TECHNICAL REPORT
R-484
September 2018

A note on oxygen, matches and fires

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

September 24, 2018

1 Introduction

This note illustrates the use of structural models in counterfactual reasoning. In particular, it demonstrates the computation of a quantity denoted *PS* – the *probability of sufficiency*, which plays an important role in commonsense reasoning, as well as in legal and medical applications.

To motivate the analysis, we will use the classical example of Oxygen, Matches, and Fire which *The Book of Why* (Pearl and Mackenzie, 2018, p. 289) describes as follows:

> "A fire broke out after someone struck a match, and the question is 'What caused the fire, striking the match or the presence of oxygen in the room?' Note that both factors are equally necessary, since the fire would not have occurred absent one of them. So, from a purely logical point of view, the two factors are equally responsible for the fire. Why, then, do we consider lighting the match a more reasonable explanation of the fire than the presence of oxygen?"

The intuitive explanation invokes the notion of *prevalence*, or anticipation:

> "The person who lit the match ought to have anticipated the presence of oxygen, whereas nobody is generally expected to pump all the oxygen out of the house in anticipation of a match-striking ceremony."

This intuition can also be captured by the notion of *sufficiency*; striking a match is more likely to be sufficient for the fire than the presence of oxygen. The language of counterfactuals permits us to make this distinction precise as follows: For any two variables, X and Y, define a quantity *PS* as "the probability that event $X = 1$ would be sufficient to producing outcome $Y = 1$." Using parenthetical counterfactual notation $Y(X = 1)$ to denote "the value that Y would attain had X been 1," *PS* can be written as (p. 289):

$$PS = P[Y(X = 1) = 1 | X = 0, Y = 0]. \tag{1}$$

In words, *PS* asks us to imagine a situation where $X = 0$ and $Y = 0$ and to test how likely it is for Y to turn into $Y = 1$ if X were to change (counterfactually) from $X = 0$ to $X = 1$. Eq. (1) thus quantifies the capacity of X to *produce* an outcome $Y = 1$ in situations where the outcome is absent. The reason that we must quantify this hypothetical event with probabilities is that both X and Y are random variables, subjected to the whims of unknown factors, some creating situations in which X produces Y, and some creating other situations where X does not produces Y. Eq. (1) quantifies production over all situations, weighted by their likelihood.

We are now going to compute *PS* for each of Oxygen and Match and compare their magnitudes. We start by specifying a structural causal model (SCM) for the variables

$$F = Fire, \quad M = Match, \quad OX = Oxygen$$

assuming that F responds to M and OX through the logical 'AND' function

$$F = \begin{cases} 1 \text{ if } OX = 1 \text{ and } M = 1 \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

Additionally, we assume that, prior to observing the fire, the probabilities for M and OX were:

$$P(OX = 1) = p_{ox} \quad P(M = 1) = p_m \tag{3}$$

with $p_{ox} \gg p_m$, since match-lighting is a rare event and the presence of oxygen is common.

We are now set to derive the probability of sufficiency (Eq. (1)) for both *Oxygen* and *Match* using a three-step procedure described in (Pearl and Mackenzie, 2018, p. 278). But before presenting this derivation, it is important that we step back and understand the significance of this exercise. Note that we are about to derive a counterfactual expression, Eq. (1), from a model that is totally void of such expressions. Instead, the model depicts the SCIENCE behind the fire story in the form of a Boolean function, Eq. (2), and two probabilities, Eq. (3), that can be estimated from the data. This stands in sharp contrast to conventional methods of estimating counterfactual quantities in the potential outcome framework which, invariably, start with counterfactual assumptions which are justified by drawing analogies to treatments assignments, or "well-defined" manipulations in controlled randomized experiments (Rubin, 1974; Robins, 1986; Angrist and Pischke, 2014; Imbens and Rubin, 2015; Morgan and Winship, 2015; Hernán and Robins, 2018).

