COMMENTARY

Note on "Generalizability of Study Results"

Judea Pearl[a] *and Elias Bareinboim*[b]

Keywords: Data fusion; External validity; Generalizability; Selection bias; Transportability

The purpose of this note is to complement the review by Lesko et al.[1] and to present readers with a comparison between two approaches to the analysis of generalizability: (1) potential outcomes and (2) model-based data fusion. We show that, while the inferential power of the former is curtailed by a priori assumptions of conditional exchangeability, the latter unveils and leverages the full range of opportunities that are licensed by both the available data and one's model of reality, regardless of whether the model supports the assumption of conditional exchangeability.

In general, the potential outcomes perspective falls short of addressing three fundamental issues in causal analysis:

(1) to determine if there exist sets of covariates W that satisfy "conditional exchangeability" (be it of treatment assignments, of populations' heterogeneity, or of selection indicators),
(2) to estimate causal parameters at the target population in cases where such sets W do not exist, and
(3) to decide if one's modeling assumptions are compatible with the available data.

These deficiencies curtail the analyses of both internal validity and external validity problems but are more pronounced in the latter, where the assumptions required are more involved. While the assumptions that ensure internal validity concern patients choice of treatments, those required for external validity concern both the way subjects are selected for the study, the factors that make the population different from the target population, and how those differences modify the effect of treatment on outcome. Technically, the standard conditional exchangeability that is assumed for neutralizing confounding, $X||Y(x)|W$, involves only three variables, X, Y, and W. By contrast, the conditional exchangeability assumed to neutralize disparities between the study and the target populations, $S||Y(x)|W$, involves four variables, Y, X, W, and S (where S is an indicator of membership in the study sample or source of unmodeled disparities). The parallels Lesko et al.[1] draw between the two tasks are syntactic at best and should not suggest that a researcher who can judge the plausibility of the former can also judge the plausibility of the latter.

From the [a]Computer Science Department, University of California Los Angeles, Los Angeles, CA; and [b]Computer Science Department, Purdue University, West Lafayette, IN.

This research was supported in parts by grants from Defense Advanced Research Projects Agency No. W911NF-16-057, National Science Foundation No. IIS-1302448, No. IIS-1527490, and No. IIS-1704932, and Office of Naval Research No. N00014-17-S-B001 (to J.P.). This research was supported in parts by grants from National Science Foundation No. 1704352 and No. IIS-1750807 (CAREER; to E.B.).

The authors report no conflicts of interest.

Correspondence: Judea Pearl, Computer Science Department, University of California Los Angeles, 3531 Boelter Hall, Los Angeles, CA 90056. E-mail: judea@cs.ucla.edu.

ISSN: 1044-3983/19/3002-0186
DOI: 10.1097/EDE.0000000000000939

Lesko et al.[1(page 555)] appear to be aware of the first limitation and state, "Judging whether a set of characteristics W is sufficient to satisfy this independence assumption (i.e., conditional exchangeability) may be a difficult task". They consequently recommend the use of a directed acyclic graph (DAG), with the help of which conditional exchangeability "can be verified by inspection." This transparency of DAG-based models has long been recognized by practicing epidemiologists who have adapted these models as effective communication tools.[2–4] What perhaps is less recognized among epidemiologists is the use of DAGs as inference tools, capable of extending the scope of analysis beyond the boundaries imposed by the potential outcome perspective.

The major difference between the two approaches lies in the fact that, whereas the potential outcomes framework confines the analysis to problems where "conditional exchangeability" can be assumed a priori, the only assumptions invoked in the model-based framework are those encoded in the structure of the DAG, and these often have broader ramifications, going beyond the exchangeability variety. Consequently, the problems discussed in Lesko et al.[1] can be given a more general solution, covering transportability as well as other generalization tasks under the same umbrella, and embracing arbitrary disparities between multiple study populations and the target population. This model-based framework, called Data Fusion,[5–7] will be described in the next section.

THE DATA FUSION FRAMEWORK: A GENERAL SOLUTION TO EXTERNAL VALIDITY

The Data Fusion framework takes an arbitrary set of heterogeneous data sources and produces a consistent estimand of the target quantity at the target population. Clearly, to embark on such ambitious task, we must first arm ourselves with a formal notation and a formal logic to assure that the resulting estimand is valid. Accordingly, the theory provides us with the notation needed to characterize the nature of each data source, the nature of the population interrogated, whether the source is an observational or experimental study, which variables are randomized and which are measured; finally, as output, the theory tells us how to fuse all these sources together to synthesize a consistent estimand of the target causal quantity at the target population. Moreover, if we feel uncomfortable about the assumed structure of any given data source, the theory tells us whether an alternative source can furnish the needed information and whether we can weaken any of the model's assumptions.

As described above, the Data Fusion framework sounds like the Philosopher's Stone of epidemiologic research, and readers would be justified in doubting the ability of any analytic method to accomplish such a general and ambitious goal. However, readers familiar with the power of the do-calculus to automate the derivation of internally valid effects[8] should not be surprised to see this power replicated and amplified when applied to problems of external validity. Space limitations permit us merely to sketch the inference strategy of Data Fusion. A gentle introduction is given in Bareinboim and Pearl[5] and Pearl and Bareinboim,[6] while a full technical account and proofs of completeness can be found in Bareinboim and Pearl.[9,10]

The inference strategy invoked by the Data Fusion framework stands almost diametrically opposed to that invoked in the potential outcome framework. Instead of starting with exchangeability type assumptions in order to justify a familiar recalibration procedure, we reverse the order and start with the target quantity itself, also called "query," and ask what estimation procedure would properly represent the query. This amounts to converting the query into a new mathematical form that would be estimable from the available data, however, heterogeneous. The conversion process relies solely on the transparent structure of the DAG, needing no external assumptions of ignorability or exchangeability. If the conversion is successful, the resulting expression provides a recipe for pooling chunks of data from their various sources and assembling them together so as to estimate the query. If the conversion is unsuccessful, we are assured that the query is inestimable given the model at hand. The conversion is done algorithmically, guided by the DAG and governed by the do-calculus.

