American Journal of Epidemiology
© The Author(s) 2018. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Vol. 188, No. 5
DOI: 10.1093/aje/kwy262
Advance Access publication:
March 16, 2019

Commentary

Systems Modeling to Advance the Promise of Data Science in Epidemiology

Magdalena Cerdá* and Katherine M. Keyes

* Correspondence to Dr. Magdalena Cerdá, Department of Population Health, New York University School of Medicine, 180 Madison Avenue, New York, NY 10016 (e-mail: Magdalena.Cerda@nyulangone.org).

Initially submitted October 10, 2018; accepted for publication November 14, 2018.

Systems science models use computer-based algorithms to model dynamic interactions between study units within and across levels and are characterized by nonlinear and feedback processes. They are particularly valuable approaches that complement the traditional epidemiologic toolbox in cases in which real data are not available and in cases in which traditional epidemiologic methods are limited by issues such as interference, spatial dependence, and dynamic feedback processes. In this commentary, we propose 2 key contributions that systems models can make to epidemiology: 1) the ability to test assumptions about underlying mechanisms that give rise to population distributions of disease; and 2) help in identifying the types of interventions that have the greatest potential to reduce population rates of disease in the future or in new sites where they have not yet been implemented. We discuss central challenges in the application of systems science approaches in epidemiology, propose potential solutions, and predict future developments in the role that systems science can play in epidemiology.

agent-based models; complex systems; public health; systems science

Abbreviations: ABMs, agent-based models; HIV, human immunodeficiency virus.

Systems science focuses on the study of complex adaptive systems and includes methodological approaches, such as systems dynamics models (1), network models (2, 3), and agent-based models (ABMs) (4). These approaches are often characterized by nonlinear relations and feedback processes, such that small changes can produce emergent properties that are not entirely predictable by individual components. From a methodological perspective, complex systems approaches use computer-based algorithms to model dynamic interactions between units, or agents, within and across multiple levels. These models have been extensively developed for modeling infectious disease transmission (5–8) and are increasingly used in chronic disease and injury epidemiology (9–15).

HOW DOES SYSTEMS SCIENCE FIT IN WITHIN THE EPIDEMIOLOGY MACHINERY?

The value of systems science models in epidemiology lies in their ability to complement traditional epidemiologic tools when real data are not available or able to answer epidemiologic questions and to answer questions that are ill-suited to traditional epidemiologic methods (6, 16–21).

First, systems science approaches can be used to test assumptions about underlying mechanisms and feedback processes that give rise to the population distribution of complex health behaviors and disease. By assigning a set of behavioral and transition probabilities, we can examine the types of conditions that produce observed disease patterns and provide quantitative bounds around the plausibility of mechanisms that explain these patterns. An understanding of these mechanisms can provide insights into why past interventions may have succeeded or failed. For example, to explain trends in prescription opioid use, misuse, and overdose, Wakeland et al. (22) modeled a complex system in which individuals initiated nonmedical use, transitioned to paying for opioids and tampering with them, and then further transitioned to heroin use. The extent to which drug misuse trajectories were influenced by the availability, accessibility, and physical properties of the prescription opioid supply was tested based on the underlying simulated dynamics that gave rise to historical trends in use and abuse of nonmedical prescription opioids. This simulation of mechanisms suggested that interventions to reduce informal sharing of opioids could reduce opioid abuse to a greater extent than could creating tamper-resistant prescription opioid formulations.

Second, we can use systems science approaches to identify the types of interventions that have the greatest potential to reduce population rates of disease, either in the future or in new sites where they have not yet been implemented. Systems science models can complement evaluations of existing interventions, integrating counterfactual questions with modeling strategies that re-weight effect estimates based on concerns about transportability (23, 24). In this case, systems models can answer research questions about interdependent causal effects that are bedeviled by concerns about feedback loops, interference, and spatial dependence (25). For example, an ABM was used to describe transmission of human immunodeficiency virus (HIV) in a dynamic network of people who injected drugs, those who used non-injection drugs, and those who did not use drugs (26). By simulating a series of hypothetical intervention scenarios, the authors found that a significant scale-up of multiple prevention programs would be necessary to reduce rates of new HIV infections in people who inject drugs over the next decades.

