
Causal epistemology: Questions

Boris Sobolev

This is an edited transcript of my talk at the Center for Clinical Epidemiology and Evaluation, Spotlight Series, June 13, 2022.[1]

1 Introduction


Hello, everyone. It's great to be here at the Spotlight today. Epistemology, the second word in my title, deals with the question, How do we know? How do we know that vaccine works, and cancer treatment prolongs life?

And here's our challenge. We see two things happening. But we don't see a causal relationship. How do we know that one causes the other? Both causation and covariation can manifest themselves as association. And the data alone can't distinguish between the two. We see treatment and we see changes in health. But we don't see causation.

It comes from reasoning with data. We reason randomization makes treatment groups similar. We reason stratification by confounders blocks covariation. And then we attribute changes in health status to treatment. We see association, rule out covariation, and claim causation.

2 The ladder of causality

In today's talk I will share with you a breakthrough in understanding causal claims. Professor Pearl made it by discovering and systematically examining the Ladder of Causality, a hierarchy of seeing, doing, and imagining. He showed how knowledge develops from association to



The Ladder of Causality

- Associations (seeing)
- Interventions (doing)
- Counterfactuals (imagining)

causal inference as we go up the ladder.[2]

Causal inference goes beyond the description of associations (*seeing*). We reason about the effect of deliberate intervention (*doing*). We also examine counterfactual scenarios (*imagining*). Counterfactuals is our way to see beyond data. Causal epistemology is a great intellectual feat. The poster at the end of this text gathers all the questions, definitions, target values, tools, and estimands of causal epistemology in one place.[3]

But today, I will focus only on the types of questions we ask as we go up the Pearl's hierarchy. The questions we ask determine the answers we get. The main learning objective today is to understand the different questions when making causal claims. To focus even more, let's look at causality in the context of health research.

I divide the talk into four parts: Associations, Interventions, Mediation and Personalized effects. Seeing Associations is our first step on the ladder.

3 Associations: comparing treatment groups

Events make up reality. Some events happens together. And sometimes they happen together for no reason. But sometimes we notice dependency between events. We notice that some events occur, and things change in response. In health research, we want to find out how much the outcome of interest changes in response to treatment. And we look at the difference in outcomes between groups receiving different treatments.

Why groups? Here's the thing. Individual results can vary. That's what drug advertisement tells us, right? The same treatment of the same disease by the same doctor may have different results.

Individual results vary. Therefore, we need to summarize individual results within the groups. The average outcome, the outcome probability, and the outcome rate, all are used as summary statistics in health research. Take for example the average outcome. It's the value around which an outcome is usually centered. That's why the most common question in health research is if there is a difference in average outcomes between treatment groups?



When outcomes have only two values, the average becomes a probability. Probability is a proportion. It's the number of times the value of interest occurs among all possible treatment outcomes.

Picture all study units as a unit square. The horizontal line in Figure 1 shows treatment groups: treated and untreated. And the vertical line shows the outcome values: *Yes* and *No*. The areas of this graph are proportional to the frequency of the combinations of treatments and outcomes. The bottom left rectangle shows how often we see treated patients with a *Yes* outcome. The top right rectangle shows how often untreated patients come with a *No* outcome.

We use their heights to test for independence between treatment and outcome. Under independence, the heights

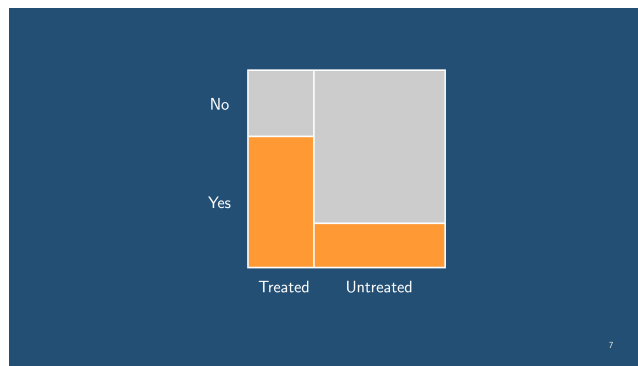


Figure 1: *Depicting associations*

are the same. Like the windows logo. The change in height shows the association. When events tend to occur together, we say they are positively related. And the taller rectangle would be on the left, as shown in this graph.

