Sept 24 2020

Complex systems models for causal inference in social epidemiology.

Hiba Kouser[1], Ruby Barnard-Mayers[1], Eleanor J Murray[1]

[1]Department of Epidemiology, Boston University School of Public Health, Boston MA

Corresponding Author:

Eleanor J Murray, ejmurray@bu.edu, Department of Epidemiology, Boston University, Boston MA, 02445

Acknowledgements: This project was supported by the NICHD (R21HD098733).

Supplementary Online Appendix: A step-by-step guide to building a systems model to estimate causal effects

In our example, we are interested in answering the larger question "*how can we best distribute limited COVID-19 testing resources in order to minimize social inequities*?". In order to design a systems model to answer this question, we need to specify a specific population of interest – for example, residents of Massachusetts in 2020. We also need to define what we mean by "social inequities" – this is more challenging, but for the purposes of our example we shall define social inequities as the degree to which COVID-19 death rates in BIPOC communities exceeds the proportional distribution of these communities in society. Finally, we need to specify possible COVID-19 testing resource distribution schemes. For example, we could specify a strategy such as "equal distribution of testing sites in census tracts based on population levels", and a comparison strategy such as "distribution of testing sites based on COVID-19 infection rates over the prior 2 weeks".

Based on these, the researchers can begin to detail the types of factors to be modeled, such as individual health characteristics, neighborhood or geographic region or movement patterns, environmental characteristics, and other social characteristics. The agents in a systems model can represent any type of unit, such as individuals, businesses, manufacturing processes, or governments. The choice of factors, or layers, to be modeled will create the world in which the agents operate. This world will then define rules agents follow, the 'environment' in which they will operate, and the level of detail needed for the model. As these models are agent-centric, the rules dictate the causal sphere in which the agents move about. This first step provides the base of the model upon which to build more complex relationships.

The second step of causal effect estimation is to design a hypothetical trial protocol based on the framework you create in step 1. This trial should be one that could, in theory, be conducted on your population of interest to obtain an answer to your specified research question, even if ethical, financial, logistical, complexity, lack of manipulability, or other barriers prevent the actual conduct of such a trial. The goal is to clarify the specific exposure settings of interest for comparison, typically structured around a decision or action that could (at least in theory) be undertaken. The components of the trial that need to be specified include: (1) eligibility, inclusion & exclusion criteria; (2) definition of the exposure strategies of interest; (3) definition of the start and end of follow-up, including any time-varying exclusion criteria that might develop; (4) specification of the causal estimand of interest (for example, intention-to-treat versus per-protocol effect [1]; and (5) analysis plan to obtain an estimate of that effect from a randomized trial.

In our example, a randomized trial designed to evaluate the benefits on inequities in COVID-19 mortality rates due to different testing schemes would likely need to be cluster-randomized, at perhaps the census tract, zip code, or other neighborhood-level identifier. Clusters would be randomized to either receive the number of testing sites dictated by an equal distribution of resource based on population size, or the number of testing sites dictated by a distribution of resources based on observed COVID-19 rates over the past 2 weeks. The number of COVID-19 deaths over a specified time period, perhaps 8 weeks, and demographic information of people who die of COVID-19 would be recorded in all clusters, and the COVID-19 death rate within race and ethnicity groups, relative to their distribution within each cluster, would be compared between clusters assigned to each group. A full description of this trial would be more detailed, but this general overview provides us with a framework to begin formulating a systems model to answer this same question.

Once the target trial has been specified, the next step is to decide on the rules that will determine how the agents change over time. These rules reflect the core set of assumptions governing the processes that the systems model is designed to emulate. Importantly, although these rules are commonly referred to as assumptions, they should be based on empirical data when possible. When these rules accurately represent the real-world in sufficient detail, the systems model can potentially be used to estimate causal effects in the real-world (while recognizing that if complete empirical data were available, a systems model would likely be unnecessary) [2]. There is a trade-off between the availability of data and the accuracy of assumptions that can be made in the modeling process which will dictate the level of uncertainty in the model results such that the more data are available, the fewer assumptions are required and the more likely the model is to give a reliable answer. [3] This trade-off has been observed in COVID-19 models, where early models had access to much less data about SARS-CoV-2 and COVID-19, and therefore relied on many more assumptions – often based upon knowledge of more well-known diseases such as influenza or SARS – while later models have been able to incorporate more COVID-specific information.