There is nothing resembling treatments or experimental manipulation in the function of Eq. (2). One can, of course envision a variety of experiments on the process described in Eq. (2) but those would be conducted to interrogate the process, not to define it. The process itself is specified independently of any envisioned manipulations. See (Pearl and Mackenzie, 2018, pp. 144–150) for discussion of how experiments interrogate Nature, rather than define it.

This difference between causal models and manipulation-based models is essential for understanding the significance of the exercise described in this note. We will assess counterfactual quantities (Eq. (1)) directly from Nature (Eq. (2)) without asking an investigator to translate Nature into a set of counterfactual statements, prior to commencing the analysis. This we now demonstrate using the three-step procedure described in (Pearl and Mackenzie, 2018, p. 278). 1. Abduction, 2. Action, and 3. Prediction.

2 Formal Derivation

Problem: Compute $PS(M)$ and $PS(OX)$, where (from 1):

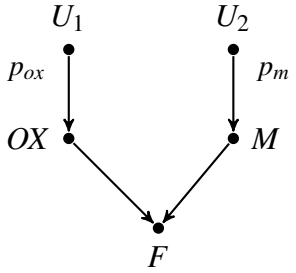
$$PS(OX) = P[F(OX = 1) = 1 | OX = 0, F = 0]$$

$$PS(M) = P[F(M = 1) = 1 | M = 0, F = 0]$$

Assumptions:

$$P(OX = 1) = p_{ox}, \quad P(M = 1) = p_m,$$

The model is given by the graph G below, where U_1 and U_2 represent unobserved factors which affect OX and M, respectively. For simplicity, we will assume these factors to be independent, as shown in Fig. 1.

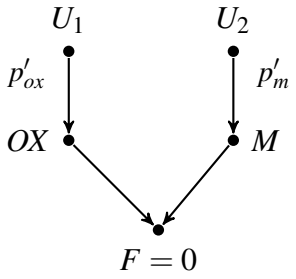


G

Figure 1

We shall now derive $PS(OX)$ and $PS(M)$ by applying the three-step algorithm to the model of Fig. 1.

1. **Abduction:** We need to update the prior probabilities p_{ox} and p_m in light of the evidence $F = 0$. This amounts to computing p'_{ox} and p'_m for model G, in which F is known to be False $(F = 0)$ (the situation prior to observing fire, as in Fig. 2.



G

Figure 2

Derivation:

$$\begin{aligned} p'_{ox} &= P(OX=1|F=0) = P(OX=1,F=0)/P(F=0) = \\ &= P(OX=1,M=0)/1-P(OX=1,M=1) \\ &= p_{ox}(1-p_m)/1-p_{ox}p_m \end{aligned}$$

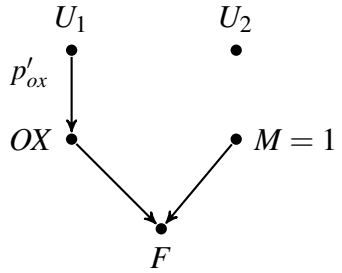
$$\begin{aligned} p'_{m} &= P(M=1|F=0) = P(M=1,F=0)/P(F=0) = \\ &= P(M=1,OX=0)/1-P(OX=1,M=1) \\ &= p_m(1-p_{ox})/1-p_{ox}p_m \end{aligned}$$

For $p_m \ll 1$ and $p_{ox} \approx 1$ we obtain:

$$p'_{ox} \approx p_{ox} \quad \text{and} \quad p'_m \approx p_m \tag{4}$$

The reason is clear; the updated priors are simply the old priors re-normalized, after excluding the event $F=1$, which is very rare. An identical result holds therefore when U_1 and U_2 are dependent (see Appendix I).

2. **Action:** To compute $PS(M)$, we take the updated model of Fig. 2 and simulate the action $do(M=1)$. This results in the graph $G_{M=1}$, of Fig. 3:



$G_{M=1}$

Figure 3

To compute $PS(OX)$, we simulate the action $do(OX=1)$, leading to graph $G_{OX=1}$, of Fig. 4.

