Causal inference and target trial emulation

Chris Holmes and Andrew Yiu

Department of Statistics, University of Oxford

IIAS, January 2023

DEPARTMENT OF
STATISTICS

Plan

1. Lecture 1 (Andrew): **Introduction to causal inference**.
   - Randomization inference, estimands as functionals, potential outcomes, graphs and Pearl's Structural Causal Model.
2. Lecture 2 (Chris): **Target trial emulation and predictive inference**.
   - Target trial emulation as a systematic approach for doing causal inference with observational data, predictive causal inference.
3. Practical:
   - Hands-on implementation of randomization inference and predictive resampling on dataset.

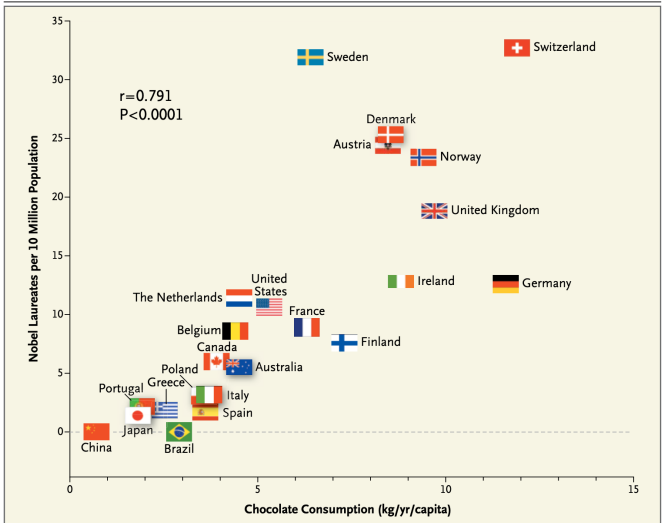


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Confectionery news.com

Eating chocolate produces Nobel prize winners, says study

By Oliver Nieburg

10-Oct-2012 - Last updated on 11-Oct-2012 at 11:51 GMT

"Association is not causation."

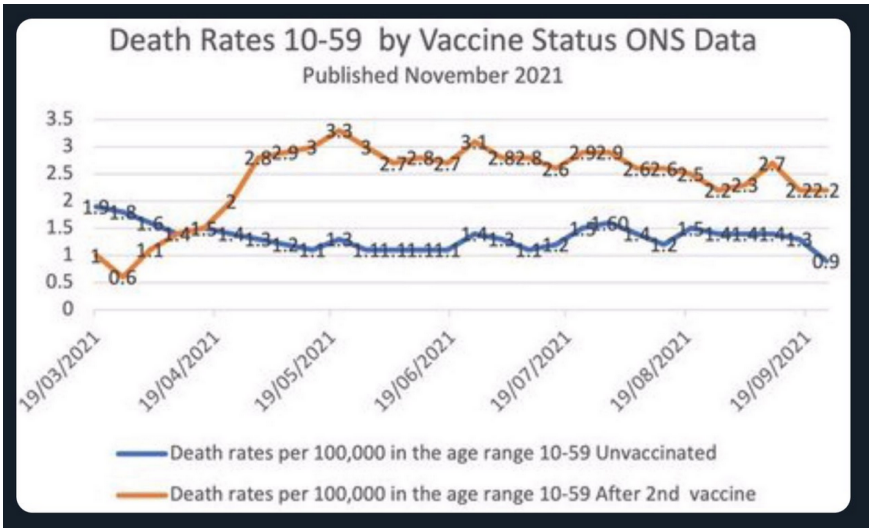
"Association is not causation."

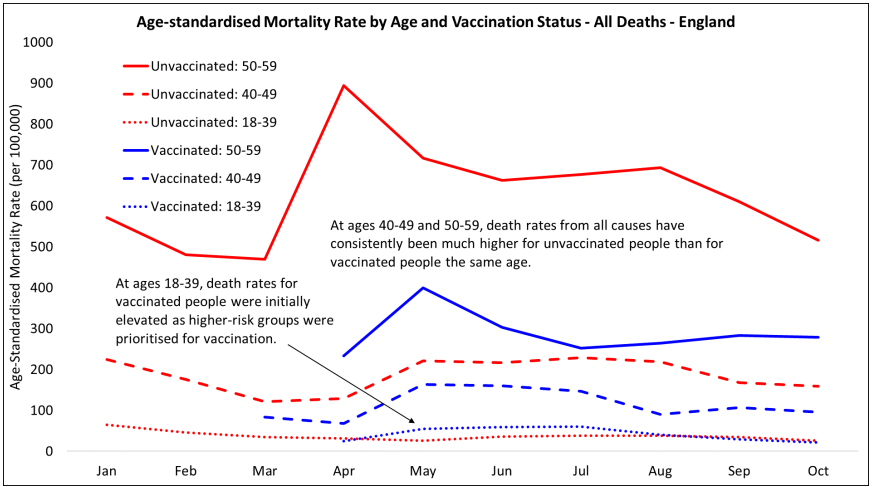
Examples of associational concepts: regression, classification, conditional independence, likelihood, p-value, confidence interval, root mean squared error, standard deviation, sufficient statistic, hazard ratio...

