

On complexity, realism and usefulness of individual-based models

Wim Delva - Epidemiologist, SACEMA, and Ghent University, Hasselt University and KU Leuven, Belgium.

With the famous quote “All models are wrong, but some are useful” and the aphorism “Everything should be made as simple as possible, but no simpler”, we are touching directly on the central challenge of all models in epidemiology: How do we build useful models that capture only the essential elements to explain the mechanisms of disease transmission and control while still doing justice to the available data? Here, we define an epidemiological model as useful if it leads to insights with actionable implications for more effective prevention or management of disease.

If we imagine the three-dimensional space of usefulness, complexity and realism, as in Figure 1 below, we can ask where in this space can or should models be situated? Assuming we are only interested in useful models, we can rephrase the question as “are complexity and realism necessary and sufficient conditions for models to be useful?” I believe that for both complexity and realism it is true that they are neither necessary nor sufficient conditions for useful modelling.

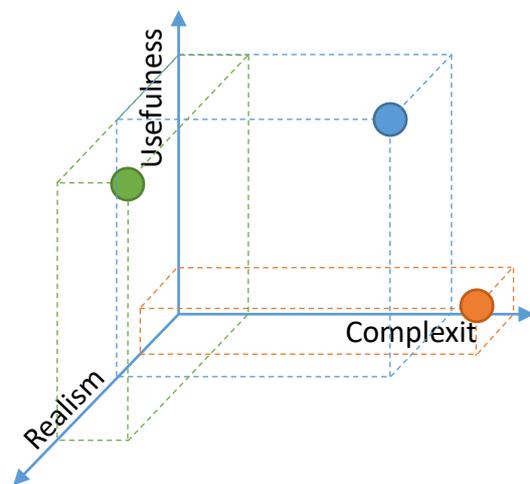


Figure 1. Three imaginary models in the three-dimensional space of usefulness, complexity and realism. The green and blue models are equally useful, but vary greatly in complexity and realism. The orange model is very complex, but the added complexity does not make it more realistic or useful.

For individual-based models (IBMs)* in epidemiology, thinking about links between a model’s complexity, its closeness to reality and its usefulness is particularly pertinent. This is because the bottom-up, modular and hierarchical structure

of this type of models makes it relatively easy to increase the level of heterogeneity and complexity represented by the model. Moreover, the rationale for doing so is often a desire to build more realistic models. The implicit belief is that by virtue of being more realistic, models also become more useful.

But is this necessarily true? The ability to let population-level features of the system emerge from individual-level processes and events, is arguably the most important quality of IBMs. It makes them a convenient, intuitive formalism to express and test hypotheses around putative biological or behavioural mechanisms of epidemic spread and control. The usefulness of the IBM then correlates with the extent to which these hypotheses – which may enjoy support or face rejection after the model output is analysed – have the potential to influence decision-making in the real world. No matter how realistic the model may be, if the relevant actors do not see how the model results have any bearing on their thinking, the model is of little use. Here, “relevant actors” do not only include obvious users of the model results such as public health officials, medical doctors, or vaccine manufacturing companies, but also biostatisticians, other modellers and the model’s own author.

Models with a narrow scope may require little complexity to represent the relevant features of the system adequately, and produce useful results. An example of this would be a model that only considers the effect of one variable, say uptake of male circumcision, on HIV incidence over a short time window in a homogeneous subgroup of school-going adolescent men. Expanding the time window, increasing heterogeneity in the study population and considering interventions with more indirect pathways (e.g. a nation-wide initiative to create jobs and improve the quality of housing and education), would naturally require a more complex model. The problem though, is that with more complexity – which translates into more input parameters to be empirically estimated, “guesstimated” or calibrated – also comes more room for uncertainty and error. Consequently, it is by no means guaranteed that more complex models correspond more faithfully with reality.