3. **Prediction:** To complete the derivation of $PS(M)$, we now compute $P(F=1)$ in $G_{M=1}$, yielding:

$$\begin{aligned} PS(M) &= P(F=1) \text{ in } G_{M=1} \\ &= p'_{ox} \\ &\approx p_{ox} \end{aligned}$$

Likewise, to compute $PS(OX)$, we compute $P(F=1)$ in $G_{OX=1}$ giving:

$$\begin{aligned} PS(OX) &= P(F=1) \text{ in } G_{OX=1} \\ &= p'_{m} \\ &\approx p_{m} \end{aligned}$$

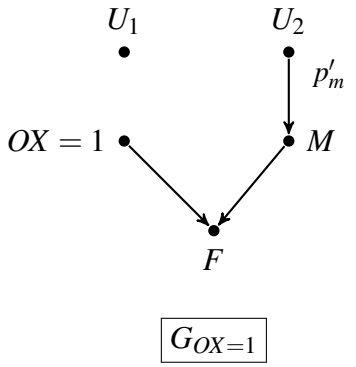


Figure 4

Thus, we have

$$PS(M) \approx p_{ox} \quad \text{and} \quad PS(OX) \approx p_m \tag{5}$$

and $PS(M) \gg PS(OX)$ as expected.

3 Conclusions and Related Works

The primary purposes of this note have been: (1) To demonstrate that counterfactuals are derivable algorithmically from common scientific knowledge, and are not needed as inputs for causal analysis. (2) To empower researchers with methods of estimating counterfactuals directly from functional description of their problems. We have demonstrated these two capabilities by computing *PS*, the probability of sufficiency, in the context of the classical Oxygen-Match-Fire example, which is pivotal for understanding causal explanations. Using this computation we obtained a formal confirmation of the intuition that lighting the match is the more plausible cause of the fire, not the presence of oxygen.

A brief historical overview of this problem and previous works towards its solution should help the reader appreciate its context and importance.

The most common conception of causation – that the effect E would not have occurred in the absence of the cause C – goes back to Hume (1748), and captures the notion of "necessary causation." The probabilistic version of necessary causation (*PN*) is behind many judicial standards. In tort law, for example, damage should be paid if and only if it is more probable than not that damage would not have occurred *but for* the defendant action.

But causation has two faces, *necessary* and *sufficient*. The distinction between the two was first articulated by John Stuart Mill (1843), and has received semi-formal explications in the 1960s, first using conditional probabilities (Good, 1961) and then using logical implications (Mackie, 1965). Both explications suffer from basic semantical difficulties (Kim, 1971; Pearl, 2000, pp. 249-256, 313-316). The popular "Sufficient Component" model of Kenneth Rothman (1976) is essentially equivalent to Mackie's "INUS condition" and inherits the semantical difficulties noted in (Kim, 1971). Nevertheless, the graphical schematics of Rothman's "causal pies" were found very effective in teaching epidemiologists how to represent interacting causes as Boolean functions in disjunctive form. Additionally, counterfactual interpretations of Rothman's model (VanderWeele and Hernán (2006) have resolved some of its semantical difficulties. In particular,

these interpretations restrict variables from entering the sufficient cause model unless they are parents of the outcome variable in the causal diagram, as depicted in Fig. 1.

Robins and Greenland (1989) gave a counterfactual definition for the probability of necessary causation taking counterfactuals as primitives, and assuming that one is in possession of a joint probability function over counterfactual events. Pearl (1999) gave definitions for the probabilities of necessary or sufficient causation (or both) based on structural model semantics which, as we have seen in this note leads to effective procedures for computing counterfactuals from a given causal theory (Balke and Pearl, 1994, 1995). Additionally, this semantics can be characterized by a complete set of axioms Galles and Pearl (1998); Halpern (1998), which can be used as inference rules in the analysis.

