TECHNICAL REPORT
R-486
February 2019

On the Interpretation of $do(x)$

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024 USA
judea@cs.ucla.edu

February 13, 2019

Abstract

This paper provides empirical interpretation of the $do(x)$ operator when applied to non-manipulable variables such as race, obesity, or cholesterol level. We view $do(x)$ as an ideal intervention that provides valuable information on the effects of manipulable variables and is thus empirically testable. We draw parallels between this interpretation and ways of enabling machines to learn effects of untried actions from those tried. We end with the conclusion that researchers need not distinguish manipulable from non-manipulable variables; both types are equally eligible to receive the $do(x)$ operator and to produce useful information for decision makers.

Keywords: Manipulability, causal effects, interventions, testability, reinforcement learning

1 Introduction

The structural causal modeling (SCM) framework described in (Pearl, 1995, 2011, 2019) defines and computes quantities of the form $Q = E[Y|do(X = x)]$ which are interpreted as the causal effect of X on Y. The computation of Q simulates a minimally invasive intervention that sets the value of X to x, and leaves all other relationships unaltered. Several critics of SCM have voiced concerns about this interpretation of Q when X is non-manipulable; that is, X is a variable whose value cannot be controlled directly by an experimenter (Cartwright, 2007; Heckman and Vytlacil, 2007; Pearl, 2009; Hernán, 2016; Pearl, 2018). Indeed, asking for the effect of setting X to a constant x makes perfect sense when X is a treatment, say "drug-1" or "diet-2," but how can we imagine an action $do(X = x)$ when X is non-manipulable, like gender, race, or even a state of a variable such as blood-pressure or cholesterol level?[1]

[1]Some critics target both the non-manipulability of X and the ambiguity created when X is a "compound treatment" or a disjunctive proposition, as in "choose diet-1 or diet-2" (Hernán, 2016;

Mathematically, the expression $Q = E[Y|do(x)]$ (short for $Q = E[Y|do(X = x)]$), is perfectly well-defined when X is part of a causal model M, for it can be computed using the surgical procedure of the *do*-operator (Pearl, 2009, p. 24). Yet conceptually, Q raises two questions when X is a state of a variable. The first question is semantical: What information does Q convey aside from being a mathematical property of our model? Since one cannot translate Q into a prediction about the effect of an executable action, what does Q tell us about reality which is not just an artifact of the model? Take for example the proposition: "The number of variables in the model is a prime number"; it is undeniably a property of M, but would hardly qualify as a feature of reality. The second question raised is empirical: Even assuming that Q conveys an important feature of reality, how can we test it empirically? And if we cannot test it, is it part of science? I will address these two questions in the following sections.

2 The Semantics of Q

Assume we are conducting an observational study guided by model M in which Q is identifiable, and is evaluated to be

$$Q = E[Y|do(x)] = q(x)$$

where $q(x)$ is some function of x, computed from the joint distribution of observed variables in the model. To what use can one put this information? I will discuss three distinct uses.

1. Q represents a theoretical limit on the causal effects of manipulable interventions that might become available in the future.

2. Q imposes constraints on the causal effects of currently manipulable variables.

3. Q serves as an auxiliary mathematical operation in the derivation of causal effects of manipulable variables.

2.1 Q as a limit on pending interventions

Consider a set $I = \{I_1, I_2, \ldots, I_n\}$ of manipulable interventions whose effects on outcome Y we wish to compare. Assume that these interventions are suspected of affecting Y through their effect on X, and X is not directly manipulable. For example, $I_1, I_2, \ldots, I_n$ could represent names of different diets we wish to investigate as a means for lowering cholesterol levels $X = x$, while Y stands for "life expectancy." Some of these interventions will have side effects and some will not. Some will change X deterministically, such that $X = f(I)$, and some will affect X stochastically. The ideal intervention will, of course, have no side effect on the

Hernán and VanderWeele, 2011; Pearl, 2011). This paper focuses on the non-manipulability of X, while compound treatments are analyzed in (Pearl, 2017).

outcome Y and will affect X deterministically. However, an ideal intervention may not be feasible given the current state of technology, but may become feasible in the future. For example, cloud seeding made "rain" manipulable in our century, and genetic engineering may render gene variations manipulable in the future. If we simulate the impact of such an ideal intervention, one with no side effects and with a deterministic f, its resultant effect on Y will be Q.

