Published: November 2011

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

Carel Pretorius - Associate, Futures Institute, Glastonbury, Connecticut, United States of America

There are few students in epidemiological modeling and analysis who can resist the temptation to fit a theoretical disease model to real epidemic data. A recent DNA fingerprinting project from Masiphumelele, a township near Cape Town, offered such a temptation. The result is a short journey into the world of statistically rare events, in this case brought about by the relatively small size of Masiphumele and by the slow reactivation rates of TB.

The dataset consists of registered TB events, corresponding to the approximate time when a TB transmission event occurred (1). Clusters formed by TB strains isolates W451 and CC100 among HIV+ cases were particularly striking. These clusters may point to ongoing transmission, which could exasperate the already desperate situation in the township. A number of questions arise. The data show apparent clusters, but what are the properties of a typical cluster? Are the clusters we see the result of correlated infection events, e.g. an infection chain between HIV+ TB cases, or do they simply appear to be clustered when in reality there is no connection (correlation in statistical parlance) between them?

Mathematical theory of point processes

The best guide for this journey is the statistical theory of point processes. It helps us frame the questions we need to ask in order to interpret TB event data correctly. We developed a point process theory for TB events starting from the simple differential equation dynamical model we previously developed to understand the population level (macroscopic) aspects of TB in this community (2). To start building a point process theory for the TB model, we developed a dynamical description of all the random events that occur in the model, comprising birth, death, primary infection, re-infection and endogenous-activation events among susceptible, latently and actively infected sub-populations. To make this step analytically tractable we used only one HIV state in the model.

We then used van Kampen's 'population size' expansion to derive a differential equation for the variances and co-variances of these random and fluctuating events (3,4). This is a so-called Fokker–Planck equation (FPE) and it describes the fluctuations as Gaussian noise around the equilibrium population level model. We solved the differential FPE with a standard ordinary differential equations (ODE) solver in Matlab, and checked the result against Gillespie's stochastic simulation technique (5). Finally, we used the FPE to study the temporal aspects of TB clusters, and obtained an understanding of the timescale between active TB events.

Insights from point process theory

The first insight from the model applies to many deterministic models (see (2) for a summary)) used nowadays to model epidemics at the community level. We showed that fluctuations in the population variables (i.e. the variables that keep track of the number of susceptible, latently and actively infected individuals) become small relative to the size of the sub-populations when the population size is closer to 40,000. Given that many TB interventions and trials are often run and evaluated at a community level, we should expect a significant level of uncertainty in population-level estimates derived from macroscopic models. This uncertainty is seldom explicitly handled in the growing field of epidemic modeling, even for models applied to small communities.

The two-time correlation function

The next important insight from the model derives from the so-called two-time correlation function for events $g_2(t_1,t_2)$ (3, p. 41). It measures increased probability of observing an event of a particular type at time t_2, given an event of a certain type at time t_1. This is equivalent to measuring the degree to which the joint density of events at t_1 and t_2 is greater than the density at t_1 and t_2. This can be thought of as reflecting the causal influence of the first event on the second via both direct (e.g. infection, reactivation) and indirect routes (chains of such events): what is the increase in probability for an event to occur at t_2 knowing that another event occurred at t_1? Note that the events do not have to be of the same type.

Implications for TB control

Thus, even though chains of causation are not explicitly tracked in this type of framework, influence can be assessed at the statistical level by examining the correlations. This gives us a fairly detailed statistical description of clusters of events. For example, events

that are separated by a timescale longer than the correlation timescale are unlikely to be part of the same transmission cluster. The method can therefore be used as the basis for understanding the temporal component of TB strain clusters, which are currently defined mostly in terms of DNA type, with geographical linkage, social interaction and other processes also playing a role.

Our analysis shows that endogenous activation events are correlated over long time scales, and are statistically likely to be part of short timescale clusters. This casts reasonable doubt around the presence of apparent clusters of isolates W451 and CC100 among HIV+ TB cases, and whether the data support the assumption of ongoing transmission of these strains. If the modeled intensities of active TB events are validated, then the model can be used to warn against spurious conclusions from measured clustered data. For example, a correlation effect may be incorrectly attributed to an infection trend, or even to a particular type of infection chain, while in reality it could be purely due to chance stemming from fluctuation. These observations have direct implication for TB control measures in the community: ongoing infection chains require more forceful intervention than reactivation events, which can be handled through standard TB control strategies.

Improvement and further work

Previous modeling work (6,7) has highlighted the potential for study time-windows and case-detection rates to bias interpretations of clustering statistics. Far less is known about how these may differ between data derived from HIV+ and HIV- TB cases. To do a complete analysis of a model with both HIV- and HIV+ TB cases is possible but cumbersome. We simply relabeled the HIV state in our model to HIV+, and changed all the dynamical parameters to ones corresponding to HIV+ individuals, who experience higher rates of primary infection, re-infection and endogenous reactivation. Higher rates of reactivation among HIV+ individuals may mean that their strains are a priori more likely to be drawn from the latent pool. Indeed, the analysis shows shorter correlation timescales among HIV+ individuals which suggests that a more stringent criterion of temporal linkage may be needed for cluster determination among HIV+ TB cases, compared with HIV- TB cases.

As more detailed data spanning a longer time interval become available it may be possible to evaluate if clustered active TB episodes are consistent with the dynamics of a closed community. If they are not, and therefore require exposure to external sources of infection to explain the observed clustering of TB events, it raises concerns for TB treatment programs. Treating TB cases only in a particular community will not reduce its TB burden: TB treatment programs would have to reach the wider community in order to be effective.

This analysis could find broad validation in epidemiological models where the transmission term is an assumed non-linear term, with few possibilities of validating the assumption against real data. The model is usually checked against aggregated population count data, which can be fit by many functional forms for the transmission term. Our approach of studying the underlying point process may shed light on whether a mass action model can produce the observed clustering of TB events. A model accounting for local contacts (household, schools) as well as global contacts (e.g. in the wider community) may be essential for modeling temporally clustered active TB events.

Note that the above is an abbreviated version of the following article: Pretorius C, Dodd P, Wood R. An investigation into the statistical properties of TB episodes in a South African community with high HIV prevalence. J Theor Biol. 2011;270(1):154-63.

Carel Pretorius, Associate, Futures Institute, Glastonbury, Connecticut, United States of America. Research interests: Developing strategic planning tools for HIV and TB epidemics.
CPretorius@futuresinstitute.org

References:

1. Middelkoop K, Bekker L-G, Mathema B, et al.Molecular Epidemiology of Mycobacterium tuberculosis in a South African Community with high HIV prevalence. J Infect. Dis. 2009;200 (5):1207–1211.
2. Bacaer N, Ouifki R, Pretorius CD, Wood R, Williams B. Modeling the joint epidemics of TB and HIV in a South African township. J Math Biol. 2008;57(4):557–593.
3. van Kampen NG. Stochastic Processes in Physics and Chemistry (second ed). Amsterdam: North-Holland; 1992
4. van Kampen NG. The expansion of the master equation. Adv Chem Phys. 1993;34:245–309.
5. Gillespie D. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 1977;81(25):2340–2361.
6. Murray M. Determinants of cluster distribution in the molecular epidemiology of tuberculosis. Proc Natl Aca Sci. 2002;99(3):1538–1543.
7. Glynn JR, Vynnycky E, Fine PEM. Influence of sampling on estimates of clustering and recent transmission of Mycobacterium tuberculosis derived from DNA fingerprinting techniques. Am J Epidemiol. 1999;149(4):366-371.