American Journal of Epidemiology

Vol. 181, No. 2
DOI: 10.1093/aje/kwu272
Advance Access publication:
December 5, 2014

Invited Commentary

Invited Commentary: Agent-Based Models for Causal Inference—Reweighting Data and Theory in Epidemiology

Miguel A. Hernán*

* Correspondence to Dr. Miguel A. Hernán, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115 (e-mail: miguel_hernan@post.harvard.edu).

Initially submitted August 10, 2014; accepted for publication August 27, 2014.

The relative weights of empirical facts (data) and assumptions (theory) in causal inference vary across disciplines. Typically, disciplines that ask more complex questions tend to better tolerate a greater role of theory and modeling in causal inference. As epidemiologists move toward increasingly complex questions, Marshall and Galea (*Am J Epidemiol*. 2015;181(2):92–99) support a reweighting of data and theory in epidemiologic research via the use of agent-based modeling. The parametric g-formula can be viewed as an intermediate step between traditional epidemiologic methods and agent-based modeling and therefore is a method that can ease the transition toward epidemiologic methods that rely heavily on modeling.

agent-based models; causal inference; parametric g-formula

Causal inferences typically combine empirical facts (data) and assumptions (theory). The acceptable relative weights of data and theory vary across scientific disciplines. Marshall and Galea (1) propose a rearrangement of the traditional roles of data and theory for causal inference in epidemiology.

To see this, let us start with an oversimplification. Consider a spectrum from "causal inference based exclusively on data" to "causal inference based exclusively on theory" and the positions of medicine and social science, 2 disciplines closely related to epidemiology, along the spectrum (Figure 1). Part of the oversimplification arises from the proposed clear-cut separation between data and theory, which has been contested by several authors (2, 3). For expediency, this commentary will sidestep the nuances of the data-theory debate.

In modern medicine, the demonstration of cause-effect relations requires randomized experiments. Statements such as "drug A is better, on average, than drug B for patients with cancer X" carry little weight unless they are supported by findings from a randomized clinical trial in which patients with cancer X are randomly assigned to either A or B. The goal is to make causal inferences as independent of theoretical arguments and expert opinion as possible. If a large, well-designed, randomized trial finds a difference, we will accept that there is an average causal effect in that population, regardless of our preconceptions. From that viewpoint, even well-conducted and analyzed observational studies are suspect because they typically require unverifiable assumptions about the comparability of the persons receiving each treatment, for example, the assumption that investigators have appropriately measured and adjusted for all confounders.

In social sciences, statements such as "economic inequality decreases the growth domestic product" are often taken seriously by many, even in the absence of an experiment that randomly assigns entire countries to different levels of inequality. Because such an experiment is impossible to carry out, social scientists making causal inferences about economic inequality need to integrate data from multiple sources—more modest experiments that test related issues, nationwide ecologic data, multiple observations across populations, time series, etc.—using some theoretical framework and possibly a mathematical model.

Thus, social sciences are generally closer to the pole "causal inference based exclusively on theory" than is medicine, which is closer to the other end of the spectrum. This relative position reflects the fact that medicine is blessed with scientific questions that can often be addressed by randomized experiments. (However, it took hard work by the pioneers of evidence-based medicine to free medicine from decisions founded exclusively upon expert opinion.) Social sciences, on the other hand, tend to ask questions that do not lend themselves to experimentation or even to emulation of an experiment

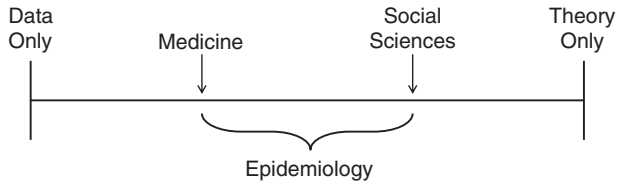

Figure 1. The relative position of several scientific disciplines along the causal inference spectrum according to the relative weights of data and theory.

using observational data. Hence, there is greater dependence on theoretical models to fill in the gaps and to provide a scaffolding to organize the various empirical findings.

What about epidemiology? Like medicine, epidemiology asks causal questions about effects on human health. Many epidemiologic questions (e.g., What are the effects of cigarette smoking?) cannot be answered via randomized experiments because of ethical, logistic, or practical constraints but could be hypothetically answered via randomized experiments in a world free of those constraints. That is, it is logically possible to imagine a randomized trial in which teenagers are randomly assigned to (and forced to comply with) either a lifetime of cigarette smoking or no smoking at all. As a result, many epidemiologic studies use observational data to mimic a hypothetical randomized experiment in a particular population at a particular time. For example, a study that compares the mortality rate between smokers and nonsmokers after adjustment for confounders is an implicit attempt to emulate a hypothetical randomized trial of cigarette smoking. Epidemiologists who combine individuals' data on smoking, lung cancer, and confounders with untestable assumptions, such as no unmeasured confounders, are, knowingly or not, adhering to the experimental paradigm to identify causal effects. (Some of us have called for a more explicit identification of the emulated or target experiment in observational studies (4, 5)).

