HHS Public Access

Author manuscript

J Theor Biol. Author manuscript; available in PMC 2020 February 11.

Published in final edited form as:

J Theor Biol. 2015 March 07; 368: 67–73. doi:10.1016/j.jtbi.2014.12.009.

Magnitude and sources of bias in the detection of mixed strain *M. tuberculosis* infection

Giacomo Plazzotta[a,*], **Ted Cohen**[b], **Caroline Colijn**[a]

[a]Imperial College London, United Kingdom

[b]Brigham and Women's Hospital, Harvard School of Public Health, United States

Abstract

High resolution tests for genetic variation reveal that individuals may simultaneously host more than one distinct strain of *Mycobacterium tuberculosis*. Previous studies find that this phenomenon, which we will refer to as "mixed infection", may affect the outcomes of treatment for infected individuals and may influence the impact of population-level interventions against tuberculosis. In areas where the incidence of TB is high, mixed infections have been found in nearly 20% of patients; these studies may underestimate the actual prevalence of mixed infection given that tests may not be sufficiently sensitive for detecting minority strains. Specific reasons for failing to detect mixed infections would include low initial numbers of minority strain cells in sputum, stochastic growth in culture and the physical division of initial samples into parts (typically only one of which is genotyped). In this paper, we develop a mathematical framework that models the study designs aimed to detect mixed infections. Using both a deterministic and a stochastic approach, we obtain posterior estimates of the prevalence of mixed infection. We find that the posterior estimate of the prevalence of mixed infection may be substantially higher than the fraction of cases in which it is detected. We characterize this bias in terms of the sensitivity of the genotyping method and the relative growth rates and initial population sizes of the different strains collected in sputum.

Keywords

Study designs aimed to detect mixed infection; Majority and minority strain; Prevalence of mixed infection; Posterior distribution of the prevalence of mixed infection; Estimates of the prevalence of mixed infection

1. Introduction

Tools for the genetic analysis of *Mycobacterium tuberculosis*, the causative agent of human tuberculosis (TB), have fundamentally altered our understanding of the natural history of this pathogen. The ability to distinguish isolates has shown that individuals can be re-

*Corresponding author. giacomo.plazzotta11@imperial.ac.uk (G. Plazzotta).

Appendix A. Supplementary data

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2014.12.009.

infected with *M. tuberculosis*, and this poses clear challenges for vaccine development since even natural infection at best provides partial immunity. Furthermore, the advent of high resolution tests for genetic variation has revealed that individuals may *simultaneously* harbor infections with more than one distinct strain of *M. tuberculosis* (Warren et al., 1999; Sola et al., 2003; Kremer et al., 1999; van Embden et al., 1993; Imaeda, 1985). This phenomenon, which we will refer to as "mixed infection", has been linked with poor treatment outcome when the co-infecting strains differ with respect to drug susceptibility (van Rie et al., 2005; Hingley-Wilson et al., 2013) and is predicted to influence the impact of population-level interventions against tuberculosis (Cohen et al., 2008; Rodrigues et al., 2007; Colijn et al., 2009; Sergeev et al., 2011; Mills et al., 2013).

Accurate estimates of the frequency with which mixed infections occur are therefore critical to understand how mixed infections impact both the natural history and the dynamics and control of this infection. However, the detection of mixed infections is challenging (Hingley-Wilson et al., 2013), even with tools that have high sensitivity for detecting minority strains and adequate resolution to discriminate between closely related (but genetically distinct) pathogens. As discussed in detail in a recent review (Cohen et al., 2012), there are many opportunities to fail to detect a mixed infection that is actually present in a host, because a minority strain might not be harvested in the collected clinical specimen, might be lost during the process of specimen transport and handling, and might fail to be detected by the particular genotyping method employed.

Despite these clear opportunities to miss the detection of true mixed infections, among the several dozen studies available, it has been found that mixed infections were often detected in as many as 10–20% (Cohen et al., 2012; Hanekom et al., 2013; Huang et al., 2010; Navarro et al., 2011) of cases in areas where the incidence of TB is high (Cohen et al., 2012). Since we believe that this statistic may underestimate the prevalence of mixed infections, we have developed a mathematical model to understand the potential sources of bias in estimates of the prevalence of mixed infection and to provide bounds for reasonable uncertainty as to the actual prevalence of mixed infections given the observed prevalence and knowledge of the laboratory protocol employed to detect mixed infections.

2. Methods

Although the designs of previous studies for detecting mixed strains have differed in important ways (Cohen et al., 2012), for the purposes of this analysis, we have generalized the study design to include several steps common to nearly all of these investigations:

1. Specimen collection from the patient (samples of 0.25 mL).
2. Specimen growth in culture.
3. Sampling of bacterial isolates from culture and extraction of mycobacterial DNA.
4. Analysis of mycobacterial DNA (see Fig. 1).

Here we focus our analysis on bias that might arise in the detection of mixed infections related to steps 2–4 above. That is, we do not consider the bias that might result from failing

to collect a minority strain from an individual, and instead we focus here on the bias that arises from failing to detect a minority strain after it has actually been collected from a patient. This is not meant to indicate that we think failing to collect a minority strain from a patient does not contribute to the underestimation of the prevalence of mixed infection. Rather, this approach allows us to provide estimates on the bias that is associated with the laboratory procedures that are distal to specimen collection. We comment further on this issue in the discussion. The probability that mixed infection is detected can be decomposed as

$$\mathbb{P}(\text{detect}) = \mathbb{P}(\text{detect} \mid \text{mixed infection present})\mathbb{P}(\text{mixed infection present}). \quad (1)$$

We define the prevalence of mixed infection to be the fraction of individuals with TB disease that are simultaneously infected by more than one distinct strain. Here we define strains by their ability to be discriminated from each other by the particular genotyping test used. Our aim is to estimate the prevalence of mixed TB infection in a population, $\rho := \mathbb{P}$(mixed infection present), from a set of data consisting of measurements aiming to detect mixed infection in individuals. To do this, we characterize $m := \mathbb{P}$(detection|mixed infection present) by modelling laboratory handling and subsequent growth of bacilli in culture. We use a stochastic model of specimen handling and growth where cell numbers are small, and a deterministic model otherwise. The model inputs are the distributions of the numbers of cells in the samples, and the growth rates of minority- and majority-type bacilli. We apply a Bayesian approach to find the posterior distribution of ρ, the prevalence of mixed infection, given data from genotyping analysis of mycobacterial DNA collected after division of sputum and subsequent solid culture.