"Association is not causation."

Examples of associational concepts: regression, classification, conditional independence, likelihood, p-value, confidence interval, root mean squared error, standard deviation, sufficient statistic, hazard ratio...

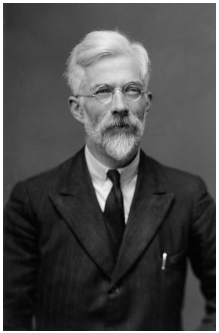
But are we ever interested in association without causation?


Death Rates 10-59 by Vaccine Status ONS Data
Published November 2021
3.5
3
2.5
2
1.5
1
0.5
0
19/03/2021
19/04/2021
19/05/2021
19/06/2021
19/07/2021
19/08/2021
19/09/2021
Death rates per 100,000 in the age range 10-59 Unvaccinated
Death rates per 100,000 in the age range 10-59 After 2nd vaccine


Age-standardised Mortality Rate by Age and Vaccination Status - All Deaths - England
Age-Standardised Mortality Rate (per 100,000)
1000
900
800
700
600
500
400
300
200
100
0
Unvaccinated: 50-59
Unvaccinated: 40-49
Unvaccinated: 18-39
Vaccinated: 50-59
Vaccinated: 40-49
Vaccinated: 18-39
At ages 40-49 and 50-59, death rates from all causes have consistently been much higher for unvaccinated people than for vaccinated people the same age.
At ages 18-39, death rates for vaccinated people were initially elevated as higher-risk groups were prioritised for vaccination.
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct

Randomized experiments

Fortunately, there's a method that guarantees association to be causation: conduct a randomized experiment.



Ronald Fisher

Rothamsted Research

The lady tasting tea

Muriel Bristol

At teabreak one day at Rothamsted, Fisher offered a cup of tea to a female colleague. She declined, stating that she preferred it when the milk had been poured first.

The lady tasting tea

Muriel Bristol

At teabreak one day at Rothamsted, Fisher offered a cup of tea to a female colleague. She declined, stating that she preferred it when the milk had been poured first.

Fisher was incredulous, thinking that there was surely no difference in taste.

The lady tasting tea

Muriel Bristol

At teabreak one day at Rothamsted, Fisher offered a cup of tea to a female colleague. She declined, stating that she preferred it when the milk had been poured first.

Fisher was incredulous, thinking that there was surely no difference in taste.

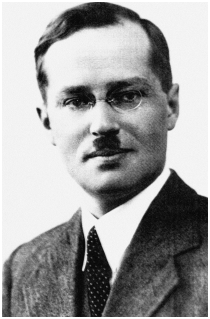
William Roach—a biochemist working in the antiseptics and insecticides lab—overheard this conversation and said: "Let's test her."

Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of what the test will consist, namely that she will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in a random order, that is in an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance, cards, dice, roulettes, etc., or, more expeditiously, from a published collection of random sampling numbers purporting to give the actual results of such manipulation. Her task is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received.

"The Design of Experiments" (1935) – R.A. Fisher

Potential outcomes

Fisher's test is often described using potential outcomes (aka: counterfactuals), which were introduced by Jerzy Neyman in his 1923 PhD thesis.



Jerzy Neyman

For an outcome variable Y (e.g. crop yield) and a binary treatment variable T (e.g. $T = 1$ if a fertilizer is used), we define a pair of variables (Y^1, Y^0): the potential outcome Y^t is the outcome that would be observed if—possibly contrary to fact—the subject receives treatment t.

If $T = 1$, then $Y = Y^1$ and we do not observe Y^0 (it becomes "counterfactual"), and vice-versa.

Fisher's test with potential outcomes

	Cup							
	1	2	3	4	5	6	7	8
Y^1	1	?	?	1	?	1	1	?
Y^0	?	0	0	?	0	?	?	0
T	1	0	0	1	0	1	1	0

- $Y = 1$: she guesses milk first; $Y = 0$ otherwise.
- $T = 1$: the tea was made with milk first; $T = 0$ otherwise.

Fisher's test with potential outcomes

	Cup							
	1	2	3	4	5	6	7	8
Y^1	1	0	0	1	0	1	1	0
Y^0	1	0	0	1	0	1	1	0
T	1	0	0	1	0	1	1	0

Under the null hypothesis

$$H_0 : Y_i^1 = Y_i^0 \text{ for } i = 1, \ldots 8,$$

we can fill in the missing data. The outcomes are treated as fixed; the only randomness comes from the randomization in T.