Now suppose we manage to identify and estimate Q in an observational study. What does it tell us about the set of pending interventions $I_1, \ldots, I_n$? The answer comes in a form of a theoretical limit: Q gives us the ultimate effect ANY intervention can possibly have on Y by leveraging Y's dependence on X. This information may not be directly usable to a decision maker trying to assess the effectiveness of any given interventions I_i, but it could be extremely valuable to one who needs to decide whether to explore new interventions to achieve greater control on X. Clearly, if Q is low, the exploration is futile, while if Q is high, the possibility exists that by finding a more effective modifier of X, we would obtain better control over Y.

Note that Q can be considered a "theoretical limit" and an "ultimate effect"—not in the sense of presenting a ceiling on the impact of I_i on Y, but rather as a ceiling on the X-attributable component of that impact. If some intervention, say I_i, shows greater impact on Y than that predicted through Q, we can safely conclude that much of that impact is due to side effects, not due to I_i affecting X.

2.2 What Q tells us about the effects of feasible interventions

We will now explore how knowing Q, the "theoretical effect" of an infeasible intervention, can be useful to policy makers who care only about the impact of "feasible interventions."

Consider a simple linear model, $I \to X \to Y$ with no unmeasured confounders and no direct link from I to Y. Let a and b stand for the structural coefficients associated with the two arrows, and let X be non-manipulable.

If we wish to predict the average causal effect $ACE(I)$ of intervention I (say a new diet) on Y (say life expectancy), then we have (after proper normalization)

$$ACE(I) = E[Y|do(I+1)] - E[Y|do(I)] = a * b.$$

Thus, b constitutes an upper bound for $ACE(I)$. Yet, since X is not manipulable, the coefficient b is purely theoretical, and the manipulativity critics will object to granting it a "causal effect" status. Oddly, this theoretical quantity does inform our target quantity $ACE(I)$ which meets all criteria of feasibility and manipulativity. Practically, if for some reason we are able to estimate b, but not a, we have some extremely valuable information about the magnitude of $ACE(I)$. In particular, if b is close to zero, we can categorically conclude that $ACE(I)$ should be zero as well. Such a prediction would be critical, for example, if intervention I is still in its developmental stage, and our study involves measurement of a surrogate intervention I' yielding a' and b'. Our model dictates that the b' estimand under I' will remain

unaltered as we move to I. Therefore, estimating the theoretical quantity b' allows us to assess $ACE(I)$ from a study conducted under I'.

The basic structure of this knowledge transfer holds for nonlinear systems as well. For example, if the chain model above is governed by arbitrary functions $X = f(I, \epsilon_x)$ and $Y = g(X, \epsilon_y)$ (with ϵ_x independent of ϵ_y), the overall causal effect of I on Y becomes a convolution of the two local causal effects. Formally,

$$E(Y|do(I)) = \sum_x P(x|do(I))E[Y|do(x)].$$

Thus, we can infer the causal effect of a practical intervention I by combining the theoretical effect of a non-manipulable variable X, with the causal effect of I on X. Note again that if the theoretical effect of X on Y is zero (i.e., $E(Y|do(x))$ is independent of x), the causal effect of the intervention I is also zero.

Let us move now from the simple chain to a more complex model (still linear) where the arrow $X \to Y$ is replaced by a complex graph, rich with mediators and unobserved confounders. Linearity dictates that $ACE(I)$ will still be given by a product ac where a is the same as before and c stands for the difference:

$$c = E[Y|do(x+1)] - E[Y|do(x)].$$

Thus, whenever we are able to identify the theoretical effect $Q = E(Y|do(x))$ we are also able to identify the causal effect of the intervention I. This statement may appear to be empty when the latter is identifiable directly from the model. However, when we consider again the task of predicting $ACE(I)$ from a surrogate study involving I', the benefit of having $Q = E(Y|do(x))$ becomes clear. It is this theoretical effect that would permit us to transfer knowledge between the two studies.