4 Interventions: deconfounding other influences

Our next step on the ladder is Interventions.

Let's go back to the difference in outcomes between treatment groups. Ultimately, we would like to use it to predict treatment results. We take what we learned from treated patients to predict what untreated patients could experience if treated.

But there is one problem such reasoning runs into. Treatment groups may be different. They may be non-comparable.

The presence of outcome determinants may differ between the groups. And we suspect that the difference in outcomes reflects their effects, and not the effect of the treatment. For example, two treatments will show different mortality if the treatment groups differ in age composition. The age-related mortality is mixed with the effect of treatment.



We usually call this *confounding*, and I follow the tradition. Just remember, non-comparable groups are the essence of confounding. Mixing of effects is a consequence of non-comparability.

It's not surprising then that at the next step we deal with de-confounding of effects. In other words, dealing with factors that may influence both the treatment selection and the outcome occurrence.

We say X causes Y , when changes in Y are to blame to changes in X . Change in X can occur by the forces of nature. Or, we can intervene and force treatment groups to be similar in every way except treatment. Through unnatural selection of units into treatment groups, we manipulate the distribution of outcome determinants. For example, in clinical trials we allocate treatments by randomization. Remember its purpose? To make treatment groups alike. By doing so, we de-confound the treatment effect and the effect of other factors.

When similar groups get different results after alternative treatments, we attribute the difference to treatments. In this sense, " X causes Y " is simply a shorthand for "group membership is responsible for different outcomes". This reasoning is based on the idea of comparing the response to applying and withholding treatment in the same patients. We compare the outcomes of untreated patients with the outcomes of the same patients as if they had received treatment.

Therefore instead of asking whether there is a difference in average outcomes between treatment groups, our question is would average outcomes be different if the same patients were in each treatment group?

Q: Would average outcomes be different if each treatment group had the same patients?

Recognition of the counterfactual nature of this question has led the regulator to clarify expectations for the proof of treatment effect. Submissions for the US FDA approval are required to demonstrate "how the results of the treatment compare with what would have happened to the same patients under an alternative treatment,"[4] see Figure 2.

III. ESTIMANDS (A.3)

Central questions for drug development and licensing are to establish the existence, and to estimate the magnitude, of treatment effects: **how the outcome of treatment compares to what would have happened to the same subjects under alternative treatment** (i.e., had they not received the treatment, or had they received a different treatment). An estimand is a precise description of the treatment effect reflecting the clinical question posed by a given clinical trial objective. It summarizes at a population level what the outcomes would be in the same patients under different treatment conditions being compared. The targets of estimation should be defined in advance of a clinical trial. Once defined, a trial can be designed to enable reliable estimation of the targeted treatment effect.



Figure 2: FDA guidance for estimands of trials[5]

Key in this requirement is the counterfactual alternative to the observed treatment: If the same patients had not have received the treatment or if they had received a different treatment.

Let's draw a square again to represent the population of units. All possible outcomes occupy the total area of this square. The outcomes of interest fill some space in it, colored in green in Figure 3. But we need two squares: one for all units treated. And the other - for the same units but remaining untreated. Then the treatment effect is the difference between green areas in these two squares.