Fisher's test with potential outcomes

Under the null, what is the probability that she guesses all cups correctly?

Test statistic:

$$S = \#\text{correct guesses} = 2\left(\sum_{T_i=1} Y_i^1\right)$$

Probability of guessing all cups correctly:

$$\mathbb{P}(\text{all correct}) = \mathbb{P}(S = 8) = \frac{1}{\binom{8}{4}}.$$

Fisher's test with potential outcomes

Under the null, what is the probability that she guesses all cups correctly?

Test statistic:

$$S = \#\text{correct guesses} = 2\left(\sum_{T_i=1} Y_i^1\right)$$

Probability of guessing all cups correctly:

$$\mathbb{P}(\text{all correct}) = \mathbb{P}(S = 8) = \frac{1}{\binom{8}{4}}.$$

- Why did Fisher choose 8 cups? With 6 cups, there are $\binom{6}{3} = 20$ ways of choosing a group of 3 cups. Fisher wanted to set a significance level of 5%, and $(1/70) \approx 1.4\% < 5\%$.

Fisher's test with potential outcomes

Under the null, what is the probability that she guesses all cups correctly?

Test statistic:

$$S = \#\text{correct guesses} = 2\left(\sum_{T_i=1} Y_i^1\right)$$

Probability of guessing all cups correctly:

$$\mathbb{P}(\text{all correct}) = \mathbb{P}(S = 8) = \frac{1}{\binom{8}{4}}.$$

- Why did Fisher choose 8 cups? With 6 cups, there are $\binom{6}{3} = 20$ ways of choosing a group of 3 cups. Fisher wanted to set a significance level of 5%, and $(1/70) \approx 1.4\% < 5\%$.
- According to Fisher's biography, Muriel Bristol did guess all cups correctly. She also got married to William Roach.

The importance of randomization

"Whole forests have been destroyed to provide paper for disputes about Fisher's analysis." Stephen Senn

Some statisticians were sceptical of the idea of physically introducing randomization into the experiment. Fisher treated experimentation as a game played against an adversary (e.g. the laws of nature). If you randomize, you guarantee that your comparison groups are indeed comparable on average, no matter how the adversary tries to conspire against you.

The importance of randomization

"Whole forests have been destroyed to provide paper for disputes about Fisher's analysis." Stephen Senn

Some statisticians were sceptical of the idea of physically introducing randomization into the experiment. Fisher treated experimentation as a game played against an adversary (e.g. the laws of nature). If you randomize, you guarantee that your comparison groups are indeed comparable on average, no matter how the adversary tries to conspire against you.

"...having used a random allocation, the sternest critic is unable to say when we eventually dash into print that quite probably the groups were differentially biased through our predilections or through our stupidity." Bradford Hill

Randomized experiments

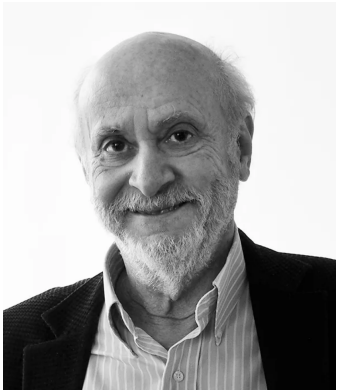
So we can draw causal conclusions from randomized experiments. But randomized experiments are not always feasible:

- Ethical concerns
- Too expensive
- Impractical

To this day, many researchers believe that causal inference is impossible unless we have a randomized experiment.

"*Dear author: Your observational study cannot prove causation. Please replace all references to causal effects by references to associations.*" from "The C-Word" (2018) - Miguel Hernán

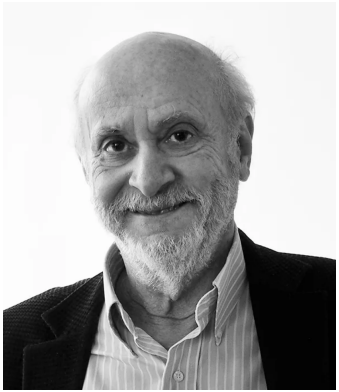
The Rubin Causal Model



Donald Rubin

In a series of papers in the 1970's, Don Rubin introduced a causal framework that extended Neyman's potential outcome model to the analysis of observational data.

The Rubin Causal Model



Donald Rubin

In a series of papers in the 1970's, Don Rubin introduced a causal framework that extended Neyman's potential outcome model to the analysis of observational data.

His philosophy was that we should use observational data to approximate (or: "reconstruct" or "emulate") a true randomized experiment as closely as possible.