To summarize these two aspects of Q, I will reiterate an example from (Pearl, 2018) where smoking was taken to represent a variable that defies direct manipulation. In that context, we concluded that "if careful scientific investigations reveal that smoking has no effect on cancer, we can comfortably conclude that increasing cigarette taxes will not decrease cancer rates, and that it is futile for schools to invest resources in anti-smoking educational programs."

2.3 *do*(*x*) as an auxiliary mathematical construct

In 2000 Phil Dawid published a paper entitled "Causal reasoning without counterfactuals" in which he objected to the use of counterfactuals on philosophical grounds. His reasons:

> "By definition, we can never observe such [counterfactual] quantities, nor can we assess empirically the validity of any modeling assumption we may make about them, even though our conclusions may be sensitive to these assumptions."

In my comment on Dawid's paper (Pearl, 2000), I agreed with Dawid's insistence on empirical validity, but stressed the difference between pragmatic and dogmatic empiricism. A pragmatic empiricist insists on asking empirically testable queries, but leaves the choice of tools to convenience and imagination; the dogmatic empiricist requires that the entire analysis, including all auxiliary symbols and all intermediate steps, "involve only terms subject to empirical scrutiny." As an extreme example, a strictly dogmatic empiricist would shun division by negative numbers because no physical object can be divided into a negative number of equal parts. In the context of causal inference, a pragmatic empiricist would welcome unobservable counterfactuals of individual units (e.g., $Y_x(u)$) as long as it leads to valid and empirically testable estimation of population effects. This is, indeed, the standard use of counterfactuals in the potential outcome framework (Rosenbaum and Rubin, 1983).

I now apply this distinction to our controversial construct Q which, in the opinion of some critics, is empirically ill-defined when X is non-manipulable. Let us regard Q—not as a causal effect or as a limit of causal effects—but as a purely mathematical construct which, like complex numbers, has no empirical content on its own, but permits us to derive empirically meaningful results.

For example, if we look at the derivation of the front-door estimate in *do*-calculus (Pearl, 2009, pp. 87–88), we can see how the operator $do(Tar)$ is used to derive the effect of smoking (assumed to be manipulable), though tar is non-manipulable. The term $do(Tar)$ enables us to apply new operations on, and new combinations of $do(Smoke)$ that eventually identify the causal effect of smoking on cancer and leaves the scene unscratched, as if tar has never been manipulated. This temporary violation of prudent empiricism is harmless, since it leads to empirically testable results, e.g., the effect of smoking on cancer.

Such auxiliary constructs are not rare in science. For example, although it is possible to derive De-Moivre's formula for $\cos n\theta$ using ordinary algebra, the derivation is immediate when we allow complex numbers and write $\cos\theta + isin\theta = \exp(i\theta)$. Indeed, complex analysis has since proven to be essential in many scientific fields—especially in engineering and quantum physics.

3 Testing $do(x)$ Claims

We are now ready to tackle the final question posed in the introduction: Granted that $Q = q$ conveys useful information to policy makers, how can we test it empirically?

Since X is non-manipulable, we must forgo verification of Q through the direct control of X, and settle instead on indirect tests as is commonly done in observational studies. This calls for devising observational or experimental studies capable of refuting the claim $Q(x) = q(x)$ and ascertaining that our data do not clash with this claim.

Since the claim $Q(x) = q(x)$ is a product of both the data and the modeling assumptions embedded in M, confirming the testable implications of those

assumptions constitutes a test for the equality $Q(x) = q(x)$.

Not all models have testable implications, but those that do advertise those implications in the model's graph and invite standard statistical tests for verification. Typical are conditional independence tests and equality constraints. For example, if $Q(x)$ is identifiable through the back-door criterion and there are several sets of covariates that satisfy the criterion, then equating the adjustment formulae generated by each of those sets provides a test for M, and hence a test for Q.

