Bayesian Evidence Synthesis for estimating HIV burden

Human Immunodeficiency Virus (HIV) is associated with high morbidity and health-care costs and, in developing countries, with significant mortality and a potentially large number of years of life lost.

Reliable estimates of the disease burden are crucial to inform planning of health-care provision and to implement and evaluate public health policies aimed at reducing transmission. The magnitude of this burden can be quantified in terms of: HIV prevalence (the proportion of the population infected); the proportion of undiagnosed HIV infections; and HIV incidence (i.e. the rate of new infections). Knowledge of these quantities is generally required in both the general population and in specific groups at high risk of infection and in different locations to maximise.

Estimation of these quantities is not straightforward: data informing them directly are not readily available. In fact, available information is typically partial, biased, refers to populations different from those of interest (or to mixtures of populations at risk) and/or is sparse. This clearly depends on the quality of public health surveillance in different countries, but, generally, estimation of HIV burden poses a real challenge even for countries with well developed surveillance systems.

However, it is often the case that, although there is paucity of direct data, sources of "indirect" information (i.e. information that inform "functions" of the quantities of interest) exist that can be usefully integrated in the estimation procedure. This additional data are usually ignored in most estimation methods as difficult to incorporate. So, on one hand, unverifiable assumptions and ad hoc adjustments might be made to compensate for the lack of information on "basic parameters" (e.g. HIV prevalence in particular groups or ages) and, on the other hand, an inefficient use is made of the information that is available. This typically results in imprecise and often biased estimates due to the selective nature of the data used in the estimation.

The statistical challenge is then to provide a method capable of combining coherently direct and indirect information from multiple sources appropriately accounting for uncertainty and biases, to derive robust estimates that are consistent all information. This combination is known as "evidence synthesis" and the Bayesian approach provides a natural inferential framework for such synthesis, being intrinsically rooted on the idea of combining evidence.

Given a parameter of interest in estimating, the Bayesian paradigm is based on Bayes' theorem: "prior" knowledge on the parameter, expressed through a "prior" distribution is updated through the likelihood of the observed data, to obtain a "posterior" distribution of the parameter, on which inference is based. The posterior distribution combines initial and experimental knowledge and reflects prior and experimental uncertainty.

This principle can be easily generalised to the situation where there are multiple parameters and multiple sources of data informing, directly or indirectly, the parameter(s) of interest: a statistical model linking (direct and indirect) data and unknown parameters (e.g. prevalence) is specified and combined with prior models for the parameters to produce posterior distributions for parameters and related quantities. In practice, these posterior distribution cannot be computed analytically, and simulation methods are employed to derive them (Markov Chain Monte Carlo (MCMC)).

The Bayesian approach has a number of advantages:

- it allows the inclusion of ancillary information (e.g. on possible biases in the data) in the form of carefully chosen prior distributions;

- it helps dealing with data sparseness issues, and consequent poorly informed parameters, by naturally allowing hierarchical structures to share information within the model;

- posterior distributions for any function of the parameters can be easily derived through MCMC.

By making a rational and exhaustive use of the full body of available data, an evidence synthesis approach would flag any inconsistency in the array of data. These conflicts are important to detect as they may highlight biases in, or misinterpretations of, the collected data. Finally, decisions around research prioritisation and service provision are more rationally and robustly taken when driven by a comprehensive, rather than selective, use of available information (Claxton et al, 2002).

Bayesian evidence synthesis has already been successfully employed in the UK (Goubar et al, 2008; Presanis et al, 2010; De Angelis et al, 2013) and the Netherlands (van Veen et al, 2010; Conti et al, 2011) to estimate HIV prevalence and incidence levels and trends over time. Nevertheless, many challenges remain in formulating appropriate models to describe the data-generating mechanisms of surveillance systems in many countries. A workshop to discuss the potential for synthesis of the data available in countries such as Poland, Spain, Denmark and Italy would be invaluable in further advancing evidence synthesis methods for epidemiology.