Social Science & Medicine xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Challenging the hegemony of randomized controlled trials: A commentary on Deaton and Cartwright

Judea Pearl

University of California, Los Angeles Computer Science Department, Los Angeles, CA 90095-1596, USA

I appreciate the opportunity to comment on the article by Angus Deaton and Nancy Cartwright (D&C) (Deaton and Cartwright, 2018), which touches on the foundations of causal inference.

My comments are a mixture of a welcome and a puzzle; I welcome D&C's stand on the status of randomized trials, and I am puzzled by how they choose to articulate the alternatives.

D&C's main theme is as follows: "We argue that any special status for RCTs is unwarranted. Which method is most likely to yield a good causal inference depends on what we are trying to discover as well as on what is already known."

As a veteran skeptic of the supremacy of the RCT, I welcome D&C's challenge wholeheartedly. Indeed, *The Book of Why* (Pearl and Mackenzie, 2018, http://bayes.cs.ucla.edu/WHY/) quotes me as saying: "If our conception of causal effects had anything to do with randomized experiments, the latter would have been invented 500 years before Fisher." In this, as well as in my other writings I go so far as claiming that the RCT earns its legitimacy by mimicking the *do*-operator,[1] not the other way around. In addition, considering the practical difficulties of conducting an ideal RCT, observational studies have a definite advantage: they interrogate populations at their natural habitats, not in artificial environments choreographed by experimental protocols.

Deaton and Cartwright's challenge of the supremacy of the RCT consists of two parts: The first (internal validity) deals with the curse of dimensionality and argues that, in any single trial, the outcome of the RCT can be quite distant from the target causal quantity, which is usually the average treatment effect (ATE). In other words, this part concerns imbalance due to finite samples, and reflects the traditional bias-precision tradeoff in statistical analysis and machine learning. The second part (external validity) deals with biases created by inevitable disparities between the conditions and populations under study versus those prevailing in the actual implementation of the treatment program or policy.

Here, Deaton and Cartwright propose alternatives to RCT, calling all out for integrating a web of multiple information sources, including observational, experimental, quasi-experimental, and theoretical inputs, all collaborating towards the goal of estimating "what we are trying to discover."

My only qualm with D&C's proposal is that, in their passion to advocate the integration strategy, they have failed to notice that, in the past decade, a formal theory of integration strategies has emerged from the brewery of causal inference and is currently ready and available for empirical researchers to use. I am referring of course to the theory of Data Fusion, which formalizes the integration scheme in the language of causal diagrams, and provides theoretical guarantees of feasibility and performance (see Bareinboim and Pearl (2016)).

Let us examine closely D&C's main motto: "Which method is most likely to yield a good causal inference depends on what we are trying to discover as well as on what is already known." Clearly, to cast this advice in practical settings, we must devise notation, vocabulary, and logic to represent "what we are trying to discover" as well as "what is already known" so that we can infer the former from the latter. To accomplish this nontrivial task we need tools, theorems and algorithms to assure us that what we conclude from our integrated study indeed follows from those precious pieces of knowledge that are "already known." D&C are notably silent about the language and methodology in which their proposal should be carried out. One is left wondering therefore whether they intend their proposal to remain an informal, heuristic guideline, similar to Bradford Hill's Criteria of the 1960's, or be explicated in some theoretical framework that can distinguish valid from invalid inference? If they aspire to embed their integration scheme within a coherent framework, then they should celebrate; such a framework has been worked out and is now fully developed.

To be more specific, the Data Fusion theory described in Bareinboim and Pearl (2016) provides us with notation 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 and, finally, the theory tells us how to fuse all these sources together to synthesize an 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.

Those familiar with Data Fusion theory will find it difficult to understand why D&C have not utilized it as a vehicle to demonstrate the

E-mail address: judea@cs.ucla.edu.

[1] For a gentle introduction to the do-operator and do-calculus, see Pearl and Bareinboim (2014) or Pearl et al. (2016).

https://doi.org/10.1016/j.socscimed.2018.04.024

feasibility of their proposed alternatives to RCT's. This enigma stands out in D&C's description of how modern analysis can rectify the deficiencies of RCTs, especially those pertaining to generalizing across populations, extrapolating across settings, and controlling for selection bias.