Step three of building a systems model is thus to specify these assumptions; causal Directed Acyclic Graphs (DAGs) can be immensely helpful in accomplishing this step. [4-9] Causal DAGs are useful for defining potential confounders, colliders, and mediators of an exposure-outcome relationship. In this way, DAGs complement and support simulation model development – Figure 2 shows the result of translating the decision tree in Figure 1 into a causal DAG. Here, we can see that pre-existing conditions, and insurance status are potential confounders for the impact of COVID-19 testing availability on mortality. Testing site location is a potential confounder for the parameter input representing the relationship between COVID-19 testing and COVID-19 diagnosis.

The fourth step is to build a model which answers the question identified in Step 1 by emulating the design of the target trial in Step 2, while incorporating the required variables identified in Step 3 as potential confounders, colliders, mediators, or effect modifiers. This can require a complex back-and-forth between the trial and model designs but is likely simpler to do than emulating a trial from observational data, since the systems model can be designed to exactly mimic features of the Target Trial. [10,11] The computational and mathematical details of creating a systems model are beyond the scope of this review, however a number of texts, software tools, and training resources exist for the interested reader. [12-15]

The fifth and final step is to design uncertainty analyses for the systems model. [16] Uncertainty in systems models stems from three major sources: stochasticity of the simulation process (introduced through random sampling), variance and uncertainty in the parameter input values, and uncertainty in the structure of the model. Stochastic uncertainty is typically addressed by increasing the simulation size or averaging results over multiple model runs, while parametric uncertainty can be quantified by performing probabilistic sensitivity analyses. [16] Structural uncertainty is more complicated to address, but can be assessed by varying the core structure of the systems model including thinking about whether there may be important omitted relationships between variables in the original model design.

References

1. Hernán MA, Robins JM. Per-protocol analyses of pragmatic trials. N Eng J Med. 2017;377(14):1391-1398. doi:10.1056/NEJMsm1605385
2. Murray EJ, Robins JM, Seage GR, Freedberg KA, Hernán MA. A comparison of agent-based models and the parametric g-formula for causal inference. Am J Epidemiol. 2017;186(2):131-142. doi:10.1093/aje/kwx091
3. Hernán MA. Invited commentary: Agent-based models for causal inference—reweighting data and theory in epidemiology. Am J Epidemiol. 2015;181(2):103-105. doi:10.1093/aje/kwu272
4. Hernan MA, Robins J. Causal Inference: What If. Chapman & Hill/CRC; 2020.

5. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period — Application to the healthy worker survivor effect. Mathematical Modelling. 1986;7:1393-1512.
6. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37-48.
7. Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82(4):669-688. doi:10.1093/biomet/82.4.669
8. Tennant PW, Harrison WJ, Murray EJ, et al. Use of directed acyclic graphs (DAGs) in applied health research: review and recommendations. medRxiv. Published online January 1, 2019:2019.12.20.19015511. doi:10.1101/2019.12.20.19015511
9. Joffe M, Gambhir M, Chadeau-Hyam M, Vineis P. Causal diagrams in systems epidemiology. Emerging Themes in Epidemiology. 2012;9(1):1. doi:10.1186/1742-7622-9-1
10. Buchanan A, King M, Bessey S, et al. Disseminated effects in agent based models: a potential outcomes framework to inform pre-exposure prophylaxis coverage levels for HIV prevention. In: Society for Epidemiologic Research Annual Conference. ; 2019.
11. Lodi S, Phillips A, Lundgren J, et al. Effect estimates in randomized trials and observational studies: comparing apples with apples. Am J Epidemiol. Published online May 7, 2019. doi:10.1093/aje/kwz100
12. Hunink MGM. Decision making in health and medicine: integrating evidence and values. In: Cambridge University Press; 2001:305-338. http://hollis.harvard.edu/?itemid=%7Clibrary/m/aleph%7C008748782
13. TreeAge Pro 2020, R2. TreeAge Software
14. Bonabeau E. Agent-based modeling: methods and techniques for simulating human systems. Proc Natl Acad Sci U S A. 2002;99 Suppl 3:7280-7287. doi:10.1073/pnas.082080899
15. Meadows DH. Thinking in Systems: A Primer. (Wright D, ed.). Chelsea Green Publishing; 2008.
16. Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB. Model transparency and validation: a report of the ISPOR-SMDM modeling good research practices task force-7. Value in Health. 2012;15(6):843-850. doi:10.1016/j.jval.2012.04.012