Eur J Epidemiol (2010) 25:13–19
DOI 10.1007/s10654-009-9397-9

METHODS

Mathematical modeling and the epidemiological research process

Mikayla C. Chubb · Kathryn H. Jacobsen

Received: 15 May 2009 / Accepted: 7 October 2009 / Published online: 27 October 2009

Abstract The authors of this paper advocate for the expanded use of mathematical models in epidemiology and provide an overview of the principles of mathematical modeling. Mathematical models can be used throughout the epidemiological research process. Initially they may help to refine study questions by visually expressing complex systems, directing literature searches, and identifying sensitive variables. In the study design phase, models can be used to test sampling strategies, to estimate sample size and power, and to predict outcomes for studies impractical due to time or ethical considerations. Once data are collected, models can assist in the interpretation of results, the exploration of causal pathways, and the combined analysis of data from multiple sources. Finally, models are commonly used in the process of applying research findings to public health practice by estimating population risk, predicting the effects of interventions, and contributing to the evaluation of ongoing programs. Mathematical modeling has the potential to make significant contributions to the field of epidemiology by enhancing the research process, serving as a tool for communicating findings to policymakers, and fostering interdisciplinary collaboration.

Keywords Mathematical model · Epidemiology · Susceptible-infectious-removed (SIR) model

M. C. Chubb · K. H. Jacobsen (✉)
Department of Global and Community Health, George Mason University, 4400 University Drive MS 5B7, Fairfax, VA 22030-4444, USA
e-mail: kjacobse@gmu.edu

Introduction

Many epidemiologists may think that statistical regression is the only modeling technique available for the epidemiologist's toolkit, but statistical models are only one of several types of analytic models that are valuable to the discipline. Spatial models make use of geographic information systems; ecological models explain population dynamics; physiological models describe cellular functions; and forecasting, simulation, and cost-benefit analyses enhance public policy decision-making. In each of these cases, models further our understanding of how social, biological, and environmental processes impact health and disease in populations.

Mathematical modeling is a set of techniques, tools, and equations that can be tailored to particular disciplines. In epidemiology, mathematical models usually define interactions between individuals or populations and other individuals, populations, or environments. By defining the rules that describe these interactions and translating those rules into equations, a complex set of processes can be broken down into components and quantified. The model can then be used to explore relationships in the modeled population, to test the impact of changed rules on the system and its components, and to examine the outcomes of various events that might have an effect on a population.

Despite these many potential uses, mathematical models are, at present, used infrequently by epidemiologists. However, modeling has already made significant contributions to the health sciences (including both clinical medicine and public health) and related disciplines, including biology, mathematics, statistics, bioinformatics, and other fields [1]. A summary of some of these areas of research is highlighted in Fig. 1. An increased familiarity with the many ways that mathematical models can be used

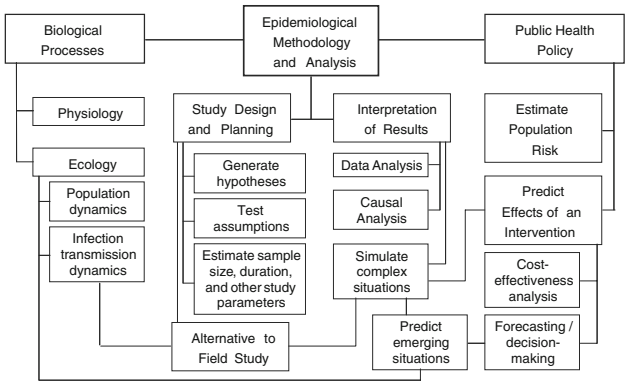

Fig. 1 Examples of uses of models in epidemiology

in epidemiological research will allow models to be used more extensively and correctly, to be accessible to a broader range of epidemiologists, and to receive more critical examination. This paper describes the great variety of uses for mathematical models within the field of epidemiology and provides an overview of the methods of modeling.

The epidemiological research process can be considered to have four key steps (Fig. 2): (1) identifying study questions, (2) designing studies and collecting data, (3) analyzing data, and (4) applying research findings to public health. The mathematical modeling process follows four corresponding steps: (1) selecting key components for the model, (2) identifying and validating the inputs that will go into the model, (3) running the model, and (4) interpreting outputs and explaining the applications of the model results. In the following four sections, we describe the applications of models to epidemiology and introduce some of the principles and techniques of modeling. Susceptible-infectious-removed (SIR) models, commonly used in infectious disease epidemiology to describe infection transmission dynamics, are used as a primary example, and additional model-based studies from many epidemiological disciplines are provided as supplemental illustrations.

