

THE CREDENTIALS OF SCIENTIFIC EVIDENCE: REBALANCING THE EPISTEMOLOGICAL SCALES

9.2.7 Pearl: Causal inference and the 'do-operator'

Abstract

This last commentary in the series gives a brief introduction to Pearl's 'causal inference' theory and provides specific examples of how it can be applied to a trial of psychotropic drugs. This is necessarily a simplistic example of the application of Pearl's complex theory. The first example is a simple explanation involving a model illustrated by the effects of intravenous phentolamine on blood pressure. The second is an example of a hypothetical test of two TCA 'antidepressants', doxepin and nortriptyline, illustrating how Pearl's methodology would distinguish between the 'antidepressant' effects of nortriptyline and doxepin, which both produced similar changes in the HDRS. It takes account of Bayesian prior probability, is based on knowledge of pharmacology and mechanisms, and illustrates the power of the *do*-operator by manipulating the putative causal variable and measuring a specific objective endpoint, thereby establishing a strong cause-effect relationship.

Comment about frequentist statistics and P values is reiterated in the context of Pearl's theories, although it is dealt with more fully in other commentaries, especially (9.2.3).

This series of commentaries results in the conclusion that Pearl's causal inference theory offers promising advantages for clinical trial work and that clinical practice experimentation is epistemologically of superior value and validity to the results of RCTs, particularly at the level of individual patient response, which is the bread-and-butter of clinicians' everyday practice.

Introduction

Judea Pearl has developed the theory of causal inference, for which he won the prestigious Turing Award¹ in 2011. Judea Pearl has stated, as did Einstein, 'Science is nothing without causality' — but RCTs contribute little or nothing to understanding causality.

Pearl [1, 2]:

It is impossible to deal with causal relationships with statistical language.' ... there is no way of doing science without causality, it is the sine qua non for all understanding and progress

The entire story of causality unfolds from just three basic principles:

- (1) causation encodes behaviour under interventions
- (2) interventions are surgeries on mechanisms
- (3) mechanisms are stable functional relationships

Pearl goes on to say:

¹ Often referred to as the Nobel Prize in computing

A mantra most scientists can recite in their sleep is that correlation does not imply causation; but they do not grasp the depth of it² ... Causality is the key: there is no way of doing science without causality, it is the *sine qua non* for all understanding and progress³

It further follows that since causality is the key, Bayesian reasoning comes inevitably into play when considering methodology: that is because the prior probability of something being correct depends on pre-existing knowledge, in this case, of ‘nature’, evolution, science, physiology, biochemistry, pharmacology — such knowledge tells us, with varying probabilities of being correct, what is likely to work and what is not likely to work.

Highly trained scientists, enormous sums of money, and time are essential requisites for scientific research. These are in limited supply. It is therefore incumbent on us to deploy them efficiently and rationally. If an applicant sought funds to do a trial it would receive scant support unless there was some pre-existing rationale to support its usefulness. Thus, a trial to test vitamin C in cancer would fall at the first hurdle. After all, we have only so much time and so much money, decisions must be made about the most efficient way of using those resources. Consciously or unconsciously, the grant assessors would be using Bayesian prior probability to decide that vitamin C trial was a waste of resources. Not only that, even if someone did such a trial and claimed results with a significance of P 0.001, few scientists would give it credence. They would assume error, bias, or dishonesty — again, consciously or unconsciously, Bayesian prior probability would be in play. Such is perhaps the simplest illustration of the absurdity of Fisherian frequentist statistics and P values, when they are used divorced from a proper scientific foundation of cause-mechanism-effect, as they so frequently are.

Karl Pearson, English mathematician (Pearson’s correlation coefficient) and prominent eugenicist⁴, was succeeded by Fisher, also a eugenicist; both treated cause as illegitimate, and believed **the mathematics of correlation was the grammar of science** — which led them both to the conclusion that other races were less intelligent, and to give plausibility to the notion of telekinesis. Epic method, epic success!