Specimen handling: The mathematical framework developed in this section is based on three assumptions regarding the handling protocol and the specimen – i.e. the sputum sample – from an individual:

1. Each specimen contains at least one strain of *M. tuberculosis*, and may contain more (but we only model detection of two at most). The strain with more bacilli in the specimen is called the majority strain, and the other the minority strain.
2. Each specimen is handled similar to any other and in two phases: sub-division and growth. Sub-division consists of dividing the sputum sample into d groups (only one of which is then cultured). Growth refers to the culture of one of the portions of the sputum sample, over a fixed time T.
3. In both sub-division and growth, the majority and minority strains are assumed to behave independently.

We use X and Y to indicate the number of minority and majority strain cells, respectively; if X and Y are indexed, the index specifies the time. For example, X_0 is the initial number of minority strain cells and Y_T is the number of majority strain cells at time T, after sub-division and growth.

2.1. The minority strain

When the sample is collected, we assume that it contains X_0 minority strain cells. During sub-division, to select a portion $1/d$ of the sample, each cell is chosen with probability $1/d$ or rejected with probability $1 - 1/d$. Therefore the total number of bacteria after sub-division follows the binomial distribution $\mathrm{Bin}(X_0, 1/d)$. Growth is modelled with a birth-only process with birth rate λ_X over a time T. We choose a birth-only process because the death rate is believed to be negligible in comparison to the birth rate in culture, and because it is preferable to minimize the complexity of the model. Birth processes are characterized by a negative binomial distribution (Bailey, 1964, p. 87); in this case, as the process starts with $\mathrm{Bin}(X_0, 1/d)$ cells from the sub-division phase, it follows that the distribution of X_T is $\mathrm{NegBin}(\mathrm{Bin}(X_0, 1/d), 2^{-\lambda_X T})$. Using the law of total probability, the explicit distribution for the number of minority cells after time T is found to be

$$\mathbb{P}(X_T = k \mid X_0) = \sum_{i=1}^{\min(X_0, k)} \binom{k-1}{i-1}\binom{X_0}{i}\left(\frac{1}{d}\right)^{i}\left(\frac{d-1}{d}\right)^{X_0 - i} p_X^{i}(1-p_X)^{k-i}, \tag{2}$$

where $p_X = 2^{-\lambda_X T}$, with λ_X being the growth rate and T the growth time; we refer to the supplement for the derivation. Eq. (2) can be rather impractical because it presents computational challenges due to the size of the binomial coefficients. For this reason, we found an asymptotic approximation:

$$\mathbb{P}(X_T = k \mid X_0) \approx C(1-p_X)^{k}(k-1)^{l(X_0 - 1)} \quad \text{for } k \to \infty, \tag{3}$$

where l and C are constant with respect to k. Interestingly l is also independent from λ_X, X_0, and T, hence it is specific to the handling protocol (see Supplement for details).

2.2. The majority strain

The majority strain is sub-divided and cultured along with the minority strain; following the same reasoning as in Section 2.1, the distribution of Y_T is found to be

$$\mathbb{P}(Y_T = k \mid Y_0) = \sum_{i=1}^{\min(Y_0, k)} \binom{k-1}{i-1}\binom{Y_0}{i}\left(\frac{1}{d}\right)^{i}\left(\frac{d-1}{d}\right)^{Y_0 - i} p_Y^{i}(1-p_Y)^{k-i}, \tag{4}$$

where $p_Y = 2^{-\lambda_Y T}$ and λ_Y is the growth rate of the majority strain. As in the previous section, Eq. (4) is impractical, but here because Y_0 is assumed to be large (Core Curriculum for Disease Control; Palaci et al., 2007) – O(1000) – it is possible to use the Weak Law of Large Numbers to approximate the distribution (4) with a normal distribution (see the Supplement for details):

$$\mathbb{P}(Y_T = k \mid Y_0) \approx \mathcal{N}(k; \mu, \sigma^2), \tag{5}$$

where

$$\mu = \frac{Y_0}{6} 2^{\lambda_Y T} \quad \text{and} \quad \sigma^2 = \frac{Y_0}{d} 2^{\lambda_Y T} (2^{\lambda_Y T} - 1) + \frac{Y_0 d(d-1)}{d^2} 4^{\lambda_Y T}.$$

We use Eqs. (3) and (5) to calculate the distributions of X_T and Y_T numerically in the next sections.

2.3. The conditional prevalence of mixed infection

After the phases of sub-division and growth in culture, genotyping is performed on DNA extracted from mycobacterial cells. In this paper we assume that the genotyping test performed is mycobacterial interspersed repetitive unit-variable number tandem repeat (MIRU-VNTR) typing (Supply et al., 2001). MIRU-VNTR typing is a convenient methodology to detect mixed infections, as multiple alleles at multiple loci are usually interpreted as the presence of mixed infection (Supply, 2005). Clearly, in order to detect a minority strain by MIRU-VNTR or any other method, the minority strain must be present in sufficient numbers. We define the threshold f as the minimum value of the proportion X_T/Y_T at which the minority strain, thus mixed infection, is detectable by MIRU-VNTR typing. It is convenient to introduce the new random variable D (for detection) which is defined by

$$D = 1 \Leftrightarrow X_T / Y_T > f \quad \text{and} \quad D = 0 \Leftrightarrow X_T / Y_T < f. \tag{6}$$

where D is a Bernoulli random variable that is used to model the positive or the negative result of the test for mixed infection for each sputum sample.

Up to this point, we have analysed the dynamics of a single sample. To estimate the prevalence of mixed infection, we need to link our model to the outcome of a study aimed to detect mixed infection. Suppose there are n individual patients in the study and each of their sputum samples is sub-divided and cultured, and then tested for mixed infection. Let the outcome be denoted D_j, for $j = 1..n$. The total number of detected mixed infection is $S_D := \sum_{j=1}^{n} D_j$. Because the D_js are Bernoulli, it follows that

$$\mathbb{P}(S_D = k \mid X_0, Y_0) = \text{Binomial}(k; n, \mathbb{P}(D = 1 \mid X_0, Y_0)), \tag{7}$$

where we recall that $\mathbb{P}(D = 1 \mid X_0, Y_0) = \mathbb{P}(X_T / Y_T > f \mid X_0, Y_0)$.