Perhaps a harder statement to defend is that highly unrealistic IBMs can still be useful. Yet, it needn’t be a difficult case to make, given the long track

record of very simple models in all areas of science that caused breakthroughs and paradigm shifts. However, because many IBMs are born out of the dissatisfaction with population-average, compartmental models that are deemed “not realistic enough”, it may be counterintuitive to developers and users of IBMs that there may be unique value in building an overly distorted caricature of the real world. An example from my own work in the area of sexual age-mixing and HIV epidemic spread is a series of “toy models” that illustrate how a sufficiently large variation in age gaps between an infected individual and their partners over the duration of the infectious period is

key for long-term persistence of the virus in the population. Whether the average partner age gap is small or large, does not determine long-term transmission dynamics. Indeed, HIV would go extinct after a few generations in a world where the average age gap between sexual partners is large but not much variation exists around this average (Figure 2). Obviously, neither of these two model worlds resemble age-mixing patterns observed in the real world, but these toy models have helped to crystallise the debate around the importance of age-mixing for HIV transmission, and have informed the statistical analysis of empirical age-mixing data.

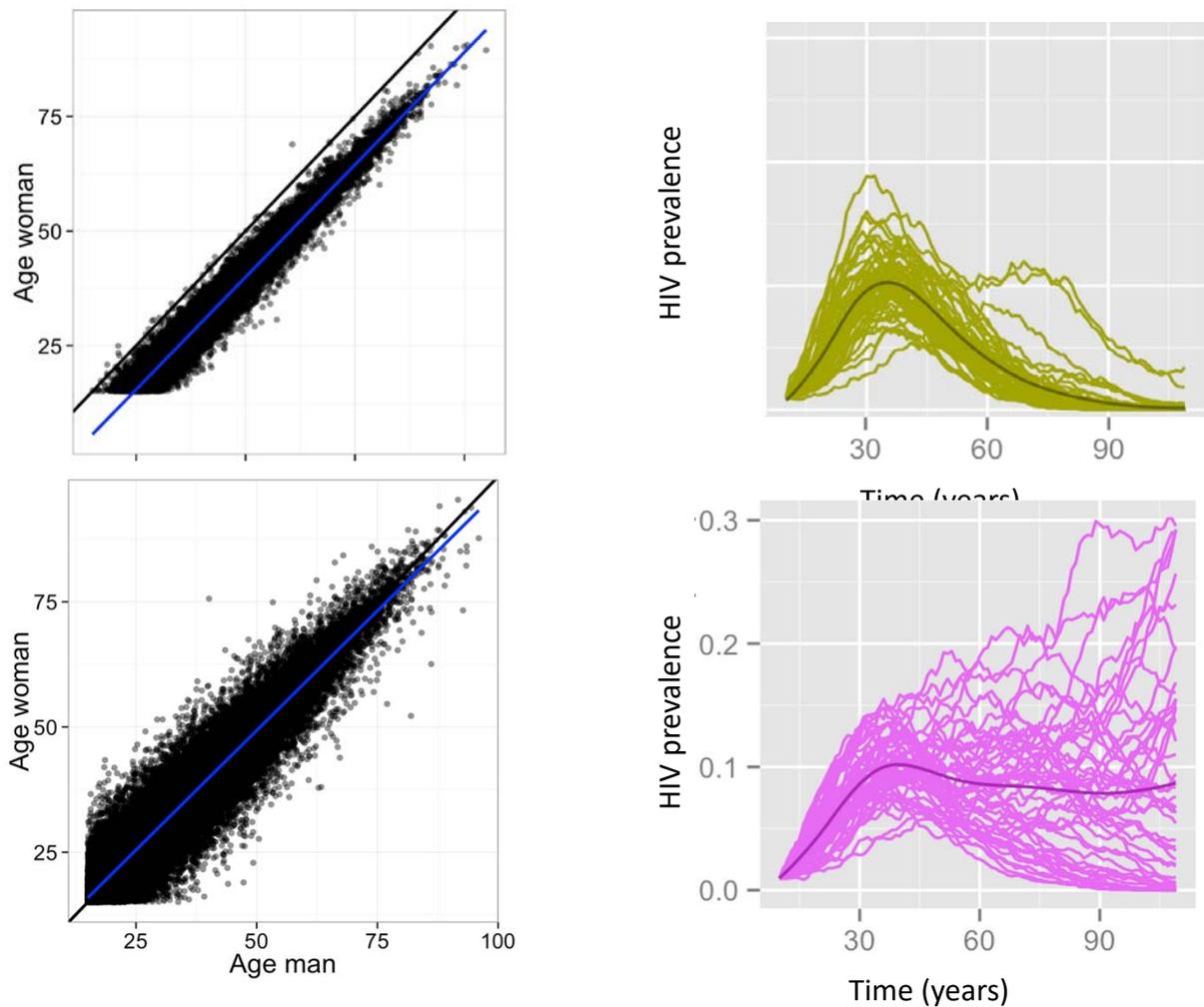


Figure 2. Two hypothetical sexual age-mixing patterns and their corresponding long-term HIV prevalence trajectories. A. Age-mixing pattern characterised by a large average age difference between sexual partners (~ 10 years) but relatively little variation in a partner age differences. B. The HIV epidemic is likely to go extinct under this age-mixing pattern. C. Age-mixing pattern characterised by a very small average partner age difference (~ zero years) but relatively large variation in partner age differences. D. HIV transmission is likely to be sustained beyond the 100 years time horizon, as the expected HIV prevalence remains high.

This issue of the SACEMA Quarterly focuses entirely on the development and use of IBMs in epidemiology. Jointly, the seven articles that make up this special issue address important questions around when and why to choose an IBM over alternative classes of models, what level of detail and complexity should be captured by the model, and whether an IBM should be realistic to be useful. Lander Willem's review paper discusses advantages and challenges of the use of IBMs for modelling chance-based transmission events among a finite set of individuals that vary widely in biological properties (e.g. their level of susceptibility or immunity to disease), social properties (e.g. their frequency of social contacts) and their behavioural response to a disease outbreak or a vaccination campaign. Leigh Johnson describes a comparison of compartmental frequency-dependent versus individual-based network models for the spread and prevention of various sexually transmitted infections (STIs). This head-to-head comparison suggests that – in the context of STI epidemics in South Africa – network models