Set-up

Suppose we have data $D_i = (Y_i, T_i, X_i)$ from an observational study:

- Y_i is the outcome variable of interest (e.g. all-cause mortality)
- T_i is a binary variable indicating treatment assignment (e.g. $T_i = 1$ if subject i receives vaccination).
- X_i is a vector of measured pre-treatment covariates (e.g. age). For expositional simplicity, we will assume throughout that the covariates are discrete.

SUTVA

As before, we define a pair of potential outcome variables (Y_i^1, Y_i^0) for each subject, which are formally linked to the observed outcomes by

$$Y_i = T_i Y_i^1 + (1 - T_i) Y_i^0.$$

Implicit in this formulation is the stable unit treatment value assumption (SUTVA), which has two components:

1. There is no interference; that is, it is assumed that a unit's potential outcomes are unaffected by the treatments assigned to other units.
2. The second is that there are no hidden variations of treatments: $Y_i^{(t)}$ does not depend on how unit i received treatment t, e.g. the hospital that a patient visited.

The fundamental problem

Under SUTVA, we can frame causal inference as a missing data problem:

Unit	Y^0	Y^1	T	X
1	?	Y_1^1	1	X_1
2	Y_2^0	?	0	X_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	?	Y_n^1	1	X_n

From Rubin's perspective, both (Y_i^1, Y_i^0) exist; it's just that we observe at most one of them. This is often referred to as the "fundamental problem of causal inference".

He compares this to the Heisenberg Uncertainty Principle in quantum mechanics, in which both the position and momentum of a particular are well-defined but it is impossible to measure both precisely.

Unconfoundedness

A further crucial assumption is called unconfoundedness.
For each unit i, we have

$$T_i \perp\!\!\!\perp (Y_i^1, Y_i^0) \mid X_i,$$

so that we can write $\mathbb{P}(T_i \mid Y_i^1, Y_i^0, X_i) = \mathbb{P}(T_i \mid X_i)$. This is sometimes called "ignorability" or "conditional exchangeability".

Unconfoundedness

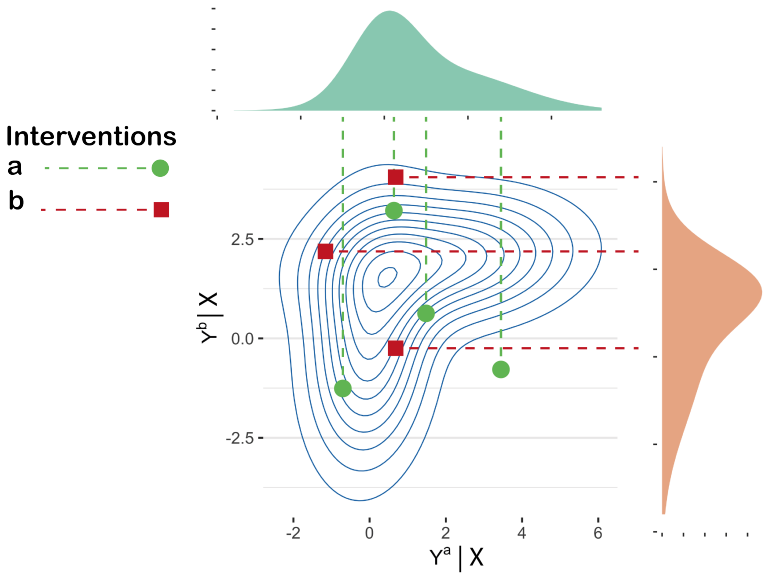
A further crucial assumption is called unconfoundedness.
For each unit i, we have

$$T_i \perp\!\!\!\perp (Y_i^1, Y_i^0) \mid X_i,$$

so that we can write $\mathbb{P}(T_i \mid Y_i^1, Y_i^0, X_i) = \mathbb{P}(T_i \mid X_i)$. This is sometimes called "ignorability" or "conditional exchangeability".

Important: this doesn't mean that the "treatment" is conditionally independent of the "outcomes"! Remember that T_i is the treatment **assignment** and (Y_i^1, Y_i^0) are the **potential** outcomes.

The way to interpret this is that we assume that there is randomization within levels of X, i.e. conditional on X, all other factors are balanced on average across both the treated and untreated groups.


Interventions
a
b
2.5
0.0
-2.5
$Y^b \mid X$
-2
0
2
4
6
$Y^a \mid X$

Separating the science from the design

Under the unconfoundedness assumption, the data distribution factorizes as

$$\begin{aligned}\mathbb{P}(T_i, Y_i^1, Y_i^0, X_i) &= \mathbb{P}(T_i \mid Y_i^1, Y_i^0, X_i)\mathbb{P}(Y_i^1, Y_i^0, X_i) \\ &= \mathbb{P}(T_i \mid X_i)\mathbb{P}(Y_i^1, Y_i^0, X_i) \quad \text{(unconfoundedness)}\end{aligned}$$

Rubin calls (Y_i^1, Y_i^0, X_i) the "science"; this is what we're actually interested in learning about.