2.4. Distributions of X_0 and Y_0

Eq. (7) is the distribution of the total number of detected mixed infections in n individuals, in a study that satisfies the initial assumptions outlined at the start of Section 2. To perform computations and statistical inference it is necessary to derive a distribution of S_D that is not conditional on X_0 and Y_0. We have chosen particular distributions for these inputs, but the overall arguments we make about the effects of stochastic growth and low starting cell numbers are not specific to these particular choices.

The number of majority type cells Y_0 is relatively large (Core Curriculum for Disease Control; Palaci et al., 2007) in the samples, with the order of magnitude 10^3. Because Y_0 is a

discrete random variable we choose a discretized gamma distribution. The shape and scale parameters are also chosen to provide a reasonable expected value and variance: recall that the sputum sample is 0.25 mL and the concentration is 5000–10,000 per mL (Core Curriculum for Disease Control):

$$\mathbb{P}(Y_0) = CDF_{Gamma(70,25)}(k) - CDF_{Gamma(70,25)}(k-1), \quad (8)$$

where *CDF* stands for the Cumulative Distribution Function.

The number of minority strain cells in the sample (X_0) is likely to be variable. It will depend on many factors, including the dynamics of bacterial populations in the host, the time of reinfection and the distribution of cell types over different TB lesions. These factors may be elucidated in the future in studies using DEP frequency or single cell technologies, but at the moment there is very little information available to inform us as to the numbers of cells present in sputum samples from diverse infections.

When a minority strain is present, we do not have empirical information about the numbers of minority cells likely to be found in the sputum. We choose a class of distributions parametrized by their expectation E_{min}, for the probability $\mathbb{P}(X_0 = k \mid X_0 \geq 1)$ of finding k minority cells in the sample given that the host has two or more strains. The numbers of minority and majority cells in sputum will depend on a complex series of growth limitations imposed by the host during the course of infection, the relative timing of infection, the extent of in-host competition between the strains, the time that has elapsed before the patient comes to clinical attention and the non-random sampling of the in-host population in sputum. The inoculum for each strain of TB is likely to consist of a relatively small number of bacilli (Balasubramanian et al., 1994), and each strain presumably undergoes a period of exponential growth at some stage. So it is likely that a substantial difference in the robustness of the two strains in the host would lead to the less robust strain either being out-competed or being present in vanishingly small fractions in the host; a minority strain would either be “drowned out” in the exponential phase, or would suffer losses through the complex course of infection if it were not sufficiently robust. Such hosts would never be detected as mixed infections. For these reasons, to maintain high enough cell numbers to comprise ≈ 1% of a sputum sample, any minority strain will likely need to be a fairly strong in-host competitor. Conversely, when more than 2–5% of a sputum sample are minority strain bacilli, they are highly likely to be detected (and this will happen only for highly robust strains that achieve a very strong balance of cell numbers in the host). The problem of bias is most relevant when a minority strain is a robust enough competitor to rise to high enough levels that there is any change of detection, but not so high that detection is effectively certain. Accordingly, we investigate the range of E_{min} in which minority strains comprise between 0 and 2% of the population of bacilli in the sputum i.e. $E_{min} \leq 40$.

Furthermore it must be taken into account that the prevalence of mixed infection corresponds to the probability $\mathbb{P}(X_0 \geq 1)$ of mixed infection present in the sample, that we called ρ. Note that ρ is fundamental for this study, as it is the parameter to be estimated. We use a Poisson distribution for X_0, parametrized by E_{min}:

$$\mathbb{P}(X_0 = k) = \begin{cases} 1-\rho & \text{if } k = 0 \\ \rho \cdot \dfrac{(E_{min}-1)^{k-1}}{(k-1)!} e^{-E_{min}+1} & \text{if } k \geq 1 \end{cases}. \quad (9)$$

2.5. Posterior distribution of the prevalence of mixed infection

In this section we use the Bayesian inference to derive the distribution for the real prevalence of mixed infection, ρ, and we present an estimate of such prevalence. At first it is necessary to evaluate the probability $\mathbb{P}(D = 1)$. The law of total probability can eliminate the condition on X_0 and Y_0 of $\mathbb{P}(D = 1 \mid X_0, Y_0)$ in Eq. (7) using $\mathbb{P}(X_0)$ and $\mathbb{P}(Y_0)$ from Eqs. (8) and (9) respectively. Note that because the distribution of X_0 is linear in ρ, the distribution of X_T and the probability $\mathbb{P}(D = 1)$ are also linear in ρ; this fact reflects the initial decomposition in Eq. (1). It follows that

$$\mathbb{P}(D = 1) = \mathbb{P}(X_T / Y_T > f) = m\rho, \quad (10)$$

where the slope m represents the probability $\mathbb{P}$(detect|mixed infection present) in (1); it depends on the parameters λ_X, λ_Y, E_{min} and T and is calculated numerically using the law of total probability (we refer to the Supplement for further details). Because D has a Bernoulli distribution with probability $m\rho$, the distribution of S_D is binomial; therefore the probability of detecting n_{mix} mixed infection in a study involving n patients is a binomial with n_{mix} successes over n trials and with success probability $m\rho$.

In Bayesian notation, the binomial distribution of S_D is the likelihood. We set an uninformative Beta prior distribution because it is a conjugate prior for the binomial (we refer to the Supplement for further details). This lead to the following posterior:

$$\mathbb{P}(\rho \mid S_D) = \frac{(m\rho)^{n_{mix}}(1-m\rho)^{n-n_{mix}}}{\mathcal{B}_m(n_{mix}+1, n-n_{mix}+1)}, \quad (11)$$

where $\mathcal{B}_m(n_{mix}+1, n-n_{mix}+1) = \int_0^m u^{n_{mix}}(1-u)^{n-n_{mix}} du$ is the incomplete beta function. Fig. 2 shows the posterior distribution of ρ for a range of values of E_{min}.

