Published: November 2017

Representing reality with individual-based models: Opportunity or illusion?

Wilbert van Panhuis - Assistant Professor of Epidemiology and Biomedical Informatics, University of Pittsburgh.

'This model is too complex, has too many parameters and assumptions; why did the authors not just use a simple statistical model to answer their question?' is common feedback on an individual-based modelling study in epidemiology. An individual-based model (IBM) of an epidemic scenario represents the mechanism of pathogen transmission between hosts in an environment and often has more parameters compared to e.g., a regression model. Both IBMs and regression models have their place in epidemiology, and they are not mutually exclusive either. Both use assumptions and need data to represent reality. Researchers often have to use a large number of patient- or individual-level observations to satisfy power requirements for e.g., measuring an association with a frequentist-based statistical model. IBMs can be created with data for a small number of system-level features, such as the pathogen transmission rate, or the fraction of the population with immunity, to test the behaviour of a system. Statistical models can fit the observed data well, sometimes even with a few parameters, but often give limited insight in the (biological or other) mechanism that generated the observed data. IBMs can represent putative mechanisms in great "realistic" detail, but may not fit the observed data very well. Which is more realistic? To what degree can IBMs represent reality; and do we need them to?

IBMs can represent detail, but are detailed real-world data available?

"What data do you need to create an individual-based model?" is a common question from people interested in developing or using an IBM. The answer, as for any scientific model, depends on the research question, i.e., the mechanism that should be represented by the IBM. To represent measles transmission, for example, the minimum data comprise information about the probability of measles transmission from one host to the next upon exposure, and the frequency and type of exposures. Exposure to measles occurs when an infectious host is within a certain proximity of a susceptible host for a certain duration of time. To represent the occurrence of exposures mechanistically, an IBM can represent host behaviour that can lead to an exposure. For example, an IBM can represent mobility and contact patterns between hosts, including contacts between infectious and susceptible hosts that can result in transmission. An IBM of a sexually transmitted disease would represent a very specific (sexual) contact pattern leading to exposures. To represent exposures even more mechanistically, an IBM can represent detailed host risk behaviour, such as decisions to stay at home during an epidemic, to seek medical care, to share needles, or to use condoms. Another possibility would be to represent environmental risk factors. For example, host mobility could be dependent on temperature, or exposure can be dependent on the presence of mosquitoes (for a mosquito-transmitted pathogen), which in turn can depend on temperature and precipitation. Since pathogen transmission requires susceptible hosts, additional mechanistic detail can be added by representing host population and immunity dynamics, e.g. by representing births, leading to new susceptible hosts, and deaths, which remove susceptible or immune hosts from the population. The amount of detail that can be represented by an IBM is almost infinite. Additional detail in an IBM will only contribute to additional realism if real-world data are available to instantiate parameters of the detailed mechanistic representation.

The good news is that data are rapidly becoming more abundant. Information about (historical) disease patterns can often be found at public health agency websites, or repositories of infectious disease data, such as Project Tycho (www.tycho.pitt.edu) at the University of Pittsburgh (1) or the International Infectious Disease Data Archive (IIDDA) at McMaster University (2). Detailed global population information is available from the WorldPop project (3) and the Integrated Public Use Microdata Series (IPUMS) database at the University of Minnesota (4). Detailed climate data are publicly available from the National Oceanic and Athmospheric Administration (NOAA) (5) and other repositories such as WorldClim (6) and PRISM (7).

Individual-based modelling can also be considered as useful method for data integration, through their ability to use many data sources to represent biological systems, such as an epidemic. As the amount and diversity of data in global health is increasing, scientists will increasingly need methods that will help them to use different data together. In case of a disease epidemic, information about the mechanism of pathogen transmission has emerged

from centuries of scientific research; the computational representation of this mechanism in an IBM can be used as a blueprint of how different datasets can be used together. For example, centuries of research has characterized the relationships between dengue transmission and host immunity, mosquito density, land use, land cover, and climate variables. In addition, research has quantified the relationship between dengue virus transmission and vaccination, vector control, and medical treatment. An IBM of dengue transmission combines all information from many different sources into a computational representation that specifies how to use different datasets together for a better understanding of dengue transmission. Similarly, creating an IBM representation of any biological system can become a method for integrating data derived from previous research, or for data emerging from new sources.

How do we know if an IBM represents reality?

If an IBM is supposed to represent an epidemic in a specific population, for example a measles epidemic in Lagos, Nigeria, then ideally, data for all parameters should be derived from this exact population, i.e. measles transmission probabilities, host risk behaviour, host mobility and contact patterns, etc. should all be empirically derived from the Lagos population. Oftentimes, such data are not available for all parameters, and sometimes, not for any parameter. In such cases, parameters can be instantiated using data derived from other populations or using made-up values based on assumptions. As the relationship between the parameter values and study population becomes less "tight", the IBM becomes potentially less realistic and the type of inference that can be made from the IBM changes. If, for example, an IBM includes parameters on the host population from Lagos, on host mobility and contact patterns from South Africa, and on measles virus probability of transmission from France, then the results hardly represent measles transmission in Lagos, and the IBM could not be used to make inference about e.g. the effect of vaccination against measles in Lagos. Some parameter values may be similar in different contexts, for example the basic reproductive rate for polio may be similar in different populations. However, unless parameter values have been empirically demonstrated to be similar in different contexts, inference based on parameter values "transposed" from a different population should not be equivalent to inference based on parameter values derived from the modelled population.