In addition, systems science models such as ABMs that embed systems dynamics and network models can answer questions about the impact of interventions given selection into social networks or spatial contexts, as well as placement of interventions. For example, Keane et al. (27) predicted what “could” happen to opioid overdose rates if they varied the number and spatial location of naloxone distribution sites and the number of naloxone kits provided in a community, given different assumptions about secondary distribution of naloxone within social networks and the local dynamics of substance use.

CURRENT CENTRAL CHALLENGES IN THE USE OF SYSTEMS SCIENCE APPROACHES

Advances in computational machinery and in the use of “big data” for public health have made it possible to answer increasingly sophisticated questions with systems science, yet challenges abound. First, systems science models make strong assumptions about behavioral and disease probabilities, the structure and function of networks, and the dynamics of disease. Hence, model calibration and validation remain a central challenge, because such data are often missing and/or confounded; further, calibration often requires putting together data elements from multiple, disparate data sets. This issue has become especially relevant as there has been a push toward models that replicate a “real-world setting” and rely on multiple assumptions about individuals, their interactions with each other, and their interactions with their environment. Second, concerns exist about the risk for bias, especially with increasingly complex models (28). Because systems models often treat parameters of intervention effects obtained from one population as causal effects in a different population, biases arise because of lack of transportability when 2 populations have different underlying risks for the study outcome or different distributions of unmeasured confounding (29). Further, treating parameters of past intervention effects obtained from one population as causal effects in a different population gives rise to the potential for collider bias in the context of time-dependent confounding (28, 29). Third, the central challenge related to the issues discussed above lies in validation. In the absence of data to calibrate model parameters, estimation of parameters is often carried out by altering parameter values until model summary measures match those of the observed data. Unfortunately, there are multiple sets of parameters that can generate similar model summary measures, making it difficult to conclude whether the model assumptions are valid.

HOW CAN WE OVERCOME THESE CHALLENGES?

Triangulation is critical to address concerns about calibration, bias, and validation in systems science models for public health. First, we need to invest in iterative efforts that combine primary data collection, systems modeling, and implementation of interventions (30). Specification of a systems dynamic, network- or agent-based model forces us to be precise about the source of each model parameter and in that way provides insight into key gaps in data that need to be collected. An ideal approach would involve modeling, identifying gaps in key parameters, investing in primary data collection to fill such gaps, and re-parameterization of a better-calibrated model. In addition, in cases in which systems modeling is used to predict the impact of a specific intervention, such models should form part of a larger evaluation effort in which initial predictions about intervention effects estimated through systems models can be tested through the implementation and evaluation of actual interventions. In turn, this could inform a more well-calibrated and valid model that could be used to answer new questions about the mechanisms through which such interventions might work and their impact on other contexts and time frames.

Second, we can incorporate epidemiologic causal inference methods to address sources of bias in the estimation of systems science models. Some have suggested that methods such as parametric g-computation can be used instead of systems science approaches to examine complex systems (28, 29), as they can handle interdependent causal effects and feedback loops but rely on a single data set to calibrate simulations and are more robust to threats of time-dependent confounding and transportability violations (29). Although such methods are quite valuable, they cannot be used, in isolation, to answer precisely the types of questions that systems science models are best suited to answer, particularly questions that require explicit modeling of the role that space and social networks could play in shaping the impact of new interventions. Yet, they can play a key role in strengthening the validity of systems science models. For example, an ABM could use parametric g-computation to estimate specific causal effects of well-defined interventions within a broader system, adding layers to the model to test how processes of social interaction, interference, and space modify the impact of the intervention. Indeed, g-computation methods excel at identifying causal effects, whereas ABMs require causal effects for valid estimation and can export such effects to illuminate the bounds of potential future intervention and prevention effects. Both are necessary and complementary rather than alternatives (31). Further, we could draw from existing efforts in epidemiology and statistics to develop transportability estimates, devising new ways to incorporate such estimates into systems science models in order to predict the potential impact of new interventions in new contexts (32).