Pearl (1999) and Tian and Pearl (2000) have derived tight bounds on *PS* and *PN* when the model is only partially specified and confounding is present. A tool kit for solving counterfactual parameters is given in Pearl et al. (2016, pp. 116–126).

Our derivation of *PS* also bears on a recent debate concerning the role of non-manipulable variables in causal inference, specifically, whether variables such as sex or race can be considered "causes" (Hernán and Taubman, 2008; Pearl, 2018). In our example, oxygen is practically non-manipulable, and yet, the structural model of Fig. 1 treats oxygen and match on equal footing, with oxygen serving as an enabler of fire (see Appendix II). The model further allows for the estimation of the counterfactuals $PS(OX)$ and $PS(M)$ by the same three-step procedure, regardless of how manipulable they are. Such counterfactuals are considered "not well-defined" in the orthodox school of potential outcome, an untenable stance that would prohibit our question "what caused the fire" from being asked, let alone being answered.

References

ANGRIST, J. D. and PISCHKE, J.-S. (2014). *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.

BALKE, A. and PEARL, J. (1994). Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, vol. I. MIT Press, Menlo Park, CA, 230–237.

BALKE, A. and PEARL, J. (1995). Counterfactuals and policy analysis in structural models. In *Uncertainty in Artificial Intelligence 11* (P. Besnard and S. Hanks, eds.). Morgan Kaufmann, San Francisco, 11–18.

GALLES, D. and PEARL, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundation of Science* **3** 151–182.

GOOD, I. J. (1961). A causal calculus (I). *British Journal for the Philosophy of Science* **11** 305–318.

HALPERN, J. Y. (1998). Axiomatizing causal reasoning. In *Uncertainty in Artificial Intelligence* (G. Cooper and S. Moral, eds.). Morgan Kaufmann, San Francisco, CA, 202–210. Also, *Journal of Artificial Intelligence Research* 12:3, 17–37, 2000.

HERNÁN, M. A. and ROBINS, J. M. (2018). *Causal Inference*. Chapman & Hall/CRC, Boca Raton. Forthcoming.

HERNÁN, M. A. and TAUBMAN, S. L. (2008). Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal Of Obesity* **32** S8 EP. <http://dx.doi.org/10.1038/ijo.2008.82>.

HUME, D. (1748). *An Enquiry Concerning Human Understanding*. Reprinted Open Court Press (1958), LaSalle, IL.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, MA.

KIM, J. (1971). Causes and events: Mackie on causation. *Journal of Philosophy* **68** 426–471. Reprinted in E. Sosa and M. Tooley (Eds.), *Causation*, Oxford University Press, 1993.

MACKIE, J. L. (1965). Causes and conditions. *American Philosophical Quarterly* **2/4** 261–264. Reprinted in E. Sosa and M. Tooley (Eds.), *Causation*, Oxford University Press, 1993.

MILL, J. S. (1843). *System of Logic*, vol. 1. John W. Parker, London.

MORGAN, S. and WINSHIP, C. (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research*. 2nd ed. Cambridge University Press, New York, NY.

PEARL, J. (1999). Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese* **121** 93–149.

PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.

PEARL, J. (2018). Does obesity shorten life? Or is it the soda? On non-manipulable causes. Tech. Rep. R-483, <http://ftp.cs.ucla.edu/pub/stat_ser/r483.pdf>, Department of Computer Science, University of California, Los Angeles, CA. Forthcoming, *Journal of Causal Inference*.

PEARL, J., GLYMOUR, M. and JEWELL, N. P. (2016). *Causal Inference in Statistics: A Primer*. Wiley, New York.

PEARL, J. and MACKENZIE, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York.

ROBINS, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. *Mathematical Modeling* **7** 1393–1512.

ROBINS, J. M. and GREENLAND, S. (1989). The probability of causation under a stochastic model for individual risk. *Biometrics* **45** 1125–1138.