If the model contains manipulable variables, then randomized controls over the manipulable variables provide additional tests for the structure of M, and hence for the validity of Q. To illustrate, consider the front-door model of Fig. 1, where I is

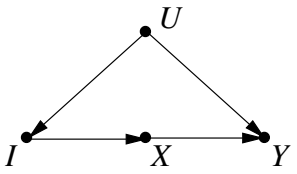

Figure 1: A model in which equating the effect of I on Y in an RCT with that obtained through the front-door formula produces a test for $Q(x)$.

manipulable, X non-manipulable, and U an unobserved confounder. The model has no testable implication in observational studies. However, randomizing I yields an estimate of $P(y|do(I))$, which should be equal to the estimand of $P(y|do(I))$ obtained through the front-door formula. (Pearl, 2009, pp. 81–83). Equating the two provides a refutable test for the assumptions embedded in the model, and hence for $Q(x)$, where

$$Q(x) = \sum_i P(y|x, I = i)P(I = i).$$

We see that, whereas direct tests of $Q(x)$ are infeasible, indirect tests are available, thus affirming the empirical content of Q. Metaphorically, these tests can be likened to the way planet Neptune was discovered (1845)—not by direct observation, but through the anomaly it caused in the trajectory of Uranus.

4 Non-manipulability and Reinforcement Learning

The role of models in handling a non-manipulable variable has interesting parallels in machine learning applications, especially in its reinforce learning (RL) variety (Sutton and Barto, 1998; Szepesvári, 2010). Skipping implementational details, a RL algorithm is given a set of actions or interventions, say $I = \{I_1, I_2, \ldots, I_n)$, and is required to find, for every observed state s of the environment an action I_k that maximizes the long-term reward achievable by acting I_k at state s. This reward function can be written as $E[Y|do(I_k), s]$, with Y the stream of future payoffs received by acting I_k.

Through trial and error training of a neural network, the RL algorithm constructs a functional mapping between each state s and the next action to be taken. In the course of this construction, however, the algorithm evaluates a huge number of reward functions of the form $E[Y|do(I_k), s]$ which, for a given s are very similar to the function $Q(x)$ that has been the focus of our discussion in this paper.

A question often asked about the RL framework is whether it is equivalent in power to SCM in terms of its ability to predict the effects of interventions.

The answer is a qualified YES. By deploying interventions in the training stage, RL allows us to infer the consequences of those interventions, but ONLY those interventions. It cannot go beyond and predict the effects of actions not tried in training. To do that, a causal model is required (Zhang and Bareinboim, 2017). This limitation is equivalent to the one faced by researchers who deny legitimacy to $Q(x)$ when X is non-manipulable. In the RL context, however, the prohibition extends to manipulable variables as well, in case they were not activated in the training phase.

A simple example illustrating this point is shown in Fig. 1, which depicts the causal structure of the environment prior to learning. X and Z are manipulable, while U_1 and U_2 are unobserved. Suppose we train a machine to learn the effect of manipulating Z on both Y and X. We now wish to infer the effect of action

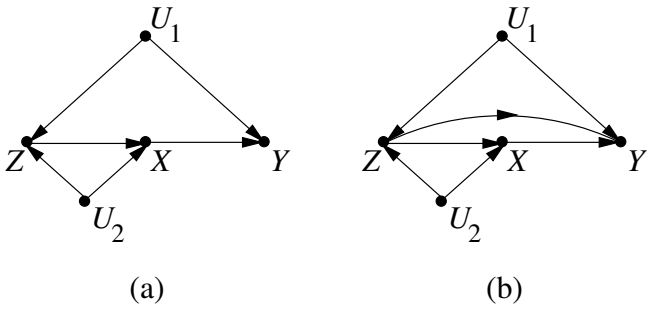

Figure 2: Model (a) permits us to learn the effect of X on Y by manipulating Z, instead of X. In Model (b) learning the effect of X on Y requires that X itself be manipulated.

$do(X = x)$ that was not accessible during training. Having a causal model, as in Fig. 2(a), the task can be accomplished through *do*-calculus (Bareinboim and Pearl, 2012, 2016), giving:

$$P(y|do(x)) = P(y, x|do(z))/P(x|do(z)).$$

Thus, the freedom to manipulate Z and estimate its effects on X and Y enables us to evaluate the effect of an action $do(X = x)$ which was never tried before.