Even if all parameters are derived from the study population, how can we know that the *computational representation* of the entire system is accurate? For example, an IBM can represent the dependency of mosquito density on temperature, and of exposures on mosquito density, and of transmission on exposures, but the resulting overall dependency of transmission on temperatures may not be accurate. This is similar to a regression equation that consists of multiple variables and interactions: the association between an outcome and multiple variables may not be identical to the sum of associations between the outcome and each variable separately. To verify the accuracy of system-level outcomes estimated by the IBM, these should be compared to observed system-level outcomes in the study population, through model testing. Testing model outcomes with observed outcomes is the ultimate test of IBM realism, indicating the degree to which we understand the transmission mechanism. If model outcomes cannot replicate observed epidemic patterns, the model likely does not reflect the real world and inference about the real world should thus be limited. When an IBM represents a historical epidemic, model outcomes can be tested with historical real-world observations. Another possibility would be to use an IBM to make forecasts about future epidemic patterns. At the time of a forecast, the parameter values for a future scenario are unknown, e.g. future mobility patterns are unknown at the time of a forecast. Making accurate future forecasts is therefore a much more challenging goal compared to accurately representing a historical pattern. The development of statistical methods to test the similarity of IBM outcomes and observed data is an active field of research. Oftentimes, IBM data are aggregated into a time series of e.g. disease cases per time interval, and statistical comparisons between model and observed time series are made (8). However, methods for comparing individual-level outcomes of an IBM with individual-level data are not well defined. Individual-level model outcomes would, for example, include the location and date of infection of each host, and the statistical comparison with observed data should include some type of comparison of model vs. observed spatial-temporal infection-space.

Representing reality is not the only use of an IBM and even without any data, an IBM can give valuable insight into the type of mechanistic processes that could lead to certain system-level outcomes. For example, an IBM could represent a hypothetical epidemic in a hypothetical population in which no births occur. If infection with a pathogen in this

population leads to complete, lifelong immunity, then eventually, the epidemic will stop when no susceptible hosts remain. An alternative scenario in the same population, but now with births occurring, can demonstrate that the supply of new susceptible hosts through births can lead to continuation or recurrence of epidemics. Without data, an IBM can be used to compare epidemic scenarios in the same hypothetical population, to explore what mechanistic processes could explain certain system-level outcomes, such as the recurrence of epidemics. Such studies lead to a different type of inference compared to IBMs that aim to represent a real-world epidemic scenario. Inference based on hypothetical IBM scenarios should be limited to the possible mechanistic explanations of observed system-level phenomena, but cannot be used to make inference about the expected course of specific real-world epidemics or the expected effects of interventions on such epidemics.

In conclusion, IBMs can be very useful to refine our understanding of the mechanism of pathogen transmission and can help make inference about real-world epidemics, if based on sufficient data. IBMs should not be discarded simply because they are complicated, have many parameters, or use assumptions. Every scientific model, including our own *"mental models"* use assumptions, albeit not always fully specified or declared (9). The type of inference made from an IBM should be closely tied to the data used to parameterize it. We have no standard format or widely used standard representation of model structure, parameter values, and assumptions, but using such a representation would make it much easier to interpret published model results (10). Some efforts are ongoing to develop standard, machine-interpretable representations of epidemic models, such as the Apollo project at the University of Pittsburgh (11). Increasingly, software, training, computational resources, and data will enable the use of IBMs in global health, adding a great asset to the repertoire of methods and tools to improve health for people around the world.

Wilbert van Panhuis *- Assistant Professor of Epidemiology and Biomedical Informatics, University of Pittsburgh. Research interests: I aim to improve the use of data to design better disease control strategies, such as vaccination programs and strategies against vector-borne diseases. I also use informatics methods to improve access to standardized global health data and direct Project Tycho. Email: wilbert.van.panhuis@pitt.edu*

References:

1. van Panhuis WG, Grefenstette J, Jung SY, Chok NS, Cross A, Eng H, et al. Contagious diseases in the United States from 1888 to the present. N Engl J Med. 2013;369: 2152–8.
2. Earn D, Dushoff J. International Infectious Disease Data Archive [IIDDA] [Internet]. [cited 1 Jan 2017]. [http://lalashan.mcmaster.ca/theobio/IIDDA/index.php/Main_Page] Accessed 23 November 2017/
3. Tatem AJ. WorldPop, open data for spatial demography. Sci Data. Nature Publishing Group; 2017;4: 170004.
4. University of Minnesota Population Center. Integrated Public Use Microdata Series, International [Internet]. 2015 [cited 25 Mar 2015]. [https://international.ipums.org/international/] Accessed 23 November 2017.
5. National Oceanic and Atmospheric Administration. National Oceanic and Atmospheric Administration [Internet]. NOAA. [http://www.noaa.gov/] Accessed 23 November 2017.
6. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol. 2005;25: 1965–1978.
7. Northwest Alliance for Computational Science and Engineering. PRISM Climate Group [Internet]. 2017. [http://www.prism.oregonstate.edu/] Accessed 23 November 2017.
8. Andraud M, Hens N, Marais C, Beutels P. Dynamic epidemiological models for dengue transmission: a systematic review of structural approaches. PLoS One. 2012;7: e49085.
9. Johnson-Laird PN. Mental models and human reasoning. Proc Natl Acad Sci U S A. National Academy of Sciences; 2010;107: 18243–50.
10. Willem L, Verelst F, Bilcke J, Hens N, Beutels P. Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006-2015). BMC Infect Dis. BioMed Central; 2017;17: 612.
11. Wagner MM, Levander JD, Brown S, Hogan WR, Millett N, Hanna J. Apollo: giving application developers a single point of access to public health models using structured vocabularies and Web services. AMIA Annu Symp Proc. 2013;2013: 1415–24.