Identifying study questions

The first step in any research project is to identify the questions that will be explored. For new studies, this may involve conducting a community needs assessment. For ongoing projects, this may take the form of a program evaluation in which several possible next steps that could be implemented are evaluated. For all studies, this step typically involves consulting the existing literature to identify what topics have previously been explored and to catalogue the gaps that remain to be filled. Some researchers find it helpful to create a simple sketch of the populations of interest, the exposures that will be examined, the relationships between these populations and/or exposures, and possible causal pathways for disease processes. This type of visual expression of what is and is not understood about a complex system can be a first step toward building a mathematical model.

For example, infection transmission dynamics can be represented using an SIR model such as the one shown in Fig. 3. SIR models are among the most commonly used

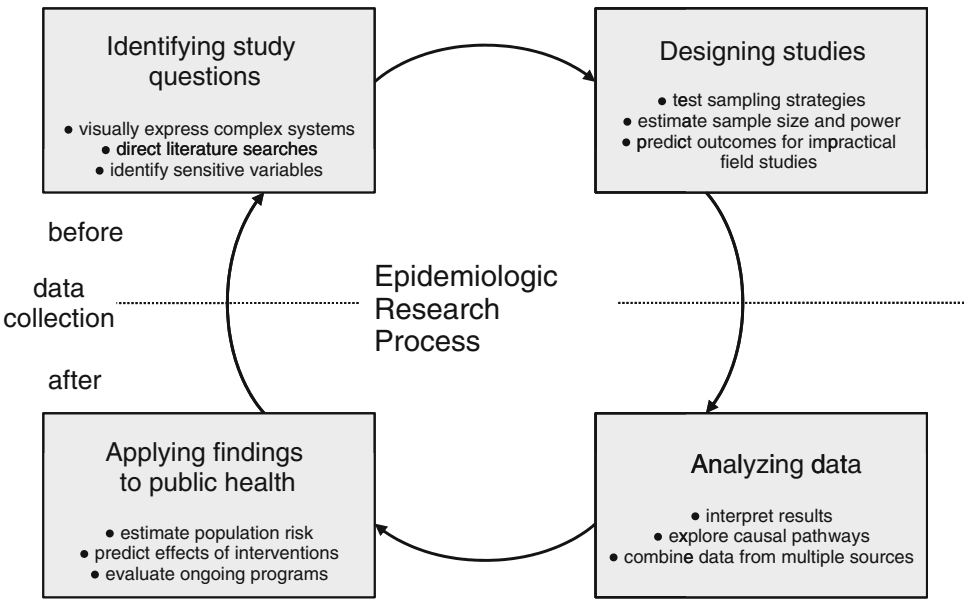

Fig. 2 Framework for the application of mathematical models throughout the epidemiological research process

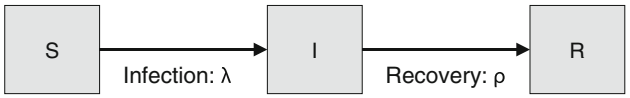

Fig. 3 Sample SIR model

models in epidemiology, and serve as a good introduction to the modeling process. In this simple model, every individual in a population is assigned to one of the three compartments: S (susceptible) for individuals at risk of infection, I (infected/infectious) for individuals who are currently infected, or R (recovered/removed) for individuals who have recovered from the infection and have immunity.

Realism can be added to the model by making it more complex (Fig. 4). If an SIR model describes changes that would be expected to occur in a population over decades or longer time periods, realism can be added by building population dynamics into the model so that susceptible individuals are "born" into the population and older adults "die" and are removed from the population. (One of the advantages of modeling is the ability to "observe" several generations' worth of data in mere minutes.) The death rate might be higher for individuals in box I, and that increased risk could be represented by an extra arrow out of the box for death due to infection. Other arrows, which represent the flow of individuals from one compartmental classification to another or flow into a population due to birth or out of a population due to death, could represent public health interventions, such as a new vaccine allowing individuals to move directly from the S box to the R box.