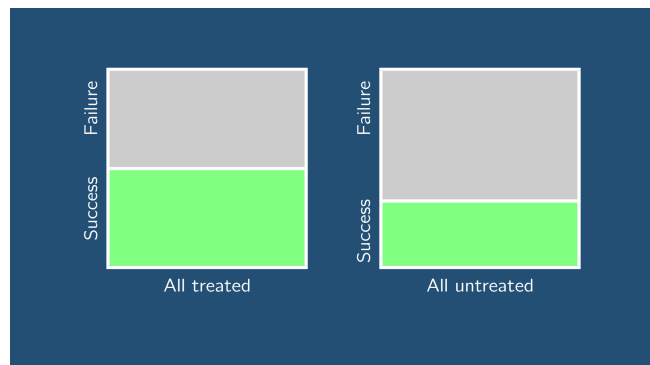


Figure 3: Comparing all units under two treatments

Contrasting all treated with all untreated is an important advance in reasoning about causality. It directly confronts a daunting problem of causal attribution: the comparability of treatment groups. We are always concerned that differences in outcomes may reflect the influence of other factors. This problem does not arise when the same patients make up the treatment groups. Indeed, their composition of outcome determinants is indistinguishable now. We are comparing like with like. Only one of which is treated.

Comparing the outcomes of identical treatment groups gives an unbiased treatment effect. In real data, we have some patients undergoing one treatment and some other patients receiving a different treatment. But we reason randomization makes treatment groups similar. The groups' sameness allows us to attribute the difference in outcomes to group membership. We credit the treatment with changes in health.

5 Mediation: deactivating intermediaries

The next step is Mediation.

Not all causal questions are answered by randomized trials. Take mediation. We can't randomize patients into groups defined by the combination of treatment and mediator, because mediator values are the result of treatment choice. We compare outcomes between treatment groups to find the treatment effect. And we always suspect differences in outcomes may reflect the influence of other factors. We look at pre-treatment factors that influence both the treatment and outcome occurrence.

Here's a new idea. Some post-treatment factors may also influence outcomes. When factors change their values in response to treatment and then, in turn, affect the outcome, we call them mediators. In a sense, mediators are treatment outcomes. Only they lie on the causal path between the treatment and the outcome of primary interest, see Figure 4.

Mediation comes with a new hypothesis: the mediator's variation explains the relationship between treatment and outcome. This is a powerful idea. In the presence of mediation, treatment and outcome, which are otherwise independent, can appear to be related. When mediators are

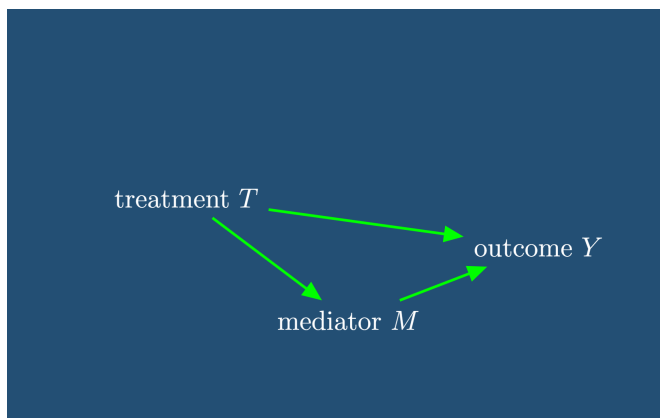


Figure 4: Basic mediation diagram

Q: Would average outcomes be different if each treatment group had the same patients with the same mediator values?

deactivated, the treatment effect can disappear.

The purpose of mediation analysis is to find out to what extent the association between treatment and outcome is due to the mediator. Naturally, conditioning on mediator will block all mediated causation. But it will not cut off the direct influence of treatment. But how can we condition on the outcome of treatment? Certainly not with a randomized experiment. We can't randomly allocate treatments and mediator values they produce. Mediators are treatment outcomes too, right?

Since Hume's time, we use counterfactual reasoning to assert causality in observed associations. We observe response to one treatment and wonder what would have happened if the same patients had received a different treatment. That's why we add an additional condition to the previous question and ask: Would average outcomes be different if each treatment group had the same patients with the same mediator values?

To answer this question, we can think of four hypothetical scenarios: (A) all patients receive treatment, (B) all patients remain untreated, (C) all patients receive treatment, but their mediators remain at values attained without treatment, (D) all patients remain untreated, but their mediators take values attained in the presence of treatment.

I label these scenarios A, B, C, and D. Then I wonder if the treatment effect will change when we compare A and C instead of A and B. Comparing A and B gives the overall treatment effect, the effect produced by the direct and mediated paths together. The difference in outcomes between A and C gives us the mediated effect. The difference between C and B will give the direct treatment effect.