It is important to note that by the Law of Large Numbers, in the limit n, $n_{mix} \to \infty$, $(1/n)S_D \to m\rho$. Because we observe $S_D = n_{mix}$ mixed infection, this implies that $m\rho \approx n_{mix}/n$, thus

$$\rho \to \frac{1}{m}\frac{n_{mix}}{n} \quad \text{as} \quad n, n_{mix} \to \infty. \quad (12)$$

Alternatively, taking the expectation of the posterior distribution in Eq. (11) and noting that the variance vanishes in the limit yield the same result. Eq. (12) provides a simple estimate for the real prevalence of mixed infection and it suggests that m gives a numerical value of the bias coefficient. Figs. 4 and 5 show the details of its behaviour and sensitivity analysis.

2.6. Deterministic approximation

If the number of initial bacteria is large for both minority and majority strains (for instance if the initial sample is large), the model can be simplified, removing most of its stochasticity. In this case we consider continuous approximations of the variables X_0 and Y_0:

$$Y_0 = \mathcal{N}(\mu_Y, \sigma_Y^2), \tag{13}$$

$$X_0 = (1-\rho)\mathcal{X}_{[0,1)} + \rho\mathcal{N}(E_{min}, E_{min})\mathcal{X}_{[1,+\infty)}. \tag{14}$$

Eqs. (13) and (14) can be considered as limit distributions of (8) and (9) respectively because the Gamma and Poisson distributions converge to a normal when the mean is large. In the deterministic approximation, sub-division and growth are not stochastic, yielding

$$X_T = \frac{2^{\lambda_X T}}{d} X_0, \tag{15}$$

$$Y_T = \frac{2^{\lambda_Y T}}{d} Y_0. \tag{16}$$

Therefore the slope m can be expressed explicitly with the following expression:

$$m = \mathbb{P}(X_T > fY_T) = \mathbb{P}\left(\frac{2^{\lambda_X t}}{d} X_0 - f\frac{2^{\lambda_Y t}}{d} Y_0 > 0\right)$$

$$\times \text{ evaluated substituting } \rho = 1 \text{ in Eq. (14)}$$

The substitution $\rho = 1$ follows from the fact that $\mathbb{P}(X_T > fY_T)$ is linear with respect to ρ, as in Eq. (10). The expression inside the brackets is a linear combination of two normal random variables and therefore is a new normal with known cumulative density function. Therefore

$$m = \frac{1}{2}\left(1 - \text{erf}\left(\frac{f2^{(\lambda_Y - \lambda_X)t}\mu_Y - E_{min}}{\sqrt{2E_{min} + f^2 2^{2(\lambda_Y - \lambda_X)t+1}\sigma_Y^2}}\right)\right). \tag{17}$$

This quantifies the bias in measurements of the prevalence of mixed infection, and how that bias depends on the relative growth rates of minority and majority type cells in culture.

Parameters: The parameters and random variables are given in Tables 1 and 2 respectively.

3. Results

We computed and analysed the posterior distribution of the prevalence of mixed infection assuming that we observe n_{mix}/n = 15% mixed infection in a study of n=500 patients. This baseline estimate of 15% represents a value of mixed infection that is in the range observed in other studies in high TB incidence areas in sub-Saharan Africa (Cohen et al., 2012; Hanekom et al., 2013). Fig. 2 displays a number of different posterior distributions of ρ related to the average number of minority type cells per sputum sample E_{min}. In particular, the smaller the E_{min} is, the larger the expectation of ρ is. This is because numerous opportunities for false negatives arise when the initial population of the minority strain is small or when its growth rate is relatively low. Our posterior estimate accounts for these possible sources of bias, and therefore the estimated mixed infection prevalence may be much higher than the observed 15%. When hosts with mixed infection consistently have a good representation of minority types in their sputum, there are fewer false negatives, m is higher, and the posterior estimate of ρ is closer to the fraction of cases in whom we detect mixed infection (n_{mix}/n).

We evaluated the posterior distribution of the prevalence of mixed infection in a specific study (Warren et al., 2004) and in Fig. 3 we presented four possible posteriors for optimistic and pessimistic values of E_{min} and the growth rates. In the most optimistic scenario the posterior ρ is higher that 0.19 with probability 0.9 and has mean over 0.23. On the other hand a more moderate choice of parameters would indicate that $0.35 < \rho < 0.65$ with probability 0.95.

Fig. 4a shows how the estimate of the prevalence of mixed infection $E[\rho]$ varies for different values of the growth rates. From Fig. 4a we conclude that $E[\rho]$ does not depend directly on λ_X and λ_Y but on their difference $\lambda_Y - \lambda_X$. This is confirmed by the deterministic approximation and in particular by the expression in Eq. (17) for m.

Fig. 4b illustrates the estimate of the prevalence of mixed infection $E[\rho]$ using a contour plot in the plane ($\lambda_Y - \lambda_X$, E_{min}). Fig. 4b demonstrates that the bias in detection of mixed infection is related to the number of minority-type bacilli the sputum sample and the difference of the growth rates. It is noteworthy that there is a region of rapid change in the estimate - for example in Fig. 4(b), if E_{min} is near 20, the estimate is very sensitive to $\lambda_Y - \lambda_X$ when the latter is near 0.1. This implies that, in some studies, the raw estimate n_{mix}/n may be uninformative. Independent estimates of E_{min} and $\lambda_Y - \lambda_X$ would greatly improve our ability to interpret such studies.

In Fig. 5 the four contour plots of the posterior estimate of mixed infection $E[\rho]$ for four different values of the sensitivity threshold f are compared. In each plot the percentage of detected mixed infection is n_{mix}/n = 19% as in Warren et al. (2004). We can see that as f increases, there is a larger area where $E[\rho] > 0.8$. This confirms that the higher the sensitivity thresholds is, the higher the chances are of non-detecting mixed infection. Consequently if a percentage n_{mix}/n is detected then it is likely that the real prevalence ρ is much higher, even close to 1. It is important to note that even if the sensitivity threshold is reasonably small, see the plot where f=0.01, the raw percentage n_{mix}/n is still not a good

estimate for a large portion of the parameter set. We conclude that the correction factor $1/m$ is necessary both when f is small and when f is large.