Additional compartments could be added to represent age groups, sex groups, income groups, or other exposure categories. For example, an SIR model with two age groups, child and adult, would need six compartments: S, I, and R boxes for children and S, I, and R boxes for adults. More complex models might include separate boxes for males and females, genetic characteristics, behavioral risk factors, or other exposures, and might require hundreds of compartments. If the infection being studied is vector-borne, the model can incorporate information about the insect vector, the life cycle of the pathogen, and human behavior and biology [2, 3]. If the disease of interest is a chronic condition, a model can incorporate information about the progression of disease [4]. Multiple exposures and multiple outcomes can be included in a model. For example, a model of the impacts of various aspects of traffic—such as traffic volume, traffic speed, the presence of safe walking and bicycling zones, and amount of vehicle emissions—can investigate a variety of health outcomes, including respiratory and cardiovascular health, osteoporosis, mental health, and injuries [5].

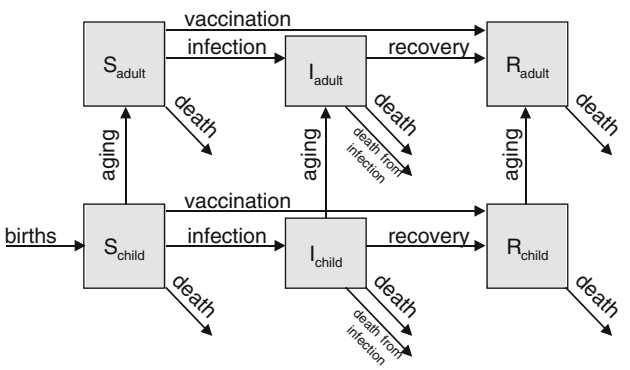

Fig. 4 More complex SIR model

Selecting the components that will be included in a model requires seeking a balance between simplicity and complexity. A model that is too simple—one that ignores critical components or relationships—will not clarify what happens in the real world. A model that is too complex is likely to be inaccurate due to the impossibility or impracticality of acquiring sufficiently detailed input data for a large number of compartments and parameters.

Whether or not a researcher intends to build and test a formal mathematical model, this kind of sketch—and this sort of "systems thinking" [6]—can clarify the relationships that the researcher wants to explore. This, in turn, may contribute to the framing of study questions, the assessment of the components and interactions that may influence a system, and the selection of variables that will be measured.

Designing studies

The second step in the research process is to design studies that will collect valid data. Models can contribute to the planning of a field study by assisting in the selection of a sampling strategy—for example, models may identify certain population groups that should be preferentially recruited based on demographic characteristics or exposure history—and in the estimation of the required sample size and study duration. Models also help to clarify the assumptions that are built into a study's design, such as the assumption that infection confers long-term immunity or the assumption that patients with chronic diseases are fully compliant with treatment regimens. There is a feedback loop between field studies and models: data from field studies are used to create models that represent the real world, and models provide information about how to best measure real-life variables.

To understand this symbiotic relationship, one must understand the process for identifying and validating the inputs that will go into the model. For SIR models, this step involves quantifying the proportion of the population located in each compartment in a model and assigning

values to the rates of flow between compartments. The arrows on Figs. 3 and 4 show the flow of individuals (or the flow of a proportion of the population in each compartment) from S to I and from I to R over time. These rates are often represented using Greek letters, such as lambda (λ) for the infection rate and rho (ρ) for the recovery rate. When possible, the values for parameters like λ and ρ are estimated from real-life data, although sometimes they must be merely educated guesses. Clearly explaining the source of these values and how closely they estimate real-life values is an important part of justifying the validity and applicability of a model.

If constant values are assigned to all parameters of an SIR model, then it is said to be deterministic and the exact same output will result every time the model is run (assuming that the same model structure, equations, and parameter values are used). If a probability distribution is assigned to these parameters to better capture the uncertainty of the estimate, then the model is said to be stochastic, which means that the output will vary each time the model is run. Stochastic models are usually run thousands of times so that the probability distribution of outputs can be examined.

A key step in testing a model is sensitivity analysis, which determines how much each model component contributes to the output of the whole model. This process is an important contribution of modeling to the design of valid field studies. Parameters that are highly sensitive strongly influence the outcomes of the model, and, presumably, real-life outcomes. Sensitive variables must be measured very carefully. Other parameters may have almost no impact on the outcome of a model. These variables may not even need to be included in data collection, although it is important to err on the side of caution when using models to identify key variables, since the results of sensitivity testing are dependent on the model structure and assumptions, the definition of parameters, and existing data. Still, the identification of parameters that appear to be highly sensitive can be critical to the development of valid data collection procedures.