Third, all systems science modeling efforts require careful calibration and validation efforts. This includes standard approaches such as conducting period-by-period calibration-test-recalibration processes, including testing whether the model can forecast study outcomes at a later time point for which there are available data; iteratively excluding study units and testing whether the model can accurately predict the outcome of interventions adopted in the excluded units; and testing how robust the model is to variations in key assumptions about model dynamics (33). Uncertainty analysis approaches, such as Latin hypercube sampling of the space of plausible parameter values, can be used to efficiently test the sensitivity of results to the choice of multiple input parameters at the same time without having to consider all the possible permutations (34). Finally, methods have been applied in infectious disease epidemiology to strengthen the reliability and validity of predictions made from systems science models. These methods include the use of Bayesian Markov chain Monte Carlo and approximate Bayesian computation methods to estimate parameters (8). These methods allow us to conduct statistical inference by searching parameter space to obtain maximum likelihood estimates and 95% credible intervals for parameters of interest (35) and to use deviance information criteria to assist in model selection (36, 37). Because systems science approaches are increasingly used in fields of epidemiology beyond infectious disease, uncertainty analyses and empirical validation methods such as these need to form an integral part of the modeling process.

A VISION FOR THE FUTURE

We predict that 3 key developments will shape the future of systems science modeling in epidemiology. First, the quality and complexity of systems science models will likely advance at an exponential pace because of the acceleration of big data (38) initiatives that provide access to detailed data on individuals, their health histories, and where they live, work, and attend school. One example is the Framework for Reconstructing Epidemiological Dynamics, an open-source agent-based modeling simulation platform (39) that uses census-based synthetic populations that are statistically equivalent to the population in any state or county (40, 41), and that simulates age-appropriate daily activities and interactions of millions of people. Increasingly sophisticated approaches to combine highly detailed location data with health histories (e.g., electronic health records, Medicaid/Medicare data) in user-friendly platforms will make systems modeling approaches more accessible to epidemiologists and will provide the means to base models on higher-quality inputs. Second, machine-learning approaches that incorporate high-dimensional sets of covariates and complex interactions for individual and place-based risk prediction could substantially improve the predictive power of systems science model inputs (42), especially as learning algorithms are increasingly sensitive to real-time data inputs for continuous adaptive learning. They can be used for risk prediction, to account for complex sets of confounders and effect modifiers (42), to predict future hotspots of disease risk (43), and to predict transportability of interventions across sites (32). We envision a scenario where constantly updated sets of big data would inform machine-learning algorithms used for individual and place-based risk prediction, which would in turn serve as inputs into a systems model. Third, the integration of standard validation approaches into systems science modeling will be critical. Although the use of methods such as Latin hypercube sampling for uncertainty analysis and Markov chain Monte Carlo methods for parameter estimation have been used in infectious disease modeling, they are relatively underexplored in the application of systems modeling in other areas of epidemiology.

As data become increasingly "big" and machine-learning systems for predictive modeling become more sophisticated, we can use systems modeling to test assumptions about the types of conditions that give rise to population distributions of disease, as well as the types of interventions that can reduce the population rates of disease, in ways we cannot yet with real data. By combining publicly accessible central data repositories with area- and person-level prediction and causal frameworks, applying careful and systematic validation methods, and integrating systems science modeling into the broader epidemiologic armamentarium through an iterative data collection—modeling—data collection—modeling—intervention—evaluation—modeling process, we can develop valid and reliable systems models that can inform policy priorities.

ACKNOWLEDGMENTS

Author affiliations: Department of Population Health, New York University School of Medicine, New York, New York (Magdalena Cerdá); and Department of Epidemiology, Columbia University Mailman School of Public Health, New York, New York (Katherine M. Keyes).