4. Discussion

We developed a mathematical framework both for assessing the conditions under which current methods underestimate the prevalence of mixed infections and for quantifying the potential magnitude of this bias. We found that the prevalence of mixed infection is biased by a factor m which depends on the growth rates and the population of the minority strain in the initial samples. With the parameters we have used, for example, if initial mixed infection sputum samples had on average 80 minority type cells per mL, the posterior estimate of the prevalence of mixed infection is 33%, compared to the direct measurement of only 15%.

Our framework combines a binomial model for the specimen sub-division with a birth model for bacterial growth in culture, treating the populations of the minority and majority strains separately. Assuming that detection occurs if and only if the ratio between the two populations is greater than a threshold f, we merged the two distributions using the law of total probability. This allowed us to obtain a posterior estimate of the prevalence of mixed infection, represented by the parameter ρ. We found that stochastic effects during specimen handling may reduce the probability of detecting mixed infections. On the other hand if the sample size were increased, fewer stochastic effects would interfere with the detection of mixed infection and the raw percentage could be a more accurate estimate.

The parameter m, and therefore the distribution of ρ, is very sensitive to variation of λ_Y and λ_X. The growth rates and, more importantly, their difference are usually not known and have important consequences for our ability to observe mixed infections in culture. Targeted experiments to measure the growth rates could help inform the extent of bias in estimation of mixed infection. These experiments could be done if it were possible to resample from initial cultures to obtain cells of both types to measure absolute and relative growth rates in culture. The parameter E_{min}, the expected number of minority cells in the specimen given that the host has mixed infection, also affects the distribution of ρ and therefore the bias, as shown in Fig. 2. In this paper we decided to treat E_{min} as a parameter and not as another random variable. In fact we have not modelled the specimen collection, but only the specimen handling: E_{min} has to be interpreted as reflecting the numbers of minority strain bacilli which, if present, will arise in the sputum sample, and this is beyond the scope of this paper. However, the diversity of TB present in a host is potentially complex and heterogeneously distributed, comprising some clonal diversity (Colijn et al., 2011) in addition to diversity resulting from multiple infections. It is reasonable to suspect that not all of the diversity will be represented in sputum samples, and that this is an additional source of bias in detecting mixed infections.

The model presented in this paper is limited in its complexity. Here, we only consider a minority and a majority strain while in reality there may be more than two different strains. Moreover we consider only strains that potentially can be detected with genotyping, i.e. strains with different MIRU types. In a real situation there can be a reinfection with bacteria having the same MIRU type and, therefore, it is impossible to detect such mixed infections

with genotyping. New studies which use methods with additional sensitivity for detecting variation between strains, such as whole genome sequencing, will likely be increasingly used to understand within-host diversity (Sun et al., 2012; Chan et al., 2013; Köser et al., 2013). However, it is important to recognize that most studies will continue to be limited by the examination of sputum samples, which may not represent the actual degree of strain heterogeneity within a host (Cohen et al., 2011). These examples suggest that mixed infections can be even more frequent than in the results reported here. On the other hand, our results also suggest that when the population size of the minority strain is large, > 3%, bias is minimal and the detected prevalence of mixed infection is very close to the real prevalence.

Mixed infection is of interest because it is informative of aspects of the epidemiology of tuberculosis, but it may be particularly relevant to the estimation of the prevalence and infectiousness of drug-resistant TB strains. Drug-sensitive and drug-resistant strains of TB can compete for susceptible hosts, and can re-infect hosts who already have one strain of TB, resulting in mixed infections. A higher estimated incidence of mixed infection could therefore suggest new estimates of the extent of reinfection, and of the level of transmission of resistant strains.

Mixed infections have been detected in nearly 15% of cases in a number of studies (Cohen et al., 2012; Hanekom et al., 2013), and have been considered to play an important role in facilitating the stable coexistence of different strains (Colijn et al., 2009), in altering treatment outcomes (van Rie et al., 2005) and undermining the effectiveness of TB control programmes (Cohen et al., 2008). In this paper we provide strong evidence that estimates of the prevalence of mixed infection can be considerably higher than the raw detection frequency. This implies that mixed infection could play an even more important role in TB epidemiology than raw estimates would suggest.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The work is supported by EPSRC Grant EP/I03626/1 and NIH Grant D2OD006663. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Bailey NT, 1964 The Elements of Stochastic Processes with Applications to the Natural Sciences. Wiley, New York.

Balasubramanian V, Wiegeshaus E, Taylor B, Smith D, 1994 Pathogenesis of tuberculosis: pathway to apical localization. Tuber. Lung Dis 75 (3), 168–178. [PubMed: 7919306]

Chan JZ-M, Sergeant MJ, Lee OY-C, Minnikin DE, Besra GS, Pap I, Spigelman M, Donoghue HD, Pallen MJ, 2013 Metagenomic analysis of tuberculosis in a mummy. New Engl. J. Med 369 (3), 289–290. [PubMed: 23863071]

Cohen T, Colijn C, Murray M, 2008 Modeling the effects of strain diversity and mechanisms of strain competition on the potential performance of new tuberculosis vaccines. Proc. Natl. Acad. Sci. USA 105 (42), 16302–16307. [PubMed: 18849476]

Cohen T, Wilson D, Wallengren K, Samuel EY, Murray M, 2011 Mixed-strain *Mycobacterium tuberculosis* infections among patients dying in a hospital in Kwazulu-Natal, South Africa. J. Clin. Microbiol 49 (1), 385–388. [PubMed: 20980576]

Cohen T, van Helden PD, Wilson D, Colijn C, McLaughlin MM, Abubakar I, Warren RM, 2012 Mixed-strain *Mycobacterium tuberculosis* infections and the implications for tuberculosis treatment and control. Clin. Microbiol. Rev 25 (4), 708–719. [PubMed: 23034327]

Colijn C, Cohen T, Murray M, 2009 Latent coinfection and the maintenance of strain diversity. Bull. Math. Biol 71 (1), 247–263. [PubMed: 19082663]

Colijn C, Cohen T, Ganesh A, Murray M, 2011 Spontaneous emergence of multiple drug resistance in tuberculosis before and during therapy. PloS One 6 (3), e18327. [PubMed: 21479171]

Core Curriculum for Disease Control, Division of Tuberculosis Elimination, Core Curriculum on Tuberculosis: What the Clinician Should Know.