Once the framework for study design and sampling has been developed, models can be used to explore the balance between sample size and statistical power or to determine whether the proposed study can be completed within budget and time constraints. In some cases, models may show that a field study is unlikely to produce meaningful results. For example, it may not be practical to conduct a study if a large population with special attributes is required and known to be unavailable, or if the most important variables are difficult or impossible to measure accurately and reliably. In situations in which a model shows that a field study is impractical, mathematical models may be able to replace field studies by generating output based on information that is already available. Replacements for field studies may be essential when time constraints or ethical concerns prevent a trial from being conducted. For example, a model of foodborne disease outbreaks based on past outbreaks can be used to estimate the effects of changes in human and pathogen behavior on population health rather than waiting to see what the outcomes of a particular emergent threat are before updating policy recommendations [7].

Analyzing data

Once a field study has been implemented and data have been collected, the third stage of the research process is data analysis. Whether or not data were collected with a mathematical model in mind, a model can be created or modified for use in interpretation of results and for causal analysis. After assigning values to compartments and parameters, a model can be run—the equations solved, usually as a function of time—and the outcome variables can be displayed visually on a computer monitor, usually in the form of graphs.

A first step in analyzing data with mathematical models is to use model representations to simplify complex data sets into manageable relationships and pathways to explore. For example, in an SIR model equations are used to define the change in the number of people (or the proportion of the total population) in each compartment during a certain time period. In the model shown in Fig. 3, there is only one arrow leaving the S box. The equation for the change in this compartment over time is written as $dS/dt = -\lambda S$, which says that the change in S per one unit of time is to lose individuals from the S box at a rate of λ. The equations for the other compartments are $dI/dt = +\lambda S - \rho I$ and $dR/dt = +\rho I$. Individuals who leave box S ($-\lambda S$) enter box I ($+\lambda S$), and individuals who leave box I ($-\rho I$) enter box R ($+\rho I$). More complex models (such as the one in Fig. 4) may require more complex equations. For example, if the infection rate is found to be related to the proportion of infectious individuals in the population at a given time, which is $I/(S+I+R)$, it would be more accurate to have the equation for the flow out of the S box to be $dS/dt = -\lambda S(I/(S+I+R))$.

It is also possible to define how different types of individuals relate to one another using structured (or preferential) mixing equations that describe how certain individuals or populations interact with one another. For example, these equations might specify that a child is more likely to have contact with another child than with an adult or that individuals who engage in high-risk behavior (such as unprotected sexual intercourse) are more likely to engage in risky behavior with other high-risk individuals

than with low-risk individuals. An even more complex model may require moving beyond compartments to individual-based simulation in which each simulated individual in the population has a personal history and a special set of rules that define how that individual interacts with other individuals and with the environment. Data can be used to refine the model components and pathways and to fill in population counts and rates of interaction. Conversely, models can be used with existing data to fill in gaps in knowledge.

Models also allow for fuller use of existing data and the concurrent analysis of data from a variety of sources. For example, models have been used to estimate the incidence of hepatitis A virus infections based on seroprevalence data from more than one hundred field studies from around the world [8] and to analyze HIV transmission dynamics in populations of injecting drug users by combining surveillance information with testing of needles used in exchange programs [9]. Other models have compiled information about pathology, immunology, and epidemiology into one model of the causes of influenza outbreaks [10]; combined bacteriological, pharmacological, and treatment information into an analysis of antibiotic resistance risks in hospitals [11]; and incorporated longitudinal data on household socioeconomic status and family violence into a model of mental health [12]. As models in epidemiology and other fields refer to and refine each other, data collected by epidemiologists becomes even more valuable for understanding population health and predicting changes in public health status.

Applying findings to public health

A typical final step in the epidemiologic research process is to identify the lessons learned from a study, which often takes the form of suggesting possible public health interventions based on the results of a field study, proposing appropriate policy measures to address public health concerns, or recommending future areas of research. Models can contribute to all three of these functions. Some of the first models in epidemiology were developed in 1760 by Bernoulli in order to promote the benefits of smallpox vaccination [13], and applied models remain popular today as tools for persuasion and enhanced decision-making.