To see the critical role that causal modeling plays in this exercise, note that the model in Fig. 2(b) does not permit such evaluation by any algorithm whatsoever, a fact verifiable from the model structure (Bareinboim and Pearl, 2012). This means that a model-blind RL algorithm would be unable to tell whether the optimal choice of untried actions can be computed from those tried.

Conclusions

We have shown that causal effects associated with non-manipulable variables have empirical semantics along several dimensions. They provide theoretical limits, as well as valuable constraints over causal effects of manipulable variables. They facilitate the derivation of causal effects of manipulable variables and, finally, they can be tested for validity, albeit indirectly.

Doubts and trepidations concerning the effects of non-manipulable variables and their empirical content should give way to appreciating the important roles that these effects play in causal inference.

Turning attention to machine learning, we have shown parallels between estimating the effects of non-manipulable variables and learning the effect of feasible yet untried actions. The role of causal modeling was shown to be critical in both frameworks.

Armed with these clarifications, researchers need not be concerned with the distinction between manipulable and non-manipulative variables, except of course in the design of actual experiments. In the analytical stage, including model specification, identification and estimation, all variables can be treated equally, and are therefore equally eligible to receive the *do*-operator and to deliver the ramifications in its effect.

Acknowledgments

Discussions with Elias Bareinboim contributed substantially to this paper. This research was supported in part by grants from Defense Advanced Research Projects Agency [#W911NF-16-057], National Science Foundation [#IIS-1302448, #IIS-1527490, and #IIS-1704932], and Office of Naval Research [#N00014-17-S-B001].

References

Bareinboim, E. and Pearl, J. (2012). Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence* (N. de Freitas and K. Murphy, eds.). AUAI Press, Corvallis, OR, 113–120.

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

Cartwright, N. (2007). *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press, New York, NY.

Dawid, A. (2000). Causal inference without counterfactuals (with comments and rejoinder). *Journal of the American Statistical Association* **95** 407–448.

Heckman, J. and Vytlacil, E. (2007). *Handbook of Econometrics*, vol. 6B, chap. Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation. Elsevier B.V., Amsterdam, 4779–4874.

Hernán, M. (2016). Does water kill? A call for less casual causal inferences. *Annals of Epidemiolgy* **26** 674–680.

Hernán, M. and VanderWeele, T. (2011). Compound treatments and transportability of causal inference. *Epidemiology* **22** 368–377.

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

Pearl, J. (2000). Comment on A.P. Dawid's, Causal inference without counterfactuals. *Journal of the American Statistical Association* **95** 428–431.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

Pearl, J. (2011). On the consistency rule in causal inference: An axiom, definition, assumption, or a theorem? *Epidemiology* **21** 872–875.

Pearl, J. (2017). Physical and metaphysical counterfactuals: Evaluating disjunctive actions. *Journal of Causal Inference,* Causal, Casual, and Curious Section **5**. Published online: <https://doi.org/10.1515/jci-2017-0018>.

Pearl, J. (2018). Does obesity shorten life? Or is it the soda? On non-manipulable causes. *Journal of Causal Inference,* Causal, Casual, and Curious Section **6**. Published online: <https://doi.org/10.1515/jci-2018-2001>.

Pearl, J. (2019). The seven pillars of causal reasoning with reflections on machine learning. Tech. Rep. R-481, <http://ftp.cs.ucla.edu/pub/stat_ser/r481.pdf>, Department of Computer Science, University of California, Los Angeles, CA. Forthcoming, *Communications of Association for Computing Machinery*.

Rosenbaum, P. and Rubin, D. (1983). The central role of propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press, Cambridge, MA.

Szepesvári, C. (2010). *Algorithms for reinforcement learning*. Morgan and Claypool, San Rafael, CA.

Zhang, J. and Bareinboim, E. (2017). Transfer learning in multi-armed bandits: A causal approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*. Minneapolis, MN.