Hanekom M, Streicher EM, van de Berg D, Cox H, McDermid C, Bosman M, van Pittius NCG, Victor TC, Kidd M, van Soolingen D, et al., 2013 Population structure of mixed *Mycobacterium tuberculosis* infection is strain genotype and culture medium dependent. PloS One 8 (7), e70178. [PubMed: 23936157]

Hingley-Wilson SM, Casey R, Connell D, Bremang S, Evans JT, Hawkey PM, Smith GE, Jepson A, Philip S, Kon OM, et al., 2013 Undetected multidrug-resistant tuberculosis amplified by first-line therapy in mixed infection. Emerg. Infect. Dis 19 (7), 1138. [PubMed: 23764343]

Huang H-Y, Tsai Y-S, Lee J-J, Chiang M-C, Chen Y-H, Chiang C-Y, Lin N-T, Tsai P-J, 2010 Mixed infection with Beijing and non-Beijing strains and drug resistance pattern of *Mycobacterium tuberculosis*. J. Clin. Microbiol 48 (12), 4474–4480. [PubMed: 20980571]

Imaeda T, 1985 Deoxyribonucleic acid relatedness among selected strains of *Mycobacterium tuberculosis*, *Mycobacterium bovis*, *Mycobacterium bovis* bcg, *Mycobacterium microti*, and *Mycobacterium africanum*. Int. J. Syst. Bacteriol 35 (2), 147–150.

Köser CU, Bryant JM, Becq J, Török ME, Ellington MJ, Marti-Renom MA, Carmichael AJ, Parkhill J, Smith GP, Peacock SJ, 2013 Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*. New Engl. J. Med 369 (3), 290–292. [PubMed: 23863072]

Kremer K, van Soolingen D, Frothingham R, Haas W, Hermans P, Martin C, Palittapongarnpim P, Plikaytis B, Riley L, Yakrus M, et al., 1999 Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: Interlaboratory study of discriminatory power and reproducibility. J. Clin. Microbiol 37 (8), 2607–2618. [PubMed: 10405410]

Mills HL, Cohen T, Colijn C, 2013 Community-wide isoniazid preventive therapy drives drug-resistant tuberculosis: a model-based analysis. Sci. Transl. Med 5 (180), 180ra49.

Navarro Y, Herranz M, Pérez-Lago L, Lirola MM, Ruiz-Serrano MJ, Bouza E, de Viedma DG, 2011 Systematic survey of clonal complexity in tuberculosis at a populational level and detailed characterization of the isolates involved. J. Clin. Microbiol 49 (12), 4131–4137. [PubMed: 21956991]

Palaci M, Dietze R, Hadad DJ, Ribeiro FKC, Peres RL, Vinhas SA, Maciel ELN, do Valle Dettoni V, Horter L, Boom WH, et al., 2007 Cavitary disease and quantitative sputum bacillary load in cases of pulmonary tuberculosis. J. Clin. Microbiol 45 (12), 4064–4066. [PubMed: 17928422]

Rodrigues P, Gomes MGM, Rebelo C, 2007 Drug resistance in tuberculosis—a reinfection model. Theor. Popul. Biol 71 (2), 196–212. [PubMed: 17174368]

Sarkar R, Lenders L, Wilkinson KA, Wilkinson RJ, Nicol MP, 2012 Modern lineages of *Mycobacterium tuberculosis* exhibit lineage-specific patterns of growth and cytokine induction in human monocyte-derived macrophages. PloS One 7 (8), e43170. [PubMed: 22916219]

Sergeev R, Colijn C, Cohen T, 2011 Models to understand the population-level impact of mixed strain m. tuberculosis infections. J. Theor. Biol 280 (1), 88–100. [PubMed: 21514304]

Sola C, Filliol I, Legrand E, Lesjean S, Locht C, Supply P, Rastogi N, 2003 Genotyping of the *Mycobacterium tuberculosis* complex using mirus: association with VNTR and spoligotyping for

molecular epidemiology and evolutionary genetics. Infect. Genet. Evol 3 (2), 125–133. [PubMed: 12809807]

Sun G, Luo T, Yang C, Dong X, Li J, Zhu Y, Zheng H, Tian W, Wang S, Barry CE, et al., 2012 Dynamic population changes in *Mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients. J. Infect. Dis 206 (11), 1724–1733. [PubMed: 22984115]

Supply P Multilocus Variable Number Tandem Repeat Genotyping of *Mycobacterium tuberculosis*. Technical Guide, 2005.

Supply P, Lesjean S, Savine E, Kremer K, Van Soolingen D, Locht C, 2001 Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. J. Clin. Microbiol 39 (10), 3563–3571. [PubMed: 11574573]

van Embden J, Cave MD, Crawford JT, Dale J, Eisenach K, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick T, 1993 Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J. Clin. Microbiol 31 (2), 406–409. [PubMed: 8381814]

van Rie A, Victor TC, Richardson M, Johnson R, van der Spuy GD, Murray EJ, Beyers N, van Pittius NCG, van Helden PD, Warren RM, 2005 Reinfection and mixed infection cause changing *Mycobacterium tuberculosis* drug-resistance patterns. Am. J. Respir. Crit. Care Med 172 (5), 636. [PubMed: 15947286]

Warren R, Richardson M, van der Spuy G, Victor T, Sampson S, Beyers N, van Helden P, 1999 DNA fingerprinting and molecular epidemiology of tuberculosis: use and interpretation in an epidemic setting. Electrophoresis 20 (8), 1807–1812. [PubMed: 10435453]

Warren RM, Victor TC, Streicher EM, Richardson M, Beyers N, van Pittius NCG, van Helden PD, 2004 Patients with active tuberculosis often have different strains in the same sputum specimen. Am. J. Respir. Crit. Care Med 169 (5), 610–614. [PubMed: 14701710]

AUTHOR-HIGHLIGHTS

- We develop a mathematical model for study designs aimed to detect TB mixed infection.
- We obtain Bayesian posterior estimates of the prevalence of mixed infection.
- The bias between the posterior estimate and the observed prevalence is discussed.
- The posterior estimate can be substantially higher than the raw percentage.