are better suited to investigate the impact of interventions on STI transmission, particularly in individuals with high numbers of sexual partners. Johnson concludes that frequency-dependent models under-estimate the importance of behaviours that increase the connections between individuals (commercial sex and concurrent partnerships) in sustaining STI transmission, while over-estimating the importance of behaviours that intensify the transmission risk between connected individuals (e.g. unprotected sex in spousal relationships). The articles by Wilbert van Panhuis, Nathan Geffen and Jennifer Lord grapple with the questions of whether and how the design of IBMs should respond to the ever-growing body of empirical data. Having more data may allow researchers to justify more complex models (i.e. more model parameters), but if these data are of poor quality or sampled from disjoint subsets of the population, then the validity of model output from such more elaborate models is not necessarily higher than that from simpler alternatives. The use of sensitivity and uncertainty analysis is suggested as a formal way of expressing to what extent the model output depends on assumptions made in the

presence of missing data and uncertainty about the values of input parameters. Additionally, structural sensitivity analysis can quantify how much the added complexity affects the conclusions drawn from the model. The article by Erin Gorsich et al. illustrates how IBMs can guide experimental data collection. By modelling hypothesised mechanisms of persistence of foot-and-mouth disease (FMD) in African buffalo in South Africa, they made explicit which aspects of the FMD transmission dynamics are insufficiently understood and require additional experiments. Elise Kuylen shows how IBMs can be used as tools for investigating the role of a particular aspect of the modelled system by comparing alternative models to counterfactual models that are purposefully deviating from what is believed to be true. Her application of the STRIDE simulator investigates the importance of transmission-relevant contacts organised in space and time according to weekdays and weekend days. She does this by contrasting a model with such an explicit contact structure against a model where the total number of weekly contacts are averaged and evenly spread over all days of the week.

In closing, I hope that this themed issue will leave you with a better understanding of the role, strengths and challenges of the use of IBMs in (infectious disease) epidemiology. For those wanting to not only read about IBMs but also get hands-on experience in conceptualising, programming, analysing and communicating IBMs, be sure to check out the announcement for the third edition of the 5-day course "Individual-based Modelling in Epidemiology: A Practical Introduction".

* In this editorial, and throughout this special issue, we are using the term individual-based model as a synonym for agent-based model.

Wim Delva - Epidemiologist, SACEMA, and Ghent University, Hasselt University and KU Leuven, Belgium. Areas of interest: statistical analysis of sexual behaviour data, stochastic and deterministic modelling of sexual network dynamics and HIV transmission. delvaw@sun.ac.za