Let's go back to the squares. In Figure 5 each square represents all units. On the left are units when they receive treatment. On the right are the same units but without treatment. Now I use the fishnet to show the proportion of units with the mediator taking one of two possible values, say 1. Notice that this proportion is the same in both squares.

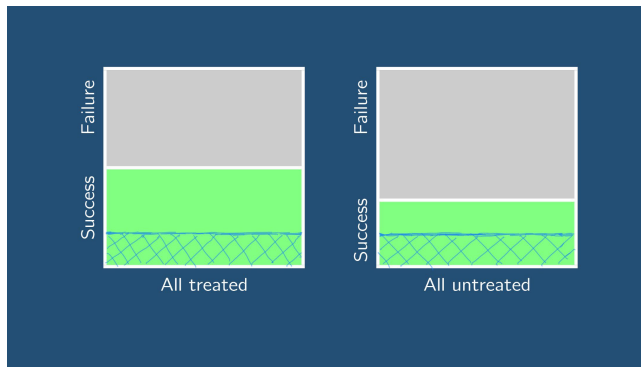


Figure 5: Treated and untreated with the same mediators

This means the right square shows the proportion of outcomes, the green area, in untreated units with naturally occurring mediator values. And the left square shows the proportion in treated units when their mediator have values that would be observed with no treatment.

The difference in the proportions of outcome between the two squares will give us the direct treatment effect. Mediator values of the control group considered in the treated units effectively remove the influence of mediator. Thus, we deactivate the mediator.

6 Personalized effects: tailoring treatment to individuals

And now we move to the top of the ladder of causality.

What is the problem that precision medicine promised to solve? The ATE crisis: the use of average treatment effects for individual treatment decisions.

Will taking a pill stop migraine? Health research never approaches this as Will taking medication stop a person having a migraine. Simply because, in health matters, we don't get to predict who will become ill, who will get better, or when, on an individual level.

ATE crisis:
Average treatment effects
for individual treatment decisions

21

Instead, testing a medical intervention involves comparing groups. A group of patients receives an experimental treatment and then another group receives a different treatment. And then the average outcomes of the groups are compared. This is the average treatment effect, ATE.

The ATE is the main piece of information used to judge whether an intervention is effective or not. But how adequately does an ATE inform the effect size (or even its existence) for an individual patient? In the patient population, some benefit from the intervention and some don't. Even the US FDA now "black-boxes" some approved interventions from clinical trials. Trials show a significant ATE, but the intervention does not benefit 40% of patients.

The individual treatment effect can be traced back to John Stuart Mill. In today's terms, Mill takes two exchangeable units and applies treatment to one. If we observe a difference in the condition of the units, we attribute it to the treatment. The units are exchangeable. In fact, they are two instances of the same unit. We can think of the change in their state as an individual treatment effect.

Individual treatment effect

John Stuart Mill, 1843

- take 2 exchangeable units
- apply treatment to one
- difference in their states is the treatment effect

22

Precision medicine tailors treatments to individuals. This raises epistemological issues that affect how we empirically test interventions. Say, we want to get individual effects from data. Then one obvious difficulty is that we only have data on one intervention applied to one person. Nor can we conduct a stratified randomized trial to get the treatment effect for all combinations of individual traits. But even if we could, stratum-specific effects would still be an ATE.

Pearl introduced a new tool to account for uniqueness in treatment decisions. Like Mill more than a century before him, Pearl looks at the outcomes of two treatments in the same unit. And then introduces the probability that the same patient would benefit from treatment and suffer without it. That's the probability of individual benefit of treatment.

Turns out it equals the probability that the outcomes of two treatments received by the same unit are different.

And it's possible to further narrow this probability down for individual characteristics.

This is a fundamental leap forward because we get to ask the right question: How likely is it that the outcome would have been different if the treatment had been different in the same patient? Or, even shorter, how likely is the treatment benefit for an individual? Or, to put it another way, what proportion of people would have benefited from the treatment?