An important aspect of the Rubin Causal Model is that the "science" is separated from the "design" of the experiment (i.e. the treatment assignment mechanism). Intuitively, $\mathbb{P}(Y_i^1, Y_i^0, X_i)$ should be invariant to the choice of experiment used to uncover its properties.

Overlap

The final assumption is called overlap (aka "positivity"). This means that

$$0 < Pr(T_i = 1 \mid X_i) < 1$$

with probability one for each unit i. In words, this requires the probability of receiving either treatment conditional on X to be strictly positive.

Overlap

The final assumption is called overlap (aka "positivity"). This means that

$$0 < Pr(T_i = 1 \mid X_i) < 1$$

with probability one for each unit i. In words, this requires the probability of receiving either treatment conditional on X to be strictly positive.

This would be violated, for example, if doctors always assign treatment to critically ill patients. In this case, it would be impossible to estimate causal effects among the critically ill without making heroic extrapolations from the data on other patients.

Identification

Given our assumptions (SUTVA, unconfoundedness and overlap), we can identify our causal estimand; that is, we can write it in terms of quantities that we can estimate from the observed data. For instance, let's suppose that we are interested in the average treatment effect $\theta = \mathbb{E}[Y^1] - [Y^0]$.

$$\mathbb{E}[Y^1] = \mathbb{E}[\mathbb{E}[Y^1 \mid X]] \quad \text{(tower law)}$$

Identification

Given our assumptions (SUTVA, unconfoundedness and overlap), we can identify our causal estimand; that is, we can write it in terms of quantities that we can estimate from the observed data. For instance, let's suppose that we are interested in the average treatment effect $\theta = \mathbb{E}[Y^1] - [Y^0]$.

$$\begin{aligned}\mathbb{E}[Y^1] &= \mathbb{E}[\mathbb{E}[Y^1 \mid X]] \quad \text{(tower law)} \\ &= \mathbb{E}[\mathbb{E}[Y^1 \mid X, T = 1]] \quad \text{(unconfoundedness and overlap)}\end{aligned}$$

Identification

Given our assumptions (SUTVA, unconfoundedness and overlap), we can identify our causal estimand; that is, we can write it in terms of quantities that we can estimate from the observed data. For instance, let's suppose that we are interested in the average treatment effect $\theta = \mathbb{E}[Y^1] - [Y^0]$.

$$\begin{aligned}
\mathbb{E}[Y^1] &= \mathbb{E}[\mathbb{E}[Y^1 \mid X]] \quad \text{(tower law)} \\
&= \mathbb{E}[\mathbb{E}[Y^1 \mid X, T = 1]] \quad \text{(unconfoundedness and overlap)} \\
&= \mathbb{E}[\mathbb{E}[Y \mid X, T = 1]] \quad \text{(SUTVA)}
\end{aligned}$$

Identification

Given our assumptions (SUTVA, unconfoundedness and overlap), we can identify our causal estimand; that is, we can write it in terms of quantities that we can estimate from the observed data. For instance, let's suppose that we are interested in the average treatment effect $\theta = \mathbb{E}[Y^1] - [Y^0]$.

$$\begin{aligned}\mathbb{E}[Y^1] &= \mathbb{E}[\mathbb{E}[Y^1 \mid X]] \quad \text{(tower law)} \\ &= \mathbb{E}[\mathbb{E}[Y^1 \mid X, T = 1]] \quad \text{(unconfoundedness and overlap)} \\ &= \mathbb{E}[\mathbb{E}[Y \mid X, T = 1]] \quad \text{(SUTVA)}\end{aligned}$$

This is called the g-formula. It's not the same as conditioning on $T = 1$!

$$\mathbb{E}[Y \mid T = 1] = \mathbb{E}[\mathbb{E}[Y \mid X, T = 1] \mid T = 1]$$

i.e. association is (generally) not causation!