M.C. and K.M.K. contributed equally to this work.

This work was funded by grant 1R21DA041154-01 from the National Institute on Drug Abuse, National Institute of Health.

Conflict of interest: none declared.

REFERENCES

1. Ip EH, Rahmandad H, Shoham DA, et al. Reconciling statistical and systems science approaches to public health. *Health Educ Behav*. 2013;40(1 suppl):123S–131S.
2. El-Sayed AM, Seemann L, Scarborough P, et al. Are network-based interventions a useful antiobesity strategy? An application of simulation models for causal inference in epidemiology. *Am J Epidemiol*. 2013;178(2):287–295.
3. Luke DA, Stamatakis KA. Systems science methods in public health: dynamics, networks, and agents. *Annu Rev Public Health*. 2012;33:357–376.
4. Pearce N, Merletti F. Complexity, simplicity, and epidemiology. *Int J Epidemiol*. 2006;35(3):515–519.
5. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. *Proc R Soc Lond Ser A Contain Papers Math Phys Char*. 1927;115(772):700–721.
6. Halloran ME, Longini IM, Jr., Nizam A, et al. Containing bioterrorist smallpox. *Science*. 2002;298(5597):1428–1432.
7. 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. 1766. *Rev Med Virol*. 2004;14(5): 275–288.

