensitivity analysis

To construct our model we made several assumptions regarding TB epidemiology, here we will explore the sensitivity of our results to these assumptions.

2.1 Recently arrived immigrants

In our model we assumed that the 933 immigrants that had been in the Netherlands for less than six months at date of diagnosis were infected elsewhere. We consider three alternatives to this assumption, and estimate the parameters $R_m$ and $k$ under these alternative assumptions.

First, we assume all recently arrived immigrants were infected in the Netherlands, and treat them the same as all other cases in our dataset. We would expect that this assumption slightly increases our estimate of $R_m$, as more infections are needed to explain all cases. Indeed, the assumption leads to estimates of $R_m = 0.5, k = 0.11$, where our original estimates were $R_m = 0.48, k = 0.10$. The slightly larger value of $k$ is probably due to an increase in the number of 'average sized' clusters (i.e. size 2,3,4,...), which point at moderate values of $k$. The increase in $k$ however, is minor.
Second, we exclude all transmission clusters that contain a recently arrived immigrant from our dataset. As larger clusters are more likely to contain such an immigrant, we expect to find a lower value of $R_m$. Indeed, this assumption leads to estimates of $R_m = 0.39, k = 0.10$. Note that the estimated value for $k$ is identical to our original estimate.

Third, again assuming that the recently arrived immigrants were infected abroad, we include any additional cases abroad that never immigrated to the Netherlands by regarding all clusters containing a recently arrived immigrant as censored. Our interpretation of the estimated parameters would then change slightly, as infections abroad are now also included. With additional censoring, we find estimates of $R_m = 0.61, k = 0.09$.

Summarizing, we find that different assumptions regarding the role of recently arrived immigrants lead to different estimations of $R_m$, but nearly identical estimates of $k$.

2.2 Extra pulmonary tuberculosis cases

In our model we assumed extra pulmonary tuberculosis (EPTB) cases were non-infectious, and discarded them from the analysis. Thus the parameters estimated in the main text describe the number of pulmonary cases caused by one pulmonary case. It is possible that some infections have been caused by (possibly misdiagnosed) EPTB cases. We consider the extreme scenario in which EPTB cases are as infectious as pulmonary TB cases.

We added the 3892 EPTB cases to our dataset and analysed the full dataset of 12222 cases. We found 8221 clusters, 7070 (86%) of which consisted of only one case. We find estimates of the parameters of $R_m = 0.50, k = 0.094$. The results appears to be insensitive to the assumption regarding EPTB cases.

2.3 Homoplasy/recurrent mutation

In our model we assumed that all RFLP types generated through a mutation are unique, i.e. lead to new clusters. However, it is possible that several mutation events lead to the same RFLP type, either through homoplasy or recurrent mutations. The net result of such mutations would be one large genotypic cluster, rather than several small clusters. Thus any given case might not have been infected within its genotypic cluster but could be the result of transmission from a different cluster and a recurrent mutation; it is then an additional index case for its genotypic cluster.

Each cluster of size $y$ with $n_{\text{obs}}$ observed index cases contains $y - n_{\text{obs}}$ cases that could actually be an index case, due to the mechanism described above. Assume this happens for each case with probability $p_r$. Then the actual number of index cases $n$ for a cluster of size $y$ with $n_{\text{obs}}$ observed index cases is distributed as $n_{\text{obs}} + n_r$, with $n_r \sim \text{Bin}(y - n_{\text{obs}}, p_r)$.

The likelihood equation in section (1.4) that we use to obtain estimates is based on $\{a_{y,n}\}$ and $\{b_{y,n}\}$, the number of clusters completely/partially observed having precisely/at least $y$ cases and $n$ index cases. Taking recurrent mutations into account, we note that some of the clusters will in reality have more index cases than observed. We define $A_{y,n}$ as the expected number of clusters of size $y$ and $n$ index cases, given the observed clusters. We get

$$ A_{y,n} = \begin{cases} 
0 & \text{if } y < n \\
\sum_{n_{\text{obs}}=1}^{n-1} a_{y,n_{\text{obs}}} \binom{y - n_{\text{obs}}}{n - n_{\text{obs}}} p_r^{n - n_{\text{obs}}} (1 - p_r)^{y - n} & \text{if } y = n \\
\sum_{n_{\text{obs}}=1}^{n} a_{y,n_{\text{obs}}} \binom{y - n_{\text{obs}}}{n - n_{\text{obs}}} p_r^{n - n_{\text{obs}}} (1 - p_r)^{y - n} & \text{if } y > n
\end{cases} $$
and equivalently we define $B_{y,n}$:

$$B_{y,n} = \begin{cases} 
0 & \text{if } y < n \\
 b_{y,n} + \sum_{i=1}^{n-1} b_{y,n_{obs}} \left( \frac{y - n_{obs}}{n - n_{obs}} \right) p_r^{n-n_{obs}} (1 - p_r)^{y-n} & \text{if } y = n \\
 \sum_{i=1}^{n} b_{y,n_{obs}} \left( \frac{y - n_{obs}}{n - n_{obs}} \right) p_r^{n-n_{obs}} (1 - p_r)^{y-n} & \text{if } y > n 
\end{cases}$$

We then replace $a_{y,n}$ and $b_{y,n}$ in the likelihood equations of section (1.4) by $A_{y,n}$ and $B_{y,n}$ to obtain estimates for $R_m$ and $k$. For $p_r \in [0, 0.5]$ the adjusted likelihood equation yields estimates of $R_m$ in $[0.40, 0.48]$ and $k$ in $[0.10, 0.052]$ (figure S1). This means the method is insensitive to recurrent mutations.

**References**


Figure S1. Estimates of the dispersion parameter $k$ under different values for $p_r$, the probability that a non-index case in an genotypic cluster was infected by a case with a different RFLP type. In the main text $p_r = 0$, yielding an estimate of $k = 0.10$. For higher values of $p_r$, i.e. higher rates of recurrent mutations, the estimate of $k$ decreases. This shows the initial estimate of $k$ was conservative.