Commentary: EPIDEMIOLOGY MS# EDE18-0227

A note on "Generalizability of Study Results"

Judea Pearl[*]

University of California Los Angeles

Computer Science Department

3531 Boelter Hall

Los Angeles, CA 90056

Tel: 310.825.3243   email: judea@cs.ucla.edu

Elias Bareinboim[†]

Purdue University

Computer Science Department

305 N. University Street 2142L

West Lafayette, IN, 47907-2107

Tel: 765.494.6010   email: eb@purdue.edu

Abbreviated running title: Model-based versus Model-free Generalizability

Keywords: External validity, generalizability, selection bias, transportability, data fusion

Introduction

The purpose of this note is to complement Lesko et al.'s review [1] and 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, $X \perp\!\!\!\perp Y(x) \mid W$, involves only three variables, X, Y, and W. By contrast, the conditional exchangeability assumed to neutralize disparities between the study and the target populations, $S \perp\!\!\!\perp Y(x) \mid W$, 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. 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.

Lesko et al. 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" [page 555]. 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, 3, 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. 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, 6, 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 problema of external validity. Space limitations permit us merely to sketch the inference strategy of Data Fusion. A gentle introduction is given in [5, 6] while a full technical account and proofs of completeness can be found in [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 re-calibration 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's et al. For example, problems in which conditional exchangeability does not hold can nevertheless be generalized [7], albeit by non-standard estimands, going beyond the simple adjustment described in Eq. (4) of Lesko et al. Complete conditions for generalizing under both confounding and selection bias are derived in [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)

3

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. 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.

References

[1] Catherine R. Lesko, Ashley L. Buchanan, Daniel Westreich, Jessie K. Edwards, Michael G. Hudgens, and Stephen R. Cole. Generalizing study results: A potential outcomes perspective. *Epidemiology*, 28(4):553–561, 2017.

[2] S. Greenland, J. Pearl, and J.M Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, 1999.

[3] Nancy Krieger and George Davey Smith. The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. *International Journal of Epidemiology*, 45(6):1787–1808, 2016.

[4] J. Pearl. Comments on 'The tale wagged by the DAG'. *International Journal of Epidemiology*, 47:81–86, 2018.

[5] E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113:7345–7352, 2016.

[6] Judea Pearl and Elias Bareinboim. External validity: From *do*-calculus to transportability across populations. *Statistical Science*, 29:579–595, 2014.

[7] Judea Pearl. Generalizing experimental findings. *Journal of Causal Inference*, 3:259–266, 2015.

[8] Judea Pearl. The deductive approach to causal inference. *Journal of Causal Inference*, 2:115–129, 2014.

[9] Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1(1):107–134, 2013.

[10] E. Bareinboim and J. Pearl. Transportability from multiple environments with limited experiments: Completeness results. In M. Welling, Z. Ghahramani, C. Cortes, and N. Lawrence, editors, *Advances of Neural Information Processing 27 (NIPS)*, pages 280–288. Curran Associates, Inc., Montreal Canada, 2014.

[11] Juan D. Correa, Jin Tian, and Elias Bareinboim. Generalized adjustment under confounding and selection biases. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press, New Orleans, LA, 2018.