Estimands as functionals

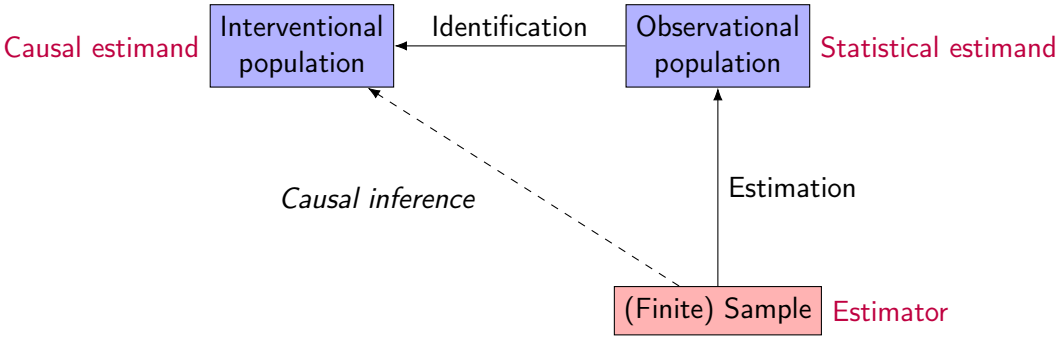
By doing the same for $\mathbb{E}[Y^0]$, the average treatment effect can be written as

$$\mathbb{E}[Y^1] - \mathbb{E}[Y^0] = \mathbb{E}[\mathbb{E}[Y \mid X, T = 1]] - \mathbb{E}[\mathbb{E}[Y \mid X, T = 0]].$$

The right-hand side is a nonparametric (or "model-free") formula. It describes the estimand directly as a mapping of the data-generating distribution $\theta = \theta(P)$ (often called a statistical functional).

The practice of defining estimands as model-free functionals is commonplace in causal inference, but it differs from the classical perspective of first choosing a statistical model (e.g. linear and logistic regression, the Cox model, additive hazards models) and then defining the estimand in terms of its parameters.

The "roadmap"


Causal estimand
Interventional population
Identification
Observational population
Statistical estimand
Causal inference
Estimation
(Finite) Sample
Estimator

Adjustment sets

To summarize: if we have found a valid set of covariates X that satisfies the unconfoundedness and overlap assumptions, then the average treatment effect is

$$\mathbb{E}[Y^1] - \mathbb{E}[Y^0] = \mathbb{E}[\mathbb{E}[Y \mid X, T = 1]] - \mathbb{E}[\mathbb{E}[Y \mid X, T = 0]].$$

We can then estimate the effect by fitting models for particular components of the data-generating distribution.

Adjustment sets

To summarize: if we have found a valid set of covariates X that satisfies the unconfoundedness and overlap assumptions, then the average treatment effect is

$$\mathbb{E}[Y^1] - \mathbb{E}[Y^0] = \mathbb{E}[\mathbb{E}[Y \mid X, T = 1]] - \mathbb{E}[\mathbb{E}[Y \mid X, T = 0]].$$

We can then estimate the effect by fitting models for particular components of the data-generating distribution.

But how do we decide on X?

- Should we adjust for all measured pre-treatment variables?
- What if there are unmeasured variables that we know are relevant to the model? Can we still estimate the causal effect?

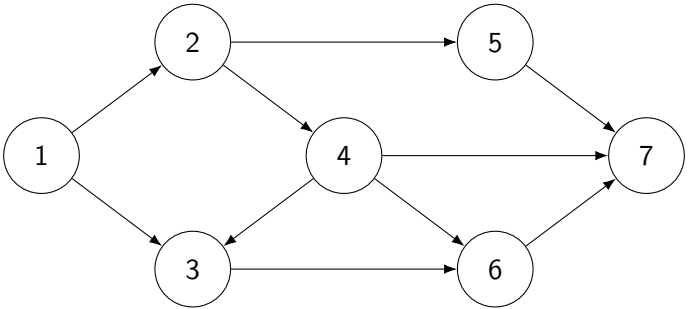
Judea Pearl

Judea Pearl

Judea Pearl—a computer scientist at UCLA—advocates a graphical approach to causality.

Among many other things, he invented a simple graphical test called the back-door criterion to choose a valid adjustment set for identification.

Causal diagrams



Definition

A path is a sequence of distinct adjacent vertices. We say that a vertex V_i is a descendant of another vertex V_j if there is a directed path $V_i \to \ldots \to V_j$. If $V_i \to V_j$, then V_i is a parent of V_j.

