Journal of Theoretical Biology 270 (2011) 154–163

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

An investigation into the statistical properties of TB episodes in a South African community with high HIV prevalence

Carel Pretorius [a,*], Peter Dodd [b], Robin Wood [c]

[a] SACEMA, c/o StIAS, DST/NRF Centre of Excellence in Epidemiological Modelling and Analysis, Stellenbosch University, 19 Jonkershoek Road, Stellenbosch 7600, South Africa
[b] Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom
[c] Desmond Tutu HIV Centre, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, South Africa

ARTICLE INFO

Article history:
Received 28 January 2010
Received in revised form
12 August 2010
Accepted 8 October 2010
Available online 4 November 2010

Keywords:
Dynamical model
Master equation
Point process
Two-time correlation function

ABSTRACT

Continuous differential equations are often applied to small populations with little time spent on understanding uncertainty brought about by small-population effects. Despite large numbers of individuals being latently infected with *Mycobacterium tuberculosis* (TB), progression from latent infection to observable disease is a relatively rare event. For small communities, this means case counts are subject to stochasticity, and deterministic models may not be appropriate tools for interpreting transmission trends. Furthermore, the nonlinear nature of the underlying dynamics means that fluctuations are autocorrelated, which can invalidate standard statistical analyses which assume independent fluctuations.

Here we extend recent work using a system of differential equations to study the HIV-TB epidemic in Masiphumelele, a community near Cape Town in South Africa [Bacaër, et al., J. Mol. Biol. 57(4), 557–593] by studying the statistical properties of active TB events. We apply van Kampen's system-size (or population-size) expansion technique to obtain an approximation to a master equation describing the dynamics. We use the resulting Fokker–Planck equation and point-process theory to derive two-time correlation functions for active TB events. This method can be used to gain insight into the temporal aspect of cluster identification, which currently relies on DNA classification only.

1. Introduction

Recent DNA fingerprinting work in Masiphumelele, a community near Cape Town, South Africa, allows us to classify TB strains, sampled from active TB cases, by their DNA type (Middelkoop et al., 2009). This data set comprises *Mycobacterium tuberculosis* (MTB) sputum samples for which the DNA types (i.e. strains) of TB have been determined by restriction fragment length polymorphism (RFLP) techniques. It shows that active TB episodes have a tendency to cluster, i.e. a few cases of the same DNA type are registered in a short time interval.

Fig. 1 shows DNA-classified sputum samples collected from 149 adults (15–60 years) TB patients from 2001 to 2005. Only the main strains as well as HIV status (with · indicating HIV_-, + indicating HIV_+ and × indicating unknown) are shown. Clusters within the W451 and CC100 strains among HIV_+ cases are particularly striking. If clusters are associated with recent infection, this indicates ongoing TB spread. This is a worry for TB control programs, particularly with drug-resistant strains. Another complication is the high HIV burden in the community, as HIV_+ individuals are more susceptible to some TB strains than others (e.g. W451-Beijing Middelkoop et al., 2009). A TB outbreak among HIV_+ cases may spread to the HIV_- population.

Mathematical models have been used to shed light on the dynamical interpretation of TB clustering data. Murray (2002) linked an individual-based dynamical model to the molecular epidemiology of TB. The study sheds light on the expected distribution of cluster sizes, which in turn is interpreted in terms of the basic reproduction number and strain diversity. The analysis of synthetic data presented in Murray (2002) suggests that interpretation of cluster sizes alone does not provide adequate information to assess the impact of TB control programs. Simulation results corroborate reports of large clusters found in studies in low-TB-incidence areas. Glynn et al. (1999) provide a statistical framework to assess the uncertainty in cluster distributions resulting from incomplete sampling. Their work suggests that detailed information of the TB cases that comprise the sample, including knowledge of the timing of infection events, is necessary in order for reliable inferences about ongoing transmission to be drawn from clustering data.

Within the scope of commonly used epidemiological modelling tools, there is currently no method available for classifying temporal clusters of TB events. Previous studies have used models to relate the distribution of clusters sizes to disease transmission. Analyses of these TB clusters make use of statistical methods for clustered data, which typically rely on the fact that the TB episodes

* Corresponding author. Tel.: +27 21 808 2589; fax: +27 21 808 2586.
E-mail address: c.d.pretorius@gmail.com (C. Pretorius).

0022-5193/$ - see front matter © 2010 Published by Elsevier Ltd.
doi:10.1016/j.jtbi.2010.10.012