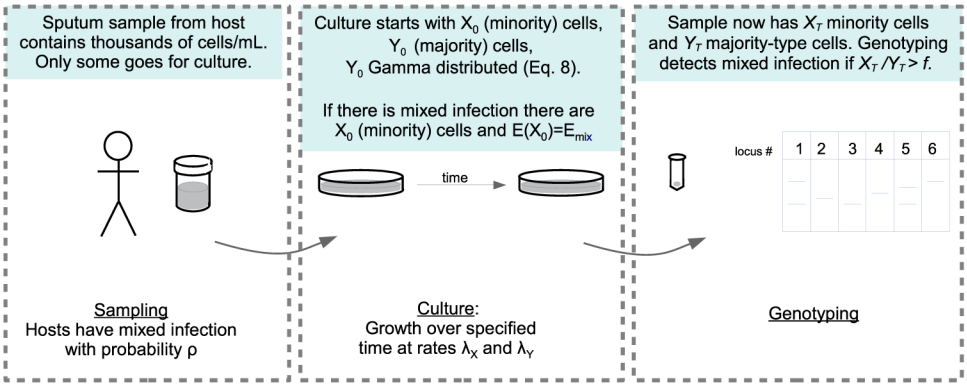

Fig. 1.

Schematic of the process of sampling, culture and genotyping.

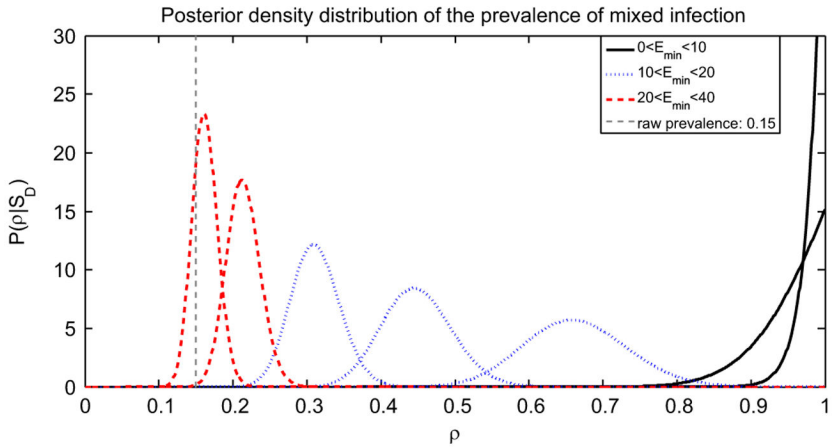

Fig. 2.

Posterior density distributions of the prevalence of mixed infection $\mathbb{P}(\rho \mid S_D)$ when both the growth rates for minority and majority type cells equal 1 and for different values of the expected number of minority cells in sputum E_{min}. Bearing in mind that the larger the E_{min} is, the smaller the mean of the distribution is, the values of E_{min} that we used are 39, 25, 18, 14, 11, 8, 4. We considered $n = 500$ patients, $n_{mix} = 75$ of whom are detected with mixed infection. A naive estimate from the data would indicate a mixed infection prevalence of approximately $n_{mix}/n = 15\%$, corresponding to $\rho = 0.15$. However the posterior distribution has mean close to 0.15 only if E_{min} is large ($E_{min} > 40$). The posterior distributions have a much higher mean as E_{min} decreases.

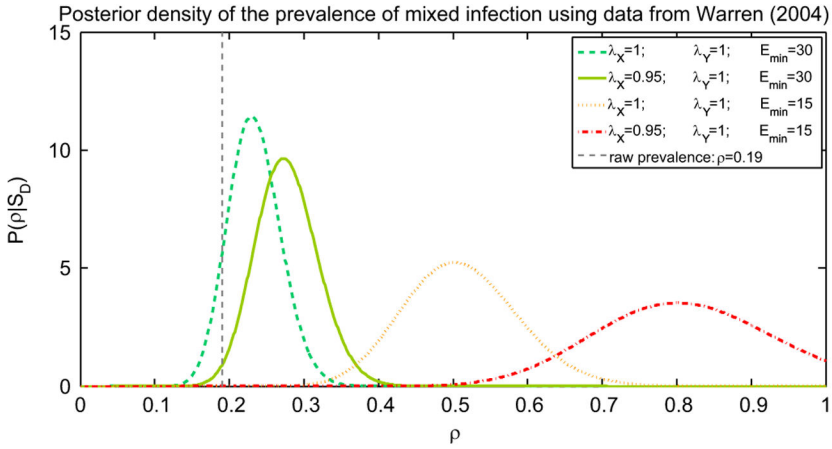

Fig. 3.

Posterior density distributions $\mathbb{P}(\rho \mid S_D)$ of the prevalence of mixed infection ρ, for different values of the expected number of minority cells in sputum E_{min}, one optimistic and one pessimistic, and two different combinations of the growth rates of minority and majority type cells, λ_X and λ_Y. We considered $n = 186$ patients, $n_{mix} = 35$ of whom are detected with mixed infection, as in Warren et al. (2004). The values for the growth rates are in line with the estimations in Sarkar et al. (2012). The raw estimate from the data would indicate a mixed infection prevalence of approximately $n_{mix}/n = 19\%$, corresponding to $\rho = 0.19$, however we observe that, even in the most optimistic scenario (green dashed line) the posterior ρ is higher that is 0.19 with probability 0.9 and has mean over 0.23. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