Here is what D&C say about extrapolation (Quoting from their discussion on, “Re-weighting and stratifying"):

> "Pearl and Bareinboim (2011, 2014) and Bareinboim and Pearl (2013, 2014) provide strategies for inferring information about new populations from trial results that are more general than re-weighting. They suppose we have available both causal information and probabilistic information for population *A* (e.g. the experimental one), while for population *B* (the target) we have only (some) probabilistic information, and also that we know that certain probabilistic and causal facts are shared between the two and certain ones are not. They offer theorems describing what causal conclusions about population *B* are thereby fixed. Their work underlines the fact that exactly what conclusions about one population can be supported by information about another depends on exactly what causal and probabilistic facts they have in common."

The text is accurate up to this point, but then it changes gears and states:

> "But as Muller (2015) notes, this, like the problem with simple re-weighting, takes us back to the situation that RCTs are designed to avoid, where we need to start from a complete and correct specification of the causal structure. RCTs can avoid this in estimation which is one of their strengths, supporting their credibility but the benefit vanishes as soon as we try to carry their results to a new context."

I believe D&C miss the point about re-weighing and stratifying.

First, it is not the case that “this takes us back to the situation that RCTs are designed to avoid.” It actually takes us to a more manageable situation. RCTs are designed to neutralize the confounding of treatments, whereas our methods are designed to neutralize differences between populations. Researchers may be totally ignorant of the structure of the former and quite knowledgeable about the structure of the latter. To neutralize selection bias, for example, we need to make assumptions about the process of recruiting subjects for the trial, a process over which we have some control. There is a fundamental difference therefore between assumptions about covariates that determine patients’ choice of treatment and those that govern the selection of subjects—the latter is (partially) under our control. Replacing one set of assumptions with another, more defensible set, does not “take us back to the situation that RCTs are designed to avoid.” It actually takes us forward, towards the ultimate goal of causal inference—to base conclusions on scrutinizable assumptions, and to base their plausibility on scientific or substantive grounds.

Second, D&C overlook the significance of the “completeness” results established for transportability problems (see Bareinboim and Pearl (2012)). Completeness tells us, in essence, that one cannot do any better. In other words, it delineates precisely the minimum set of assumptions that are needed to establish consistent estimate of causal effects in the target population. If any of those assumptions are 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."

Finally, the completeness result highlights the broader implications of the Data Fusion theory, and how it brings D&C's desiderata closer to becoming a working methodology. Completeness tells us that any envisioned strategy of study integration is either embraceable in the structure-based 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 “It supposes we have causal information that we are not likely to have.” 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 new assumptions elsewhere, and the algorithm can tell you where.

It is hard to see therefore why any of D&C's proposed strategies would resist formalization, analysis and solution within the current logic of Data Fusion theory.

It took more than a dozen years for researchers to accept the notion of completeness in the context of internal validity, as it emerged from the *do*-calculus (see Pearl (1995); Shpitser and Pearl (2008); Tian and Pearl (2002)). Here, completeness tells us what assumptions are absolutely needed for nonparametric identification of causal effects, how to tell if they are satisfied in any specific problem description, and how to use them to extract causal parameters from non-experimental studies. Completeness in external validity context is a relatively new result (see Bareinboim and Pearl (2013)), which will probably take a few more years for enlightened researchers to accept, appreciate and to fully utilize. One purpose of this commentary is to urge the research community, especially Deaton and Cartwright to study the recent mathematization of external validity and to benefit from its implications.

Those familiar with Data Fusion theory will find it difficult to understand why D&C have not utilized it as a vehicle to demonstrate the feasibility of their proposed alternatives to RCT's. Those unfamiliar with the theory would probably say: “Who needs a new theory to do what statistics does so well?” “Once we recognize the importance of diverse sources of data, statistics can be helpful in making decisions and quantifying uncertainty.” [Quoted from Andrew Gelman's blog]. The reason I question the sufficiency of statistics to manage the integration of diverse sources of data is that statistics lack the vocabulary needed for the job. I will demonstrate it in a couple of toy examples taken from Bareinboim and Pearl (2016).

Example 1

Suppose we wish to estimate the causal effect of X on Y, and we have two diverse sources of data: (1) an RCT in which Z, not X, is randomized, and (2) an observational study in which X, Y, Z and perhaps other variables are measured. What substantive assumptions are needed to facilitate a solution to our problem? Put another way, how can we be sure that, once we make those assumptions, we can pool data from both studies and construct an (consistent) estimate of our target effect.