ROTHMAN, K. J. (1976). Causes. *American Journal of Epidemiology* **104** 587–592.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66** 688–701.

TIAN, J. and PEARL, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence* **28** 287–313.

VANDERWEELE, T. J. (2017). Invited commentary: The continuing need for the sufficient cause model today. *American Journal of Epidemiology* **185** 1041–1043.

VANDERWEELE, T. J. and HERNÁN, M. A. (2006). From counterfactuals to sufficient component causes and vice versa. *European Journal of Epidemiology* **21** 855–858.

Appendix I: The Importance of the Abductive Step

The infinitesimal probability of no oxygen, $p_{ox} \ll 1$, led to the equality

$$p'_{ox} \approx p_{ox} \quad \text{and} \quad p'_m \approx p_m$$

which may give readers the impression that the abduction step is superfluous, and that we could have gotten Eq. (5) directly, by computing the causal effects $P[Y(OX = 1) = 1]$ and $P[Y(M = 1) = 1]$ instead of Eq. (1). Indeed, intervening to secure oxygen in the house will have very low probability p_m of resulting in fire, and intervening to light a match will result in fire with high probability, p_{ox}. To appreciate the importance of the abduction step: let us compute *PS* for a hypothetical scenario in which p_m and p_{ox} are determined by two independent fair coins, resulting in $p_{ox} = p_m = 1/2$.

The causal effects in this case would compute to

$$P[Y(OX = 1) = 1] = P[Y(M = 1) = 1] = 1/2 \tag{6}$$

because once we assure the presence of oxygen fire will occur 50% of the time, when a match is struck. Conversely, once a match is struck, fire will occur 50% of the time, when oxygen is present.

However, the probability of actually producing fire in situation where fire is initially absent is in fact lower than $1/2$. Going through the abduction exercise, we get

$$p'_m = P(M = 1|F = 0) = P(U_2 = 1|U_2 = 0 \text{ or } U_1 = 0) = 1/3$$
$$p'_{ox} = P(OX = 1|F = 0) = P(U_1 = 1|U_2 = 0 \text{ or } U_1 = 0) = 1/3$$

and, accordingly, the probabilities of sufficiency become:

$$PS(M) = PS(OX) = 1/3$$

lower than the causal effects in (6).

Much wider difference between p_m and p'_m will obtain if we let U_1 affect U_2 in a significant way. For example, let U_1 be a fair coin and let U_2 track U_1. The marginal probabilities of OX and M will remain the same, $p_m = p_{ox} = 1/2$. and the causal effects, likewise, will be the same

as in Eq. (6). However, the posterior probabilities will be vastly different, yielding $p'_m = p'_{ox} = 0$, because both $M = 1$ and $OX = 1$ must be false in any situation where $F = 0$. Accordingly, the probabilities of sufficiency must both vanish

$$PS(M) = PS(OX) = 0$$

as we can see from Figs. 3 and 4, using $p'_m = p'_{ox} = 0$. Indeed, prior to the fire, either U_1 or U_2 must be absent, but since they track each other, both must be absent, so lighting a match will not trigger a fire. The abductive stage is crucial for extracting this information before envisioning interventional scenarios.

This example demonstrates another feature of SCM. It contains more information than its potential outcome ramifications. VanderWeele (2017) called it "many to one" mapping; see also (Pearl, 2000, p. 35). If we vary the dependence between U_1 and U_2 PS would vary, but the causal effects may remain unaltered. The latter depends solely on the marginal distributions $P(M = 1)$ and $P(OX = 1)$.