Until one appreciates the absurdity of failing to recognise causality as being an essential element of the scientific method, one will not be able ‘do science’ — I am with the famous Princetown statistician Tukey (see 9.2.6) in regarding Fisher as representing the naïve infancy of statistics. Another famous statistician Meehl [3]) suggested Fisher had lead scientists ‘**down the Primrose path**’ (that expression is not much use now and in fact means the opposite of what Meehl obviously intended to say). He meant ‘**up the garden path**’, **which means what he intended, i.e. to deceive or mislead.**

Pearl’s causal inference theory can be regarded as ‘the mathematics of common sense’: much follows from this understanding. The ‘Fisherian’ view is that statistics and correlation are the grammar of science, and that is antithetical to Pearl’s thesis that, ‘**It is impossible to deal with causal relationships with statistical language.**’ Perhaps this understanding reveals the kernel of truth, expressing the widespread (justified) ‘common-sense’ scepticism about statistics, contained in the old quotation, variously attributed (I can certainly picture Mark Twain saying it⁵):

² I would add that, whilst many recite that mantra, they conspicuously fail to acknowledge its implications, or modify their speculations and conclusions accordingly

³ The meagre contribution of RCTs to Scientific progress is an eloquent testimony to that fact

⁴ It may be relevant that he changed the spelling of his forename from ‘Carl’ to ‘Karl’ after spending time in Germany, with which he was enamored, and this may relate to his eugenic tendencies

⁵ Also attributed to Disraeli

| There are three kinds of lies: lies, damned lies, and statistics

Pearl's thesis clearly contradicts Fisher's assertion: but remember, that assertion, and the turbulent stream of mis-thinking that yet flows from it, has been the major contributor to the ubiquitous P-value chaos to which we have been subjected by 'medical scientists' for many years. One may also note here that the statistics in most medical research is neither conceived by, nor carried out by, expert statisticians, nor does the peer review process usually involve review by a statistician.

Pearl, again:

| The central theme is to view causality as a computational scheme devised to facilitate prediction of the effects of actions.

Pearl introduced a new mathematical language of 'doing' (*do*[x]), allowing us to better distinguish between cause and effect — intervening on the cause will change the effect, but any intervention on the effect will not change the cause. His genius, and the reason he got the Turing prize, was to be able to put this, and much more, into **exact mathematical language and causal diagrams**. The 'post hoc, ergo propter hoc' trope is of the cockerel/rooster crowing causing the sun to rise... strangle the bloody cockerel: the sun still rises. Deceive the cockerel into crowing earlier or later, the sun still rises at the same time...

That is the essence of causality and put like that it is a statement of the [blindingly] obvious, but nonetheless it is one that frequentist statisticians, having been led up the garden path by Pearson and Fisher (Meehl [3]), have long failed to properly appreciate.

Causality is something that those statisticians evinced no consideration of — indeed, books and papers on statistics that include the word 'cause' are rare as hen's teeth, Pearl states that there are none.

Pearl [4]:

| It is critical to realize that data are profoundly dumb about causal relationships. We will also explain why any query about the mechanism by which causes transmit their effects—the most prototypical "Why?" question—is actually a counterfactual question in disguise. **Statistics inflicted causal blindness on itself. Bayesian networks remain a very important tool for AI and encapsulate much of the mathematical foundation of causal diagrams.** The smoking debate brings the importance of causality into its sharpest focus. To understand the world of causes and effects, you need to combine models and data, a rather neglected exercise in mainstream statistics.

'Statistics inflicted causal blindness on itself' — what a perspicacious comment

Also note this key statement: **'Bayesian networks remain a very important tool for AI and encapsulate much of the mathematical foundation of causal diagrams'**.

We now come back to the old trope, alluded to above, of **Post hoc ergo propter hoc**, especially because it is usually advanced and cited as an example of a logical fallacy: we also come back to it because it is a key factor and element in clinical practice. It is not necessarily or invariably a logical fallacy. It is only a logical fallacy to those ignoramuses who do not understand prior probability, causality, Pearl's '*do*-operator', mechanisms, and the power of manipulating putative causal factors.

Strangle the bloody cockerel

One key part of Pearl's idea is the '*do*-operator' [4] — strangling the cockerel — an idea that formalises a method of simulating a randomised control trial using

only observational data, and that allows causal connections to be extracted, not just correlations.