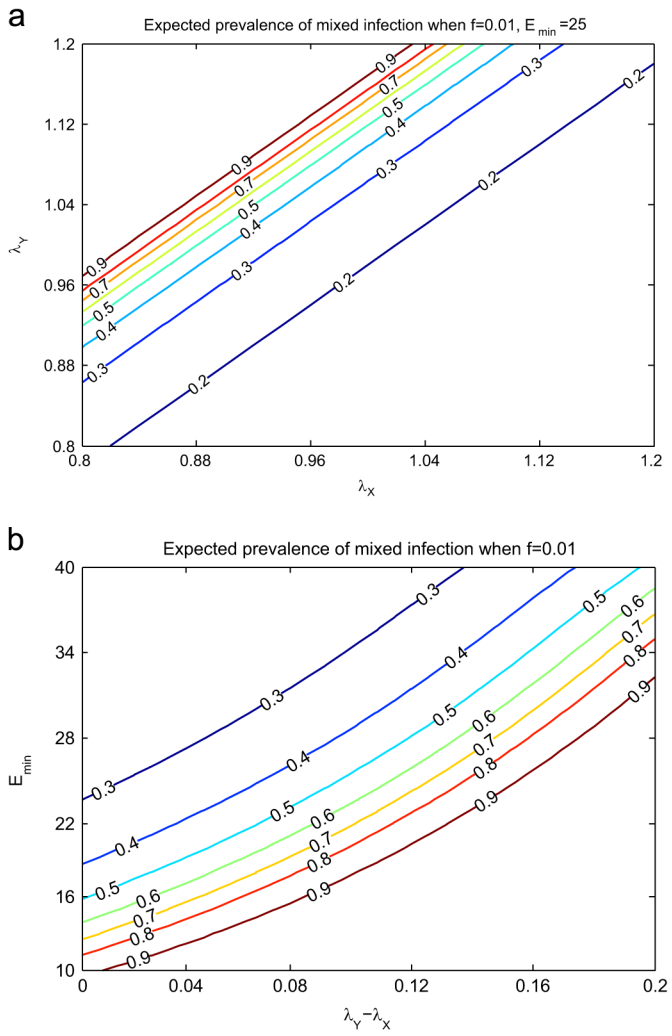

Fig. 4.

We calculated $E[\rho]$, the expected prevalence of mixed infection, in a study with $n_{mix} = 75$ individuals detected with mixed infection among $n = 500$ patients. In (a) for fixed values of the average number of minority type cells in sputum, $E_{min} = 25$, and of the sensitivity threshold $f = 0.01$ we can see a numerical evidence that $E[\rho]$ depends on the difference of the growth rates $\lambda_Y - \lambda_X$ and not on the two growth rates independently; this is confirmed by the deterministic approximation, Eq. (17). In panel (b) how $E[\rho]$ varies taking into account the difference $\lambda_Y - \lambda_X$ on the x-axis and the parameter E_{min} on the y-axis is shown. From panel (b) we note that there is a large area (bottom-right) in the parameter space where $E[\rho]$ is close to 1, estimate very far from the detected 0.15. Although $E[\rho]$ decrease rapidly from 0.9 to 0.6, most part of the parameter space features an expected prevalence of mixed infection larger than 0.3.

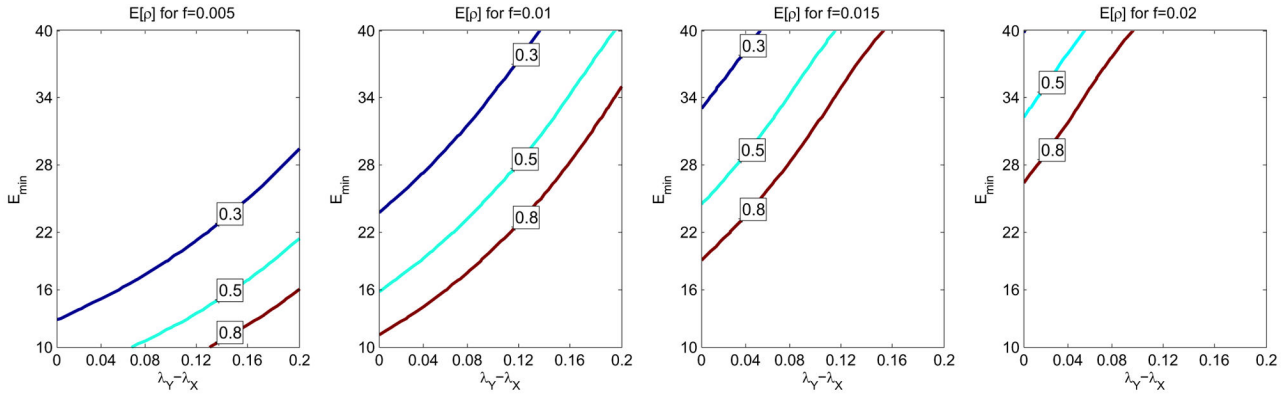

Fig. 5.

Contour lines of the expected prevalence of mixed infection $E[\rho]$ are drawn for four different values of the sensitivity threshold f of the genotyping method. Every plot shows three elevation levels (0.3, 0.5 and 0.8) when the difference of the growth rates $\lambda_y - \lambda_X$ spans between 0 and 0.2 (x-axis) and the expected number of minority strain cells in sputum E_{min} spans between 10 and 40 (y-axis). To produce each plot we simulated a study involving 500 patients among whom 75 are detected with mixed infection. From the comparison of the contour plot we evince that as the threshold f increases, a greater portion of the parameter space features a high (> 0.8) expected prevalence of mixed infection. On the other hand, when f is small, $E[\rho]$ is closer to the detected prevalence 15%. This not only confirms that a small sensitivity threshold allows more precise results, but also shows that even when such threshold is small, the raw percentage 15% should be corrected to give a good estimate of the real prevalence of mixed infection.

Table 1

Parameters.

Parameters	Description	Range/expression
ρ	Probability of presence of mixed infection in sputum sample	Eq. (9)
E_{min}	Mean of minority type cell in the sputum sample given presence of mixed infection	1–40
λ_X	Growth rate of minority strain cells	1–1.2, from Sarkar et al. (2012)
λ_Y	Growth rate of majority strain cells	1–1.2, from Sarkar et al. (2012)
T	Growth time	7 days
l	Exponent of the approximation function (3)	0.336 (see Supplement)
C	Coefficient in the approximation function (3)	Supplement Eqs. (8) and (10)
f	Threshold for detection of mixed infection	0.005–0.02
d	Number of parts the sputum sample is divided in during handling	4

Table 2

Random variables.

Random variable	Description	Expression
X_0	Number of minority strain cells in the sputum sample assumed mean: ρE_{mix}	Eq. (9)
Y_0	Number of majority strain cells in the sputum sample assumed mean: 1750 from Core Curriculum for Disease Control (), Palaci et al. (2007)	Eq. (8)
X_T	Number of minority strain cells after specimen handling	Eq. (2)
Y_T	Number of majority strain cells after specimen handling	Eq. (4)
D	Bernoulli random variable representing the test result for mixed infection	Eq. (6)
S_D	Total mixed infection detected in a study with h patients	Eq. (7)