Example 2

Suppose we wish to estimate the average causal effect (ACE) of X on Y, and we have two diverse sources of data: (1) an RCT in which the effect of X on both Y and Z is measured, but the recruited subjects had an unusually high Z, and (2) an observational study conducted in the target population, in which both X and Z (but not Y) were measured. What substantive assumptions would enable us to estimate ACE, and how should we combine data from the two studies so as to synthesize a consistent estimate of ACE.

The nice thing about a toy example is that the solution is known to us in advance, and so, we can check any proposed solution for correctness. Curious readers can find the solutions for these two examples in Bareinboim and Pearl (2016). More traditional readers will probably try to solve them using statistic techniques, such as meta analysis or partial pooling. The reason I am confident that the second group will end up with disappointment comes from a profound statement made by Nancy Cartwright in 1989: “No Causes In, No Causes Out”. It means not only that you need substantive assumptions to derive causal conclusions; it also means that the vocabulary of statistical analysis, since it is built entirely on properties of distribution functions, is inadequate for expressing those substantive assumptions that are needed for getting causal conclusions. Although part of the data in our examples is provided by an RCT, hence it provides causal information, one can show mathematically that the additional assumptions needed for solving the problems above must invoke causal vocabulary; distributional

assumptions are insufficient. In other words, two statistically indistinguishable problems may require two different estimates, depending on their underlying causal structures. As someone versed in both graphical modeling and counterfactuals, I would go even further and state that it would be a miracle if anyone succeeds in translating the needed assumptions into a meaningful language other than causal diagrams (Scenario 3 in Pearl (2015), for example, shows why the language of potential outcomes and ignorability expressions are inadequate for expressing these assumptions.).

Armed with these examples and findings, we can go back and examine why D&C do not embrace the Data Fusion methodology in their quest for integrating diverse sources of data. The answer, I conjecture, is that D&C were not intimately familiar with what this methodology offers us and how vastly different it is from previous attempts to operationalize Cartwright's dictum: “No causes in, no causes out."

References

Bareinboim, E., Pearl, J., 2012. Transportability of causal effects: completeness results. In: Proceedings of the Twenty-sixth Conference on Artificial Intelligence (AAAI-12), Menlo Park, CA.

Bareinboim, E., Pearl, J., 2013. A general algorithm for deciding transportability of experimental results. J. Causal Inference 1, 107–134.

Bareinboim, E., Pearl, J., 2014. Transportability from multiple environments with limited experiments: completeness results. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc, pp. 280–288.

Bareinboim, E., Pearl, J., 2016. Causal inference and the data-fusion problem. Proc. Natl. Acad. Sci. Unit. States Am. 113, 7345–7352.

Cartwright, N., 1989. Nature's Capacities and Their Measurement. Clarendon Press, Oxford.

Deaton, A., Cartwright, N., 2018. Understanding and misunderstanding randomized controlled trials. Social Science and Medicine. https://doi.org/10.1016/j.socscimed.2017.12.005.

Pearl, J., 1995. Causal diagrams for empirical research. Biometrika 82, 669–710.

Pearl, J., 2015. Generalizing experimental findings. J. Causal Inference 3, 259–266.

Pearl, J., Bareinboim, E., 2011. Transportability of causal and statistical relations: a formal approach. In: Proceedings of the Twenty-fifth Conference on Artificial Intelligence (AAAI-11), Menlo Park, CA.

Pearl, J., Bareinboim, E., 2014. External validity: from *do*-calculus to transportability across populations. Stat. Sci. 29, 579–595.

Pearl, J., Glymour, M., Jewell, N., 2016. Causal Inference in Statistics: a Primer. Wiley, New York.

Pearl, J., Mackenzie, D., 2018. The Book of Why: the New Science of Cause and Effect. Basic Books, New York.

Shpitser, I., Pearl, J., 2008. Complete identification methods for the causal hierarchy. J. Mach. Learn. Res. 9, 1941–1979.

Tian, J., Pearl, J., 2002. A general identification condition for causal effects. In: Proceedings of the Eighteenth National Conference on Artificial Intelligence. AAAI Press/The MIT Press, Menlo Park, CA, pp. 567–573.