This represents the mathematics of common sense.

The Bayesian notion of taking a prior probability into account when planning and interpreting an experiment involves pre-existing knowledge: that often provides a plausible mechanistic model behind the experiment in question, and that is essentially a prior probability — if you cannot construct a plausible hypotheses indicating why your idea might be true, then it is unlikely there will be any useful purpose to your ‘trial’, or that your experiment will have any heuristic utility, or provide evidence that it is ‘true’. I used to refer to this as ‘so-what’ research?

There you have it, Pearl’s *do*-operator, and a causal mechanism leading to a measurable outcome. That is science, real science.

Pearl:

You can’t have a method to reach causal claims without causal assumptions, and since we use these causal assumptions every day, we need an efficient and formal language to encode and discuss them. Please inspect your most favourite statistics textbook; does it have ‘causal’ in the index? ... **Observational studies have a definite advantage: they interrogate populations at their natural habitats, not in artificial environments choreographed by experimental protocols.**

Additionally, further developments by Pearl in his paper *Causal inference and the data-fusion problem* facilitate improvements in the conclusions that can be extracted from meta-analysis [5].

There is an important reason why RCTs, comparing one drug to another, are of restricted scientific usefulness and validity — **they are rarely constructed to answer causal and mechanistic questions; indeed, it is frequently not even possible to so-construct them** — and that is especially true when an associated mechanism is ill-defined and heterogenous.

You cannot have a method to reach causal claims without causal assumptions (Pearl)

Pearl’s ‘*do*-operator’ in clinical practice

My inchoate understanding of the complexities of Pearl’s methodologies notwithstanding, I will give an example.

Our first example is with intravenous drugs to lower blood pressure, and one can then extrapolate to the timing of the improvement of symptoms after ECT, MAOIs, etc.

We know about the pharmaco-kinetics of phentolamine when injected intravenously, and about its mechanism of action on alpha receptors, and we have a clear idea of the timescale over which those mechanisms produce a change in blood pressure — all as result of our scientific knowledge gained from physiology and experimental pharmacology (which allow us to construct and estimate a prior probability).

We inject phentolamine, we observe serial measurements of the reduction in blood pressure; we expect these to occur within five minutes, and we know the duration of the drug action is about thirty minutes. We expect this because of our prior knowledge of physiology and pharmacology, and this constitutes the putative mechanism, which influences our estimation of the prior probability, cf. Pearl, **You cannot have a method to reach causal claims without causal assumptions.**

We set up a saline infusion and in iterations of the experiment we inject the drug after the saline infusion has been running for, 5 minutes, 10 minutes, 20 minutes **We observe the time interval between administration and the change in blood pressure.** We expect the interval between the administration of the active drug and the reduction in blood pressure, and the duration of action, to remain the same, irrespective of whether it is given 5 minutes or 30 minutes after the start of the infusion.

The key concept is a predefined stimulus, a causal mechanism with a specified time of onset, magnitude, and duration, and a specific predefined, objectively measurable, effect.
That establishes the cause–mechanism–effect relationship

The timing of the administration of the drug is the ‘do-operator’. If it is causal in relation to the change we are measuring, and the mechanism remains the same, which we are justified in assuming it would, then the time interval will remain the same. We are fixing other variables and changing the application of the proposed causal agent. If the time relationship between administration and blood pressure change remains constant within the variation expected and predicted in such experiments, then we have established a strong causal link.

In principle we can establish the same *modus operandi* for looking at the rate of improvement of pre-specified signs (e.g., psycho-motor retardation) and symptoms after ECT or MAOIs, in relation to when the treatment is started. We can thus establish a cause-effect relationship between the administration and its effect.