Q: How likely is it that the outcome would have been different if the treatment had been different?

And this is not just an academic question. We can find bounds for this proportion using experimental and sometimes observational data. The bounds can be further found for the characteristics of the individual.

7 Main message

Alright, that's it for today. The main take-away: as we move up the ladder of causal claims, we get answers to different questions. The questions we ask determine the answers we get.

Associations tell us whether there is a difference in average outcomes between treatment groups. Interventions tell whether average outcomes would be different if each treatment group had the same patients. In mediation analysis, we ask if each treatment group had the same patients with the same mediator values? Finally, in individual treatment decisions, we ask how likely is the treatment benefit for an individual?

Take-aways

- difference in outcomes between treatment groups
- what if each treatment group had the same patients
- what if the same patients with the same mediators
- How common is the treatment benefit?

References

- [1] Sobolev B. Causal epistemology: questions; 2022. <http://tiny.cc/Cblog1>.
- [2] Pearl J. Causality: models, reasoning, and inference. Cambridge University Press; 2000.
- [3] Pearl J. Causes of Effects and Effects of Causes. Sociological Methods & Research. 2015;44(1):149-64.
- [4] US FDA. E9(R1) Statistical principles for clinical trials. Addendum: Estimands and sensitivity analysis in clinical trials; 2021. <https://www.regulations.gov/docket/FDA-2017-D-6113>.
- [5] ICH. Addendum on estimands and sensitivity analysis in clinical trials E9(R1); 2019. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf.