Note however that the extra information that enables us to compute PS requires the specification of the functional relationships between the variables involved (as in Eq. (2)) as well as the distribution of the unobserved error terms U_1 and U_2. These two specifications elevates SCM to the top level (rung 3) of the Ladder of Causation (Pearl and Mackenzie, 2018) which supports counterfactuals. Lacking any of these two, as in the potential outcomes framework (or in Causal Bayesian Networks (Pearl, 2000, Sec. 1.3.1)) may allow us to evaluate causal effects, but not counterfactuals. Tyler VanderWeele (2017) stresses this distinction succinctly: "The potential outcomes framework considers the effects of causes, whereas the sufficient outcomes framework considers the causes of effect." To this one must add a correction; Rothman's sufficient outcomes framework in itself, lacking the distribution of the error terms, does not allow the evaluation of causes of effect. A full SCM specification is needed for the task, which includes both, Eq. (2) and the distribution of the U terms.

Appendix II: Causes vs. Enablers

Epidemiologists reading this note may complain that the analysis of PS may confer causal power onto variables that are merely "effect modifiers" but not genuine "causes." Indeed, in ordinary epidemiological conversions oxygen would be classified as an effect modifier, not as a cause of fire. So will variables such as humidity, atmospheric pressure and wind velocity. They are perceived to be assisting or hindering the fire, not causing it. From a chemical viewpoint however the opposite is true; fire is a process of oxidation, hence oxygen is an active agent in the process, while match striking merely creates a local rise in temperature which is an enabling condition, not an active cause of fire. If we further look at the logical function defining the process, Eq. (2), we find total symmetry. Moreover, examining Rothman's "pie diagrams" which many epidemiologists consider a faithful depiction of their conceptual framework, we find each of Match and Oxygen labeled a "sufficient cause component" in a 2-component pie

$$\{Oxygen, Match\}.$$

What then governs the distinction between "cause" and "effect modifier" or "enabler" in epidemiology? Is it the manipulability of the former, or the higher PS measure that the former

earns from prevalence considerations? I believe both considerations contribute to the distinctions and, certainly, we should not refrain from calling a nonmanipulable effect modifier "a cause," if its *PS* value justifies the name.

Effect modifiers, contrary to opinions of some epidemiologies (Hernán and Taubman, 2008) do have well defined causal effects, defined by the *do*-operator and the model in which they are embedded. The same goes to notions such as confounding and mediation. Whatever property the model bestows upon a manipulated variable it also bestows upon an effect-modifier, since the two are not marked differently in the model. The interpretation of such causal effects may not translate into policies that directly manipulate these modifiers, yet they enter the evaluation of policies that control the presence of these modifiers so as to regulate their consequences (Pearl, 2018).

Lastly, it is interesting to note that the capacity of an event $X = 1$ to produce an outcome $Y = 1$ can be uncovered directly from the structural equation model. We can proclaim $X = 1$ a "producer" of $Y = 1$ iff there exists a context C such that

$$Y(X = 0, C) = 0 \quad \text{and} \quad Y(X = 1, C) = 1.$$

For example, each of $M = 1$ and $OX = 1$ is a producer of $F = 1$ in the model of Eq. (2), because $OX = 1$ serves as an enabling context for $M = 1$, and $M = 1$ serves as an enabling context for $OX = 1$. Events $M = 0$ and $OX = 0$ cannot be producers of $F = 1$ since no enabling contexts exist.

One may be tempted to surmise that the property of production coincides with the presence of an event as a component in Rothman's "sufficient component model." But this is not the case. Consider the 3-pie model:

$$\{A = 0, B = 1, C = 1\}, \{A = 1, B = 1\}, \{A = 1, B = 0, C = 0\}$$

Event $A = 0$ appears in the first pie and, yet, it is not a producer of $Y = 1$ because no context exists which would make Y switch from 0 to 1 as A switches from 1 to 0. The same is true for $B = 0$ which appears in the 3rd pie. All other events however are producers of $Y = 1$. For example, $C = 0$ is a producer of $Y = 1$ because the context $\{A = 1, B = 0\}$ will see Y switch from $Y = 0$ to $Y = 1$ as C changes from $C = 1$ to $C = 0$.