If there is a specific mechanism, then the predicted changes must occur within a **pre-specified timeframe** in order for the cause–effect connection to be established

Over many decades clinicians have done exactly this ‘inadvertently’: for instance, when titrating the ECT stimulus that actually causes EEG activity and observing that until it is sufficient to produce a fit one cannot expect improvement [6]. I did an analogous experiment repeatedly in the days when TCAs were the main treatment. I swapped patients from non-SNRI-TCAs to clomipramine and observed an improvement within two weeks. It is crucial to recognise that this scenario fulfils the criteria described in this series for epistemological validity. Patients with a stable long-term illness (having failed to respond to a therapeutic course of the previous antidepressant) thereby avoid the RTM issue (cf. McDonald [7]). It also minimises the possibility of spontaneous remission, since the chance of that occurring during the two weeks after the change of treatment, in a patient who has been ill for months, or even years, is low. **Such a clinical experiment scenario fulfils the criteria of good cause-effect relationship science better than an RCT — it therefore has greater epistemological validity and allows clinicians to ascribe greater confidence to their clinical experience.**

Lasagna, referred to by some as ‘the dean of American pharmacology’, consulted for the FDA in the 1960s, also suggested [8, 9], in what we must consider an impressive example of prescience, precisely that:

we had best gird our loins and work on the methodology for validating naturalistic clinical experience

this comment was in tongue-in-cheek article entitled *Clinical trials of drugs from the viewpoint of the academic investigator (a satire)*[9].

Lasagna further observed [8]:

There may be a good deal of clinical experience suggesting that your hypothetical drug does work. I am not willing to throw out a lot of naturalistic experience on the basis of one or two negative double-blind trials. I have seen too many negative double-blind trials. . . . There are a number of drugs for which we don't have double-blind control placebo tests, for instance, digitalis, antiepileptic drugs, antibiotics, and anti-Parkinsonian drugs. In these cases, the medical profession and the academic experts have decided, "I have enough feeling for this drug on the basis of what I have seen in ordinary 'uncontrolled' clinical experience to conclude that it is O.K.

An understanding of these principles helps one to see why Lasagna's perspicacious comment 50 years ago was bang on the nail, decades before Pearl's causal inference formulas were developed.

This is a lightbulb moment, because it tells you that informed clinical experience utilises Pearl's do-operator much of the time — that is precisely why it is often more informative than an RCT, and more valid than an RCT. Utilising such causal inference means that each instance of an experiment is its own control: that, at a stroke, removes many of the complications of RCTs, both ethical and practical. It also negates the concept of the placebo response (cf. 9.2.5), whatever that is conceived to be, and neutralises its relevance.

A Model example: Bayes and Pearl's *Dox*

In this (not so) hypothetical model we have patients with anxiety and depression, and we cannot ascertain accurately which of those two is 'primary'. We use two different drugs to effect an improvement in their symptomatology, which is measured by a rating scale that does not distinguish definitively between symptoms that are related to anxiety as opposed to depression — in other words the scores of these items are significantly conflated (the HDRS fits that description). The two so-called antidepressant drugs we test are doxepin and nortriptyline⁶; these have quite different mechanisms of action, one being a selective histamine H1 antagonist (doxepin) and the other a selective re-uptake inhibitor of noradrenaline (NRI). They both produce a small improvement in the rating scale score. However, it is not possible to make a deduction about either, 1) the role of these drugs in the improvement of the illness, or 2) the mechanism by which they might affect that, or 3) if they have affected the core changes of the illness, rather than only ameliorating symptomatology.

A Bayesian would assess that the prior probability of the two drugs being equally effective is low and, therefore whatever any P-value is, as traditionally calculated, it should be modified by this prior probability. Our Bayesian would also point out that basic science and pharmacology indicates that histamine probably is not involved in the fundamental pathology of depression and that H1 antagonists are hypnotic and anxiolytic; therefore, their effect is probably due to those properties. Like the effect of phentolamine, this sedation appears as soon as effective plasma levels are achieved (viz. 2-4 hrs).

A Bayesian would also suggest that boosting noradrenaline probably helps depressive illness and that this takes some days, rather than hours, as informed by basic science. It may also be changing different symptoms, perhaps those more closely related to the core symptoms of depression (viz anergia and an anhedonia). They might also suggest that the timing (rapidity) and type of symptom improvement should be examined more carefully (which indeed it should have been).

⁶ DMI is an even better Choice, as in the Amsterdam reference, because it is a more specific NRI.