Models can be useful to both scientists and policymakers, and are helpful for demonstrating the value of public health programs to stakeholders. For example, a model of HPV vaccination in Finland that compared the effectiveness of vaccinating different populations at different ages determined that programs targeting females alone were almost as effective as programs for both sexes [14], while an evaluation of a possible HPV vaccination program for 12-year-old girls in the United States determined that the proposed program would have somewhat higher cost than existing childhood vaccination programs but would provide a similarly high benefit [15, 16]. Other studies have used surveillance data to predict the effects of policies or programs on the incidence and prevalence of other sexually transmitted infections in the general population [17–19]; the SimSmoke simulation model uses assessments of the impact of past tobacco control policies to predict the impact of new policies on smoking prevalence in the future [20]; the BOLD model feeds data collected under rigorous standards at sites around the world into a model of the burden of chronic obstructive pulmonary diseases [21]; DISMOD II, a program available through the World Health Organization, checks the internal validity of burden of disease estimates [22]; the Prevent model examines the impact of risk factors on chronic disease [23]; and studies exploring the best ways to allocate health resources have, for example, examined the relative impacts of resources used for preventing the onset of chronic diseases versus preventing the complications of existing cases [24]. Models have also been used to identify high-risk populations, and to predict the impact of demographic shifts or behavioral changes on disease incidence and prevalence.

Models using data collected during a study or intervention can inspire further related interventions, trigger investigation of outcomes that are not understood, lead to changes in an intervention effort as it progresses, and provide reassurance that intervention programs are on track for success. For example, models were used to improve the Onchocerciasis Control Program in West Africa midstream. The effectiveness of the expensive, large-scale program to reduce the black fly vector that transmitted the parasite that causes onchocerciasis (also known as river blindness) was questioned when nearly a decade into the program there was little change in the prevalence of onchocerciasis in the treatment area. A model of the decreasing intensity of infections based on data collected during the intervention showed that continuing the program could lead to the elimination of onchocerciasis from the study area in just five additional years. The model proved to be correct [25], and the OnchoSim program is now being used for surveillance and planning in other regions of Africa [26].

Other studies have compared the projected health impacts of various types of interventions based on collected data. For example, a dynamic population model used to explore the relative outcomes of various types of smoking cessation interventions found that minimal counseling by a physician was the most cost-effective way to reduce tobacco use, but it was responsible for only a small portion of those who quit smoking; intensive counseling plus use of a pharmaceutical smoking cessation aid was

more expensive, but was significantly more effective [27]. These applications of study results to public health show that the research process rarely ends at this stage, but instead flows naturally back to the identification of new questions to explore.

Conclusion

In the cycle of epidemiological research, mathematical models can provide many benefits, such as simplifying and presenting complex information, evaluating the significance of variables, performing additional analysis on data, and forecasting outcomes for a project or population (Fig. 2). The publication of epidemiological models can be of great benefit to the epidemiological community when researchers describe their frameworks, assumptions, analyses, and interpretations in clear and quantifiable terms.

At present, one of the main challenges to the expanded use of mathematical models in epidemiology is the limited pool of epidemiologists with the advanced mathematical training required to design and conduct high-level analysis. This impediment can be substantially alleviated by expanding collaborative research with experts in related disciplines, such as computer science, mathematics, bioinformatics, geography, and engineering. Epidemiology will benefit from more and broader collaborations, and interdisciplinary work will contribute to the development and application of both new tools and novel uses for existing analytic techniques. A related concern is the need for epidemiological modelers to clearly explain both the outcomes and the limitations of their work to the public, to politicians, and to public health professionals. As the number of epidemiologists comfortable with the use and interpretation of models grows, the number of researchers able to effectively communicate this information will also increase. This will enable researchers to make even fuller use of mathematical models during all stages of the epidemiologic research process.