In contrast to the potential outcome strategy, we impose no prior restriction on the form of the resulting estimand but, rather, allow it to emerge naturally and algorithmically from the problem description itself. Our strategy amounts to extracting from the problem description all opportunities for valid generalizations, many of which would be excluded under the restrictions of Lesko et al.[1] For example, problems in which conditional exchangeability does not hold can nevertheless be generalized,[7] albeit by nonstandard estimands, going beyond the simple adjustment described in equation 4 of Lesko et al.[1] Complete conditions for generalizing under both confounding and selection bias are derived in Correa et al.[11]

To summarize, the Data Fusion framework substantially extends the class of problems analyzable by the potential outcome framework and bases all its conclusions on transparent assumptions encoded in the DAG structure. Moreover, it provide us with guarantees of "completeness," which tells us, in essence, that one cannot do any better. In other words, it delineates precisely the minimum set of assumptions needed to establish consistent estimate of causal effects in the target population. If any of those assumptions is violated, we know that we can do only worse. From a mathematical (and philosophical) viewpoint, this is the most one can expect analysis to do for us, and therefore, completeness renders the generalizability problem "solved."

Completeness also tells us that any strategy of generalizing study results is either embraceable in the framework of Data Fusion or it is not workable in any framework. This means that one cannot dismiss the conclusions of Data Fusion theory on the grounds that "Its assumptions are too strong," or

"I do not have the knowledge to specify the DAG." If a set of assumptions is deemed necessary in the Data Fusion analysis, then it is necessary, period; it cannot be avoided or relaxed, unless it is supplemented by other assumptions elsewhere, and the algorithm can tell us where.

CONCLUSIONS

We commend Lesko et al.[1] for an illuminating survey of how the potential outcome framework deals with the problem of generalizing study results to target populations. In this note, we highlighted the general limitations inherent in the potential outcome framework and the specific limitations that apply to generalization problems. We provided a brief description of an alternative framework, called Data Fusion, which circumvents the limitations above and provides a complete solution to the problem of external validity using transparent and testable assumptions.

ABOUT THE AUTHORS

JUDEA PEARL is Chancellor's professor of computer science and statistics at University of California Los Angeles (UCLA), where he directs the Cognitive Systems Laboratory and conducts research in artificial intelligence, human cognition, and philosophy of science. He has authored numerous scientific papers and three books, Heuristics (1983), Probabilistic Reasoning (1988), and Causality (2000, 2009) which won the London School of Economics Lakatos Award in 2002. More recently, he coauthored Causal Inference in Statistics (2016, with M. Glymour and N. Jewell) and The Book of Why (2018, with D. Mackenzie) which brings causal analysis to general audience. Pearl is a member of the National Academy of Sciences and the National Academy of Engineering, a fellow of the Cognitive Science Society and a founding fellow of the Association for the Advancement of Artificial Intelligence. In 2012, he won the Technion's Harvey Prize and the ACM Alan Turing Award for the development of a calculus for probabilistic and causal reasoning. ELIAS BAREINBOIM is an assistant professor in the Department of Computer Science at Purdue University. His research focuses on causal and counterfactual inference and their applications to data-driven fields. Bareinboim received a PhD in Computer Science from UCLA. His doctoral thesis was the first to propose a general solution to the problem of "data fusion" and to provide practical methods for combining datasets generated under different experimental conditions. Bareinboim's recent explorations include causally enhanced design of experiments and fairness analysis. Bareinboim's recognitions include the NSF Faculty Early Career Development Program (CAREER) award, IEEE AI's 10 to Watch, the Dan David Prize Scholarship, the Yahoo! Key Scientific Challenges Award, and the 2014 AAAI Outstanding Paper Award.

REFERENCES

1. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. *Epidemiology*. 2017;28:553–561.
2. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10:37–48.
3. Krieger N, Smith GD. The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. *Int J Epidemiol*. 2016;45:1787–1808.
4. Pearl J. Comments on 'The tale wagged by the DAG'. *Int J Epidemiol*. 2018;47:81–86.
5. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci*. 2016;113:7345–7352.
6. Pearl J, Bareinboim E. External validity: from do-calculus to transportability across populations. *Stat Sci*. 2014;29:579–595.
7. Pearl J. Generalizing experimental findings. *JCI*. 2015;3:259–266.
8. Pearl J. The deductive approach to causal inference. *JCI*. 2014;2:115–129.
9. Bareinboim E, Pearl J. A general algorithm for deciding transportability of experimental results. *JCI*. 2013;1:107–134.
10. Bareinboim E, Pearl J. Transportability from multiple environments with limited experiments: completeness results. In: Welling M, Ghahramani Z, Cortes C, Lawrence N, eds. *Advances of Neural Information Processing*. Montréal, Canada, Curran Associates, Inc.; 2014:280–288.
11. Correa JD, Tian J, Bareinboim E. Generalized adjustment under confounding and selection biases. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, LA: AAAI Press; 2018.