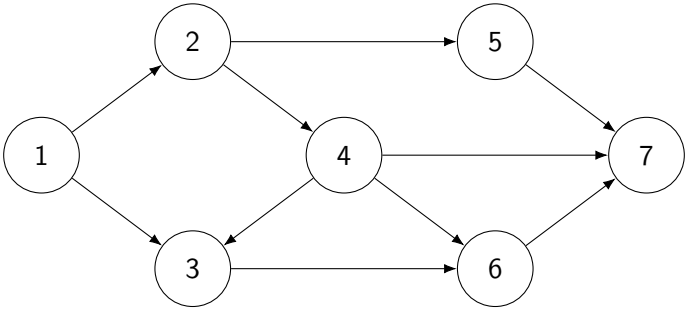
Blocking

Definition

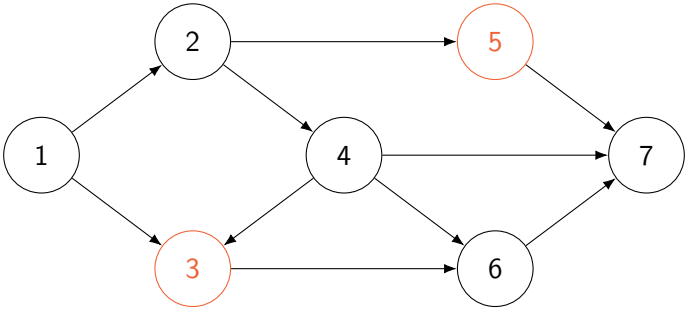
A path between V_1 and V_n is blocked by a set S (containing neither V_1 nor V_n) whenever there is a node V_k satisfying either

1. $V_k \in S$ and
   - $V_{k-1} \to V_k \to V_{k+1}$ (chain)
   - $V_{k-1} \leftarrow V_k \leftarrow V_{k+1}$ (chain)
   - $V_{k-1} \leftarrow V_k \to V_{k+1}$ (fork)
2. $V_{k-1} \to V_k \leftarrow V_{k+1}$ (collider) with neither V_k nor any of its descendants contained in S.

Example

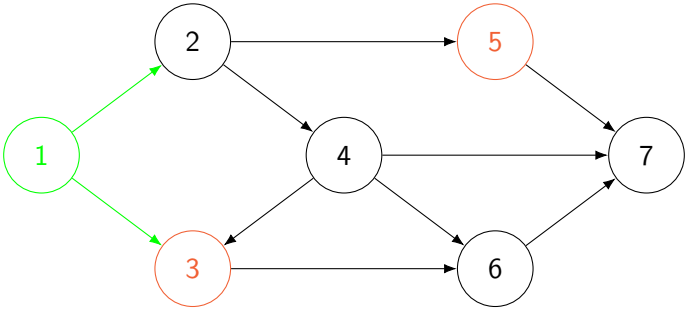


Example



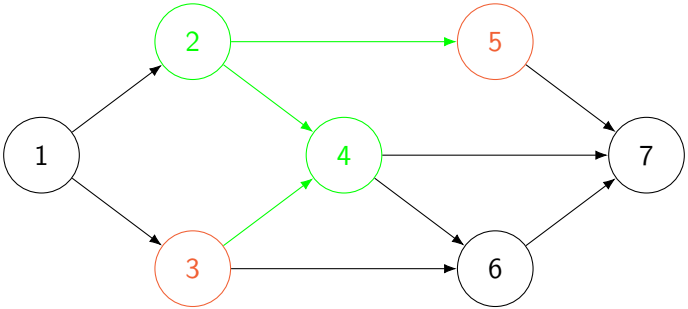
Claim: Every path from $\{3\}$ and $\{5\}$ is blocked by $\{1\}$.

Example



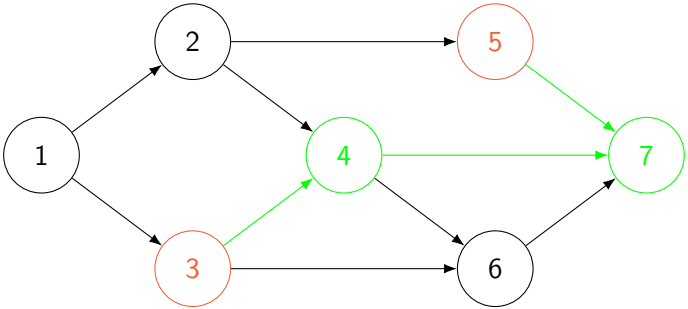
Claim: Every path from $\{3\}$ and $\{5\}$ is blocked by $\{1\}$.
Any path that goes through $\{1\}$ is blocked.

Example



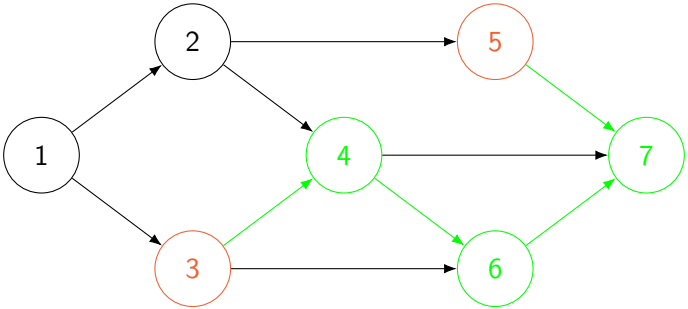
Claim: Every path from $\{3\}$ and $\{5\}$ is blocked by $\{1\}$.
$\{4\}$ is a collider and neither $\{4\}$ nor $\{7\}$ are contained in $\{1\}$.