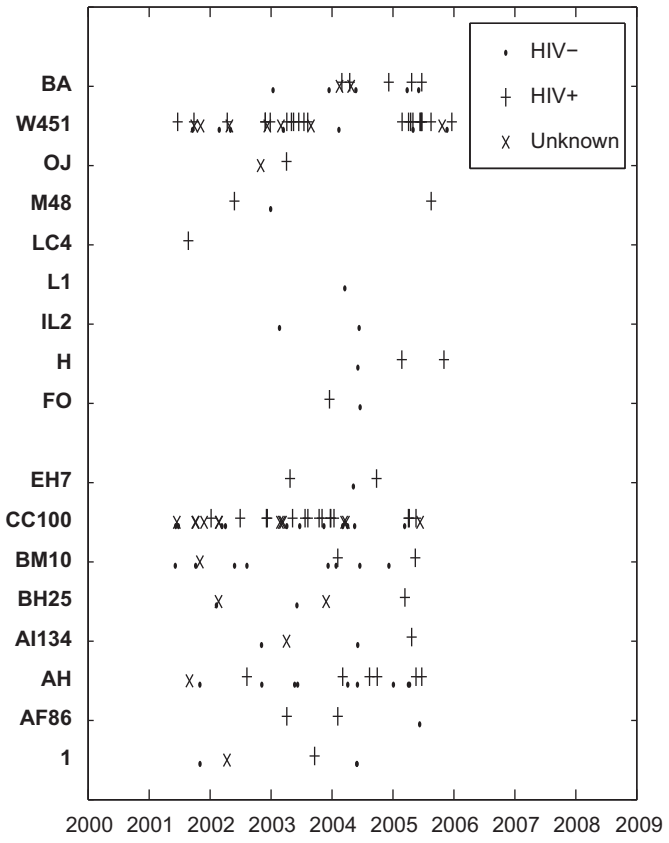

Fig. 1. *Mycobacterium tuberculosis* strains from the study community, adapted from Middelkoop et al. (2009). The data show the distribution of active TB events by HIV status and MTB genotype.

are generated by an underlying dynamical model. From a statistical point of view, a larger than average density of events implies clustering.

What is the average density of active TB events in a time interval and can this reliably be estimated by visual inspection? What is the timescale beyond which two events are not statistically correlated, when they are of the same DNA type? Is the study 'window' of five years, during which the MTB data published in Middelkoop et al. (2009) was recorded, sufficient to capture typical MTB clusters in the community? If it is not sufficient, when will the study (which continues to collect TB strain data), have recorded 'enough' of the epidemic history? This information is vital when associations between MTB, strain type, HIV status, age, gender are investigated.

Positive correlations mean that TB cases have a tendency to cluster in time and this must be taken into account even when assessing whether a series of cases indicates an increasing underlying trend. Standard tests such as the *t*-test assume that fluctuations at different points in time are independent and therefore do not apply. The presence of HIV changes many of the rates between different TB disease states, and therefore affects the correlation structure.

A system of ODEs was recently used to model dual HIV-TB epidemics in the same community. The system provides an accurate macroscopic description when the population is 'large', mixes homogeneously and fluctuations due to individual stochasticity are small. However, the population modelled here consists of only 10,000 individuals and small-population effects may limit the validity of the macroscopic equations. One of the questions we seek to answer here is what do 'large' and 'small' mean in this context? How large do we expect the stochastic corrections to the deterministic model used in Bacaer et al. (2008) to be?

We use a 'system size expansion' technique developed by van Kampen (1992, Chapter 10) to study the stochastic process underlying the dynamical model, and to gain insight into the statistical properties of the system's fluctuations. A Fokker–Planck approximation for the master equation of the system is derived. Using this equation we are able derive differential equations for the variances and co-variances of the fluctuations. This method thus gives a handle on both the deterministic and stochastic descriptions of the system (Plischke and Bergersen, 2006, Chapter 8). We use point process theory to explore some of the above questions regarding the interpretation of TB clustering data. Probability density functions for TB events are derived directly from the underlying dynamical model, for events that lead to active TB episodes. Two-time correlation functions are calculated using these density functions and are used to study the timescale over which active TB events can be expected to cluster.

Section 2 defines a population where each individual is in one of three MTB states: susceptible to MTB infection, latently infected with MTB and actively infected. It derives the nonlinear master equation which governs how the probability distribution of the population over these states changes with time. Section 3 applies a population-size expansion technique to approximate the master equation. Simulation results are presented in Section 4, which show the relative size of fluctuations in the system, and compares the analytically modelled fluctuations to that which can be obtained by stochastic simulation. Section 5 uses point process theory to derive two-time correlations for active TB events.

2. The master equation of a TB-only model

A summary of 25 mathematical and simulation models developed and analyzed in different epidemic contexts between 1992 and 2008 is presented in Bacaer et al. (2008, Table 2). Many of these demonstrate the mathematical properties of HIV-TB models (Schulzer et al., 1992; Raimundo et al., 2002; Moghadas and Gumel, 2003; Naresh and Tripathi, 2005; Lungu, 2007). Some evaluate the potential impact of control strategies advocated by the World Health Organization (WHO) (Dye et al., 1998; Murray and Salomon, 1998). A small subset (Currie et al., 2003, 2005; Hughes et al., 2006; Raimundo et al., 2003; West and Thompson, 1997) provides guidance for fitting models to real time-series data of HIV-TB epidemics. Drawing particularly upon these works, Bacaër et al. developed and tailored a model to the HIV-TB epidemic in Masiphumelele. We adopt this model in our analysis of TB events in the same community.

Consider the following compartmental model of TB dynamics, with S, E and I representing susceptible, latent and active TB cases, respectively. In this model all TB cases have the same HIV status, either HIV_- or HIV_+. The total population size is given by $\Omega = S+E+I$. The parameters for transition between compartments are shown in Table 1. The model captures the TB processes of primary infection (resp. reinfection), endogenous reactivation, enrollment for TB treatment, cure through treatment, natural recovery without treatment and TB mortality. Using this notation, a dynamical model for HIV_- TB cases only is given by the following system of coupled differential equations:

$$\begin{aligned}\frac{dS}{dt} &= B - kSI/\Omega - \mu S,\\ \frac{dE}{dt} &= (p'S - qE)kI/\Omega - (a+\mu)E + bI,\\ \frac{dI}{dt} &= (pS+qE)kI/\Omega - (b+m)I + aE.\end{aligned} \tag{1}$$

To study this system microscopically we have to derive a master equation for the time-dependent transition probability of the state