Thus, the practice of epidemiology suggests that many epidemiologists desire to be closer to the methods of medicine than to those of the social sciences; the focus is on obtaining high-quality data from many individuals while relying as little as possible on theory and modeling. However, epidemiologists pay a hefty price for this strategy of emulation of experiments with observational data. As Glass et al. noted, "Epidemiologists and public health practitioners can be induced to prioritize the study of proximal, downstream interventions at the individual level. For example, it is easier to conduct, or emulate using observational data, randomized trials of smoking-cessation programs that target individuals than to conduct trials about the behavior of well-funded corporate entities with vested interests and political connections" (6, p. 70).

Such is the cost of minimizing the role of theory: It necessitates addressing narrower questions. As an example, we ask questions about the health effects of hypothetical interventions on individuals' behavior, such as diet and physical activity, more often than questions about the health effects of hypothetical interventions on key components of societal structure, such as our tax system, educational policy, corporate behavior, and intergenerational wealth redistribution. What can be done by epidemiologists who are interested in more upstream interventions for which the randomized experiment cannot be conducted or emulated but that might have the greatest potential to change health outcomes? What about epidemiologists interested in interventions in complex systems for which data to produce informative estimates do not exist? Marshall and Galea, like others before, provide an answer: Increase the weight of theory relative to data by using agent-based modeling.

Modelers create a mathematical model of reality. To do so, they combine empirical findings with essentially unverifiable assumptions (theory) about how the world works. By using data from multiple sources to guess the parameters that define the model, agent-based models become a scientific collage that can be used to make causal inferences across populations, calendar periods, exposures, and levels of intervention. As Marshall and Galea remind us, agent-based models can be naturally studied within the counterfactual framework; these models are obviously used to estimate counterfactual quantities, that is, to estimate how much the world would have changed if we had implemented a particular intervention.

Mathematical models, including agent-based models, are common tools in scientific disciplines that ask complicated causal questions, such as social sciences, systems biology, climate science, health policy, and neuroscience. These models describe systems that exhibit dynamically complex properties, such as interdependence of causal effects, feedback loops, and interference. If approximately correct, the model becomes a powerful tool to answer questions so complex that no data set in the world can answer them directly. For example, agent-based models can be used to determine the optimal way to prevent coronary heart disease in the United States by combining parameters estimated in the Framingham Heart Study with expert knowledge about disease progression in human populations; to find the optimal schedule for colon cancer screening by combining parameters estimated from randomized trials with basic sciences findings about tumor biology; or to compare the nationwide effects (on health and cost) of several policies to implement personalized strategies for antiretroviral therapy initiation and maintenance in South Africa over a 60-year period by combining information from multiple sources.

On the other hand, the inferences from the model cannot be experimentally tested or even approximated from a traditional observational study in a timely fashion. Otherwise, we would have used traditional epidemiologic study methods—the inferences of which depend less on theory and more on data—from the start. That is, if the model is essentially incorrect, we may never find out about it. To tackle model misspecification, disciplines with a wide use of agent-based models have developed a set of generally accepted principles to build models, populate and calibrate their parameters, and test their predictions.

Marshall and Galea recognize that a consensus on best practice methods to evaluate the validity of agent-based models is also needed in epidemiologic research. It is to be hoped that future work will focus on hard questions related to the combination of theory and data in agent-based models. Where will the data come from? What data are allowed? How can we combine data from different populations, time periods, and definitions of exposure that are collected at different levels (cellular, individual, local, regional, global)? How do we decide which sensitivity analyses must be conducted? Most

importantly, which complex systems do we understand well enough for modeling to be a reasonable option? (Much acrimony has erupted over the billion-Euro Human Brain Project because many neuroscientists simply do not think that it is possible to simulate the brain at this time (7).)

Many practicing epidemiologists, attached to their cherished data, may not be prepared to jump head first into the world of agent-based modeling. Interestingly, there is a middle ground between causal inference from mathematical models and traditional epidemiologic studies: the parametric g-formula (8). The parametric g-formula is a hybrid approach that uses a mathematical model (formally equivalent to an agent-based model) to estimate (via regression and Monte Carlo simulation) the effects of hypothetical interventions from a single data set. This method allows us to naturally handle interdependences of causal effects, feedback loops (including time-varying confounders affected by prior exposure), and other components of complex dynamic systems described by Marshall and Galea.