We know that H1 antagonism is an immediate effect which causes sedation and that NRI effects are not immediately apparent but may be apparent after a week or two.

That is where Amsterdam's [10] paper from over 40 years ago is revealing, even if that was not understood at the time. They demonstrated that in a comparison of doxepin vs DMI, doxepin worked more quickly, probably because of greater sedative effects. However, since the precise timeline of improvement, and the separation of distinct improvements in different psychopathological domains of symptoms, were not incisively examined, no more useful conclusions can be drawn. Nevertheless, this serves as an exemplar of how a 'Pearl-inspired' investigation could proceed.

The Bayesian notion of prior probability, which has been causing statisticians with a frequentist focus to foam at the mouth for so many years, is straightforward. It is analogous to the calculation of the predictive probability of a blood test indicating the presence of an illness; and that in turn is dependent on the prevalence of the condition in question in the index population — that is equivalent to a Bayesian prior probability: see this reference for a refresher course about that [11]. Statistical frequentists do not seem to object to this, since they regard the prevalence in the population as reliable data, rather than a Bayesian prior probability. That illustrates their 'naivety', as Tukey described Fisher's approach, because the susceptibility of various sub-groups in that society to one or other illnesses may well vary, so the prevalence of the condition in the population is a Bayesian prior probability, not 'pure' data. Indeed, as Tukey would have said to Fisher, there is no such thing as pure data, all data comes with presumptions and prior probabilities.

There is no such thing as pure data, all data comes with presumptions and prior probabilities

If one treats the improvement in the rating scale score as a true reflection of antidepressant response then the Bayesian will interpret the value in relation to doxepin differently because of the assumption that the prior probability of that being an antidepressant is lower than that of nortriptyline. If the prior probability for doxepin was 1/20 and nortriptyline was 1/2, then the respective modified values at P 0.05 would convert doxepin to a P value of 1.0 and NTP to 0.1.

Now we introduce a little more science, relating to pharmacology and mechanisms; we know that the effect of drugs blocking post-synaptic receptors is rapid, as for example the effect of a benzodiazepine, or a sedative antihistamine. On the other hand, we know that the effect of re-uptake inhibitors is delayed. Thus, we hypothesise that the measurable symptom changes induced by these drugs will show a different rate of change and time course. If we design our experiment to dissect more finely the rate of change of different symptoms over time, then we would expect the effects of an H1 antagonist to be observable more quickly than those of a re-uptake inhibitor.

And there you have it, Pearl's *do*-operator, and a mechanism leading to a measurable outcome. That is science, real science.

Summary and conclusions

Much of medical research and clinical-trial work has been blighted by a blinkered approach centered on RCT-based drug efficacy-testing, which has come to dominate both research activity and the literature; that is because of the massive financial influence of pharmaceutical companies, whose objectives and priorities

are different to those of a treating clinician. Considering the broad sweep of scientific investigation, RCTs have a narrow and restricted role which is negatively counterbalanced by their resource-consuming inefficiency and the seriously misleading results they have frequently produced, which are sometimes forgotten or minimised by RCT advocates. Such advocates are at risk of being excessively influenced by the sunk cost fallacy — the cost of RCTs involving psychotropic agents over the last few decades has indeed been enormous, estimated in tens of billions of dollars in the last decade alone. Such expenditure requires sound justification, which simply does not exist.

Pearl himself states that Bayesian prior probability is the first step on the road to causal inference and that Bayesian networks incorporate some of the concepts that are elaborated in his theory of causal inference:

Bayesian networks remain a very important tool for AI and encapsulate much of the mathematical foundation of causal diagrams

The enduring failure of frequentist statisticians to acknowledge the role of both prior probability and causality is a crippling legacy of Fisher's influence — Pearl described it as:

statistics inflicting causal blindness on itself

Pearl's causal inference theory represents a major advance in epistemological thinking which offers a way forward that has considerable advantages. His 'do-operator' allows individual patient-level data to be extracted from clinical practice experimentation and enables causal inferences to be made concerning treatment effects — the advantages of thereby making redundant placebo effects, randomisation, stratification, time, expense, ethical difficulties, etc. are dramatic and clear. As Pearl himself puts it:

It allows us to interrogate populations in their natural habitats, not in artificial environments choreographed by experimental protocols

A key concept is a predefined stimulus and a causal mechanism with a specified time of onset, magnitude, and duration, and a specific predefined, objectively measurable effect (cf. phentolamine example above). That establishes the cause-mechanism-effect nexus, thereby imbuing it with the power of real science.