PEARL'S CAUSAL EPISTEMOLOGY

		CAUSES OF EFFECTS	
		COUNTERFACTUALS	PERSONALIZED EFFECTS
		MEDIATION	
		EFFECTS OF CAUSES	
		ASSOCIATIONS	
Activity	Comparing groups	Deactivating mediators	Imagining a different past
Query	Is there a difference in average outcomes between treatment groups?	Would average outcomes be different if each treatment group had the same patients with the same mediator values?	How likely is it that the outcome would have been different if treatment had been different? but not the other?
Target	Difference between groups	Natural direct effect	Bounds of an individual effect
Application	To report	To explain	To tailor
Tools	<p>Experimental treatment allocation</p> <ul style="list-style-type: none"> Treatment outcome $Y^x(u)$ if unit u were assigned to treatment group $X=x$ The marginal probability of $Y^x=y$ if all units had treatment $X=x$ $P(Y^x=y) \stackrel{\text{def}}{=} \sum_{(u)} I(Y^x(u)=y) P(u)$ <p>Non-experimental treatment allocation</p> <ul style="list-style-type: none"> Model G links pre-treatment variables Z, treatment X and outcome Y <ul style="list-style-type: none"> Conditional independence structure $P(y, x, z) = P(y x, z) P(x z) P(z)$ The joint distribution of variables X, Y, Z when $P(X=a z)$ is set to 1, counterfactual to the value assigned by G for a given z $P(y, z do(a)) = \frac{P(y, x, z)}{P(x z)} \Big _{x=a}$ <p>Association measure</p> <ul style="list-style-type: none"> Absolute risk reduction $P(Y=1 X=1) - P(Y=1 X=0)$ <p>Control for covariate Z</p> <ul style="list-style-type: none"> Stratification $P(Y=1 X=1, Z) - P(Y=1 X=0, Z)$ Inverse treatment-probability weighting $\frac{P(Y=1, X=1, Z)}{P(X=1 Z)} - \frac{P(Y=1, X=0, Z)}{P(X=0 Z)}$ <p>Covariate selection</p> <ul style="list-style-type: none"> What covariates we should control? What covariates we should not control? What is minimally necessary set? 	<p>Two-factor group assignment</p> <ul style="list-style-type: none"> Outcome $Y^{x,m}(u)$ if unit u had treatment $X=x$ and had mediator set to level m <p>Mediation in a thought experiment</p> <ul style="list-style-type: none"> Outcome $Y^x(u)$ and mediator $M_x(u)$ if unit u had been assigned to group $X=x$ The marginal probability of outcome if all units had treatment $X=1$ and mediator values as in group $X=0$ $P(Y^{1, M_0} = y)$ <p>Three hypothetical populations</p> <ul style="list-style-type: none"> All units were assigned to group $X=0$ All units were assigned to group $X=1$ All units were assigned to group $X=1$ with mediator levels as in group $X=0$ <p>Natural direct effect</p> <ul style="list-style-type: none"> The difference in average outcomes in group $X=1$ and $X=0$, if mediator values in both groups were the same as in group $X=0$ $NDE \stackrel{\text{def}}{=} E(Y^{1, M_0}) - E(Y^0)$ <p>Natural indirect effect</p> <ul style="list-style-type: none"> A change in average outcome for $X=0$ if mediator values were as for $X=1$ $NIE \stackrel{\text{def}}{=} E(Y^{0, M_1}) - E(Y^0)$ <p>Identification in non-experimental setting</p> <ul style="list-style-type: none"> NDE is identifiable by conditioning on Z X blocks all backdoor paths from M to Y, except through X Z-specific effect of $\{X, M\}$ on Y is identifiable Z-specific effect of X on M is identifiable no covariate from Z descends from X 	
Estimands	$P(Y=1 X=1) - P(Y=1 X=0)$ $E(Y X=1) - E(Y X=0)$	$P(Y^{1, M_0} = 1) - P(Y^0 = 1)$ $E(Y^{1, M_0}) - E(Y^0)$	$P(Y^1 - Y^0 = 1)$ $P(Y^1 = 1 X = 0, Z) - P(Y^0 = 1 X = 0, Z)$ $P(Y^1 = 1 X = 1, Z) - P(Y^0 = 1 X = 1, Z)$ <ul style="list-style-type: none"> If Z is a pure mediator $X \rightarrow Z \rightarrow Y$ $PNS(z) \leq \sum_{z, z' \neq z} \min\{P(Y=1 z); P(Y=0 z')\}$
			<p>Subtractive and additive counterfactuals</p> <ul style="list-style-type: none"> Necessity refers to expectations outcome would not have occurred if no treatment had been available when we see treatment and outcome occurring together The probability of being a necessary cause $PN = P(Y^0=0 Y=1, X=1) \stackrel{\text{def}}{=} \sum_{u'} P(Y^0(u)=0) P(u Y=1, X=1)$ Sufficiency refers to expectations outcome would have occurred had treatment been available when we see neither treatment nor outcome $PS = P(Y^1=1 Y=0, X=0) \stackrel{\text{def}}{=} \sum_{u'} P(Y^1(u)=1) P(u Y=0, X=0)$ <p>Attributable proportion</p> <ul style="list-style-type: none"> Under non-exogeneity and monotonicity $PN = 1 - \frac{P(Y=1 X=0)}{P(Y=1 X=1)}$ <p>Probability of necessity and sufficiency</p> <ul style="list-style-type: none"> The probability that outcome will occur with treatment and will not occur without it $PNS = P(Y^1=1, Y^0=0)$ It equals the average treatment effect measured in randomized experiment under the monotonicity assumption $PNS = P(Y^1=1) - P(Y^0=1)$
			<p>Individual effect</p> <ul style="list-style-type: none"> The difference in outcomes of treatment 1 and treatment 0 in unit u $Y^1(u) - Y^0(u)$ The probability of benefiting from treatment 1 relative to treatment 0 $P(Y^1 - Y^0 = 1) \equiv PNS$ <p>Bounds in an experimental study</p> <ul style="list-style-type: none"> The lower bound of PNS $\max\{0, P(Y^1=1) - P(Y^0=1)\}$ The upper bound of PNS $\min\{P(Y^1=1), P(Y^0=0)\}$ <p>Personalized treatment benefit</p> <ul style="list-style-type: none"> The probability of benefiting from treatment 1 compared with 0 among units sharing attribute $Z=z$ with unit u $PNS(z) = P(Y^1 - Y^0 = 1 Z(u) = z)$ If Z meets the backdoor criterion $X \leftarrow Z \rightarrow Y$