The parametric g-formula can be used to answer complex questions in a particular population without wandering too far from the study data. For example, we have used the parametric g-formula to compare the effects of joint dietary and lifestyle interventions on coronary heart disease (9, 10), diabetes (11), and asthma (12) in the Nurses' Health Study, as well as to compare the effects of therapeutic interventions in persons infected with the human immunodeficiency virus (13, 14). As of today, there seem to be no applications of the g-formula that incorporate interference or transmission, which hints at promising avenues for methodologic research.

The reliance on a single data set may make the parametric g-formula a more palatable form of modeling for epidemiologists because the method ensures less extrapolation from the data and a more direct calibration of the model. However, this protection against extrapolation comes at a cost: The parametric g-formula is restricted to causal inferences that do not go beyond the context and timeframe of the studied population. In contrast, agent-based models can be used to make causal inferences for any population, setting, and timeframe of interest. For example, one could apply the parametric g-formula to an observational cohort of Spaniards infected with the human immunodeficiency virus and followed for 5 years to estimate the 5-year survival curve under 3 dynamic strategies of antiretroviral therapy initiation that are actually observed in the cohort. However, one would not apply the parametric g-formula to that cohort to estimate the lifetime survival curve in all countries under dynamic strategies of antiretroviral therapy initiation that have never been tried in humans. That is precisely the type of questions that agent-based models attempt to answer, at the risk of extrapolation error.

In summary, Marshall and Galea have joined other investigators who believe that answering complex causal questions requires shifting the dial of causal inference away from data and toward a greater reliance of theory. They seem to be genuinely correct. However, using agent-based modeling will require a tectonic change in many epidemiologists' attitude about the relative roles of data and theory. The parametric g-formula may be an intermediate step that more epidemiologists are willing to try, and one that can ease the transition into full-blown agent-based modeling.

ACKNOWLEDGMENTS

Author affiliation: Department of Epidemiology Harvard School of Public Health, Boston, Massachusetts (Miguel A. Hernán); Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts (Miguel A. Hernán); and Harvard-MIT Division of Health Sciences and Technology, Boston, Massachusetts (Miguel A. Hernán).

This work was partly funded by National Institutes of Health grant R01 AI102634.

I thank Drs. Sander Greenland and Nancy Krieger for their helpful comments.

Conflict of interest: none declared.

REFERENCES

1. Marshall BDL, Galea S. Formalizing the role of agent-based modeling in causal inference and epidemiology. *Am J Epidemiol*. 2015;181(2):92–99.
2. Ziman JM. *Real Science: What It Is and What It Means*. Cambridge, UK: Cambridge University Press; 2000.
3. Krieger N. Does epidemiologic theory exist? On science, data, and explaining disease distribution. In: Krieger N, ed. *Epidemiology and the People's Health: Theory and Context*. New York, NY: Oxford University Press; 2011.
4. Hernán MA. With great data comes great responsibility: publishing comparative effectiveness research in epidemiology [editorial]. *Epidemiology*. 2011;22(3):290–291.
5. Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology*. 2014;25(3):418–426.
6. Glass TA, Goodman SN, Hernán MA, et al. Causal inference in public health. *Annu Rev Public Health*. 2013;34:61–75.
7. Brain fog [editorial]. *Nature*. 2014;511:125.
8. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Model*. 1986;7(9-12):1393–1512.
9. Taubman SL, Robins JM, Mittleman MA, et al. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *Int J Epidemiol*. 2009;38(6):1599–1611.
10. Lajous M, Willett WC, Robins JM, et al. Changes in fish consumption in midlife and the risk of coronary heart disease in men and women. *Am J Epidemiol*. 2013;178(3):382–391.
11. Danaei G, Pan A, Hu FB, et al. Hypothetical midlife interventions in women and risk of type 2 diabetes. *Epidemiology*. 2013;24(1):122–128.
12. Garcia-Aymerich J, Varraso R, Danaei G, et al. Incidence of adult-onset asthma after hypothetical interventions on body mass index and physical activity: an application of the parametric g-formula. *Am J Epidemiol*. 2014;179(1):20–26.
13. Young JG, Cain LE, Robins JM, et al. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Stat Biosci*. 2011;3(1):119–143.
14. Westreich D, Cole SR, Young JG, et al. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Stat Med*. 2012;31(18):2000–2009.