8. Heesterbeek H, Anderson RM, Andreasen V, et al. Modeling infectious disease dynamics in the complex landscape of global health. *Science*. 2015;347(6227):aaa4339.
9. Galea S, Hall C, Kaplan GA. Social epidemiology and complex system dynamic modelling as applied to health behaviour and drug use research. *Int J Drug Policy*. 2009; 20(3):209–216.
10. Galea S, Riddle M, Kaplan GA. Causal thinking and complex system approaches in epidemiology. *Int J Epidemiol*. 2010; 39(1):97–106.
11. Marshall B, Paczkowski M, Seeman L, et al. A complex systems approach to evaluate HIV prevention in metropolitan areas: preliminary implications for combination intervention strategies. *PLoS One*. 2012;7(9):e44833.
12. Tracy M, Cerdá M, Galea S. Causal thinking and complex systems approaches for understanding the consequences of trauma. In: Spatz Widom C, ed. *Trauma, Psychopathology and Violence: Causes, Consequences, or Correlates?* New York, NY: Oxford University Press; 2011:233–264.
13. Tracy M, Cerdá M, Keyes KM. Agent-based modeling in public health: current applications and future directions. *Annu Rev Public Health*. 2018;39:77–94.
14. Auchincloss AH, Diez-Roux AV. A new tool for epidemiology: the usefulness of dynamic agent-models in understanding place effects on health. *Am J Epidemiol*. 2008; 168(1):1–8.
15. Diez-Roux AV. Complex systems thinking and current impasses in health disparities research. *Am J Public Health*. 2011;101(9):1627–1634.
16. Cerdá M, Tracy M, Keyes KM. Reducing urban violence: a contrast of public health and criminal justice approaches. *Epidemiology*. 2018;29(1):142–150.
17. Cerdá M, Tracy M, Ahern J, et al. Addressing population health and health inequalities: the role of fundamental causes. *Am J Public Health*. 2014;104(suppl 4):S609–S619.
18. Nianogo RA, Arah OA. Agent-based modeling of noncommunicable diseases: a systematic review. *Am J Public Health*. 2015;105(3):e20–e31.
19. Yang Y, Diez-Roux A, Evenson KR, et al. Examining the impact of the walking school bus with an agent-based model. *Am J Public Health*. 2014;104(7):1196–1203.
20. Yonas MA, Burke JG, Brown ST, et al. Dynamic simulation of crime perpetration and reporting to examine community intervention strategies. *Health Educ Behav*. 2013;40(1 suppl): 87S–97S.
21. Scott N, Hart A, Wilson J, et al. The effects of extended public transport operating hours and venue lockout policies on drinking-related harms in Melbourne, Australia: results from SimDrink, an agent-based simulation model. *Int J Drug Policy*. 2016;32:44–49.
22. Wakeland W, Nielsen A, Geissert P. Dynamic model of nonmedical opioid use trajectories and potential policy interventions. *Am J Drug Alcohol Abuse*. 2015;41(6):508–518.
23. Lesko CR, Buchanan AL, Westreich D, et al. Generalizing study results: a potential outcomes perspective. *Epidemiology*. 2017;28(4):553–561.
24. Westreich D, Edwards JK, Lesko CR, et al. Transportability of trial results using inverse odds of sampling weights. *Am J Epidemiol*. 2017;186(8):1010–1014.
25. Marshall BD, Galea S. Formalizing the role of agent-based modeling in causal inference and epidemiology. *Am J Epidemiol*. 2015;181(2):92–99.
26. Marshall BD, Friedman SR, Monteiro JF, et al. Prevention and treatment produced large decreases in HIV incidence in a model of people who inject drugs. *Health Aff (Millwood)*. 2014;33(3):401–409.
27. Keane C, Egan JE, Hawk M. Effects of naloxone distribution to likely bystanders: results of an agent-based model. *Int J Drug Policy*. 2018;55:61–69.
28. Hernán MA. Invited commentary: agent-based models for causal inference-reweighting data and theory in epidemiology. *Am J Epidemiol*. 2015;181(2):103–105.
29. Murray EJ, Robins JM, Seage GR, et al. A comparison of agent-based models and the parametric G-formula for causal inference. *Am J Epidemiol*. 2017;186(2):131–142.
30. Diez-Roux AV. Invited commentary: the virtual epidemiologist-promise and peril. *Am J Epidemiol*. 2015; 181(2):100–102.
31. Naimi AI. Commentary: integrating complex systems thinking into epidemiologic research. *Epidemiology*. 2016;27(6): 843–847.
32. Rudolph KE, Schmidt NM, Glymour MM, et al. Composition or context: using transportability to understand drivers of site differences in a large-scale housing experiment. *Epidemiology*. 2018;29(2):199–206.
33. Windrum P, Fagiolo G, Moneta A. Empirical validation of agent-based models: alternatives and prospects. *J Artif Soc Soc Simul*. 2007;10(2):1–8.
34. McKay M, Beckman R, Conover W. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*. 1979;21(2):239–245.
35. Hamra G, MacLehose R, Richardson D. Markov chain Monte Carlo: an introduction for epidemiologists. *Int J Epidemiol*. 2013;42(2):627–634.
36. Toni T, Welch D, Strelkowa N, et al. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interface*. 2009;6(31): 187–202.
37. Akbari OS, Matzen KD, Marshall JM, et al. A synthetic gene drive system for local, reversible modification and suppression of insect populations. *Curr Biol*. 2013;23(8):671–677.
38. Mooney SJ, Westreich DJ, El-Sayed AM. Commentary: epidemiology in the era of big data. *Epidemiology*. 2015;26(3): 390–394.
39. Grefenstette JJ, Brown ST, Rosenfeld R, et al. FRED (a Framework for Reconstructing epidemic Dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC Public Health*. 2013;13:940.
40. Cajka JC, Cooley PC, Wheaton WD. Attribute assignment to a synthetic population in support of agent-based disease modeling. *Methods Rep RTI Press*. 2010;19(1009):1–14.
41. Wheaton WD, Cajka JC, Chasteen BM, et al. Synthesized population databases: a US geospatial database for agent-based models. *Methods Rep RTI Press*. 2009;2009(10):905.
42. Goin DE, Rudolph KE, Ahern J. Predictors of firearm violence in urban communities: a machine-learning approach. *Health Place*. 2018;51:61–67.
43. Neill DB, Herlands W. Machine learning for drug overdose surveillance. *J Technol Hum Serv*. 2018;36(1):8–14.