References

1. Temime L, Hejblum G, Setbon M, Valleron AJ. The rising impact of mathematical modelling in epidemiology: antibiotic resistance research as a case study. Epidemiol Infect. 2008;136:289–98.
2. McKenzie FE, Samba EM. The role of mathematical modeling in evidence-based malaria control. Am J Trop Med Hyg. 2004;71:94–6.
3. Michael E, Malecela-Lazaro MN, Kabali C, Snow LC, Kazura JW. Mathematical models and lymphatic filariasis control: endpoints and optimal interventions. Trends Parasitol. 2006;22:226–33.
4. Shih HC, Chou P, Liu CM, Tung TH. Estimation of progression of multi-state chronic disease using the Markov model and prevalence pool concept. BMC Med Inform Decis Mak. 2007;7:34.
5. Joffe M, Mindell J. Complex causal process diagrams for analyzing the health impacts of policy interventions. Am J Public Health. 2006;96:473–9.
6. Leischow SJ, Best A, Trochim WM, Clark PI, Gallagher RS, Marcus SE, et al. Systems thinking to improve the public's health. Am J Prev Med. 2008;35:S196–203.
7. Cummins EJ. The role of quantitative risk assessment in the management of foodborne biological hazards. Int J Risk Assess Manag. 2008;8:318–30.
8. Jacobsen KH, Koopman JS. The effects of socioeconomic development on worldwide hepatitis A virus seroprevalence patterns. Int J Epidemiol. 2005;34:600–9.
9. Kretzschmar M, Wiessing L. New challenges for mathematical and statistical modeling of HIV and hepatitis C virus in injecting drug users. AIDS. 2008;22:1527–37.
10. Moghadas SM. Gaining insights into human viral diseases through mathematics. Eur J Epid. 2006;21:337–42.
11. Grundmann H, Hellriegel B. Mathematical modelling: a tool for hospital infection control. Lancet Infect Dis. 2006;6:39–45.
12. Banyard VL, Williams LM, Saunders BE, Fitzgerald MM. The complexity of trauma types in the lives of women in families referred for family violence: multiple mediators of mental health. Am J Orthopsychiatry. 2008;78:394–404.
13. Blower S, Bernoulli D. An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. Rev Med Viro. 2004;14:275–88.
14. French KM, Barnabas RV, Lehtinen M, Kontula O, Pukkala E, Dillner J, et al. Strategies for the introduction of human papillomavirus vaccination: modelling the optimum age- and sex-specific pattern of vaccination in Finland. Br J Cancer. 2007;96:514–8.
15. Sanders GD, Taira AV. Cost effectiveness of a potential vaccine for human papillomavirus. Emerg Infect Dis. 2003;9:37–48.
16. Dasbach EJ, Elbasha EH, Insinga RP. Mathematical models for predicting the epidemiologic and economic impact of vaccination against Human Papilloma infection and disease. Epidemiol Rev. 2006;28:88–100.
17. Garnett GP. An introduction to mathematical models in sexually transmitted disease epidemiology. Sex Transm Inf. 2002;78:7–12.
18. Grassly NC, Fraser C. Mathematical models of infectious disease transmission. Nat Rev Microbiol. 2008;6:477–87.
19. Mills S, Saidel T, Magnani R, Brown T. Surveillance and modelling of HIV, STI and risk behaviors in concentrated HIV epidemics. Sex Transm Inf. 2004;80:ii57–62.
20. Levy DT, Bauer JE, Lee H. Simulation modeling and tobacco control: creating more robust public health policies. Am J Public Health. 2006;96:494–8.
21. Buist AS, Vollmer WM, Sullivan SD, Weiss KB, Lee TA, Menezes AM, et al. The burden of Obstructive Lung Disease Initiative (BOLD): rationale and design. COPD. 2005;2:277–83.
22. Barendregt JJ, van Oortmarrsen GJ, Vos T, Murray CJL. A generic model for the assessment of disease epidemiology: the computational basis of DisMod II. Popul Health Metr. 2003;1:4.
23. Brønnum-Hansen H. How good is the Prevent model for estimating the health benefits of prevention? J Epidemiol Community Health. 1999;53:300–5.
24. Homer JB, Hirsch GB. System dynamics modeling for public health: background and opportunities. Am J Public Health. 2006;96:452–8.
25. Remme JHF. Research for control: the onchocerciasis experience. Trop Med Int Health. 2004;9:243–54.

26. Alley WS, van Oortmarssen GJ, Boatin BA, Nagelkerke NJD, Plaisier AP, Remme JHF, et al. Macrofilaricides and onchocerciasis control, mathematical modeling of the prospects for elimination. BMC Public Health. 2001;1:12.

27. Feenstra TL, Hamberg-van Reenen HH, Hoogenveen RT, Rutten-van Mölken MP. Cost-effectiveness of face-to-face smoking cessation interventions: a dynamic modeling study. Value Health. 2005;8:178–90.