A practical example of this, directly relevant to clinical practice, and observable by any practicing clinician, is the power to better separate the different effects of drugs on illness symptoms — as in the example of doxepin versus nortriptyline given above — which demonstrates the power of 'mechanism-orientated' thinking combined with causal inference.

The hegemony of RCT-practice has eclipsed not just Bayes and Pearl, but other important approaches and techniques: it has stunted mechanism-orientated and cause-orientated scientific investigation — since cause and mechanism are the essence of science, that is a consequential and far-reaching misdirection of resources and effort. Much scientific work has become divorced from the fundamentals of medical science and from the objectives of investigating causes and effects by experiment, which is the *sine qua non* of serious science.

This has influenced a progressive insidious denigration of the value and validity of clinical science practice and experimentation with individual patients — clinicians have been deceived, some might say bullied, into thinking that their real-world clinical experience is of inferior value, when in fact it is more valuable than the results of RCTs.

Tukey's view that frequentist statistics represent a superficial and flawed approach to investigating hypotheses has subsequently been endorsed by a legion of influential thinkers — it must be given careful consideration, especially

since it has not yet adequately permeated the corpus of medical thinking. The misuse of, and misunderstandings associated with, P-values has been of epidemic proportions — P-values have been responsible for a plague of nonsensicality and second-rate thinking, exemplified by papers that claim a significant P value for demonstrating the existence of irrational and unscientific notions (e.g., homeopathy, telepathy) contained in the maelstrom of misleading papers that have overwhelmed the literature, and drowned out that which is good. Frequentist statistics and P-values are so disconnected from science and reality as to be effectively useless — they frequently produce deceptive or meaningless results, a fact that was recognized by the early investigators like Lasagna in the FDA way back in the mid 1960s.

Anything that can be done to speed the adoption and acceptance of mechanism-orientated and cause-orientated scientific investigation utilising Pearl's causal inference theory will be a great service to medical science.

References

1. Pearl, J., M. Glymour, and N.P. Jewell, *Causal inference in statistics: A primer*. 2016: John Wiley & Sons.
2. Pearl, J., *On the Interpretation of do(x)*. Journal of Causal Inference, 2019: p. Feb.
3. Meehl, P.E., *Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology*. 1992.
4. Pearl, J. and E. Mackenzie, *The book of why: the new science of cause and effect*. 2018: Basic books.
5. Bareinboim, E. and J. Pearl, *Causal inference and the data-fusion problem*. Proceedings of the National Academy of Sciences, 2016. **113**(27): p. 7345-7352.
6. Sackeim, H.A., et al., *Effects of pulse width and electrode placement on the efficacy and cognitive effects of electroconvulsive therapy*. Brain Stimul, 2008. **1**(2): p. 71-83.
7. McDonald, C.J., S.A. Mazzuca, and G.P. McCabe, Jr., *How much of the placebo 'effect' is really statistical regression?* Stat Med, 1983. **2**(4): p. 417-27.
8. Lasagna, L., *The Efficacy of Self-Medication, Philosophy and Technology of Drug Assessment discussion*. Smithsonian Institution, Washington, DC: Joseph D. Cooper, ed, 1973. **168**: p. 156.
9. Lasagna, L., *Clinical trials of drugs from the viewpoint of the academic investigator (a satire)*. Clin Pharmacol Ther, 1975. **18**(5 Pt 2): p. 629-33.
10. Amsterdam, J.D., et al., *Double-blind comparison of doxepin and desipramine in patients with primary affective disorder*. Acta Psychiatr Scand, 1982. **65**(4): p. 292-300.
11. Akobeng, A.K., *Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice*. Acta Paediatr, 2007. **96**(4): p. 487-91.