Example



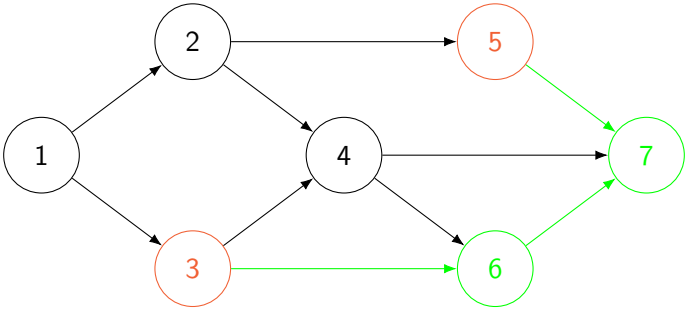
Claim: Every path from $\{3\}$ and $\{5\}$ is blocked by $\{1\}$.
$\{7\}$ is a collider and $\{7\}$ is not contained in $\{1\}$. In fact, any path that goes through 7 makes it a collider.

Example



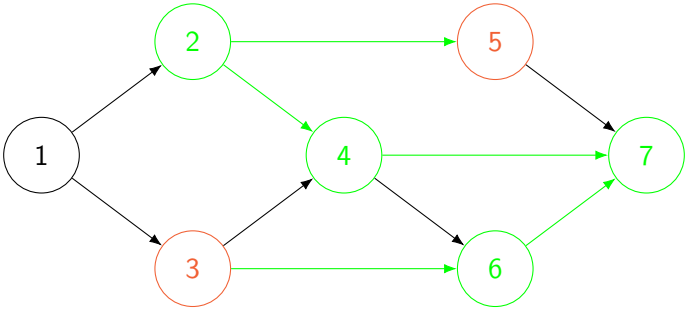
Claim: Every path from $\{3\}$ and $\{5\}$ is blocked by $\{1\}$.
$\{7\}$ is a collider and $\{7\}$ is not contained in $\{1\}$. In fact, any path that goes through 7 makes it a collider.

Example



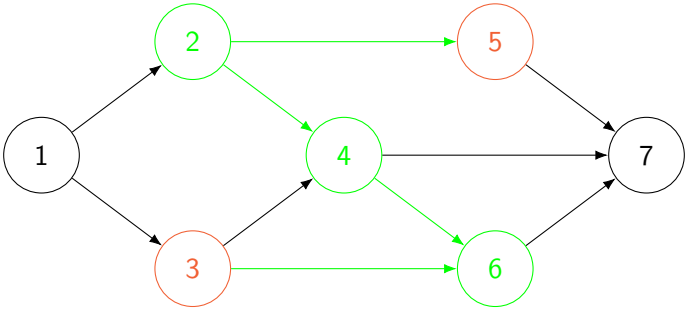
Claim: Every path from $\{3\}$ and $\{5\}$ is blocked by $\{1\}$.
$\{7\}$ is a collider and $\{7\}$ is not contained in $\{1\}$. In fact, any path that goes through 7 makes it a collider.

Example



Claim: Every path from $\{3\}$ and $\{5\}$ is blocked by $\{1\}$.
$\{7\}$ is a collider and $\{7\}$ is not contained in $\{1\}$. In fact, any path that goes through 7 makes it a collider.

Example



Claim: Every path from $\{3\}$ and $\{5\}$ is blocked by $\{1\}$.
$\{6\}$ is a collider and $\{6\}$ and $\{7\}$ are not contained in $\{1\}$.

The back-door criterion

Definition

A set of variables W satisfies the back-door criterion relative to (T, Y) if

- no node in W contains a descendant of T
- W blocks all paths from T to Y entering T through the back-door ($T \leftarrow \ldots$).

The back-door criterion

Definition

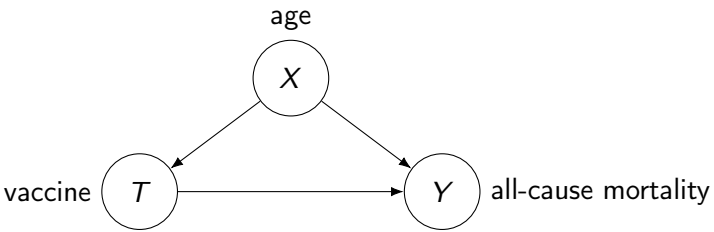
A set of variables W satisfies the back-door criterion relative to (T, Y) if

- no node in W contains a descendant of T
- W blocks all paths from T to Y entering T through the back-door ($T \leftarrow \dots$).

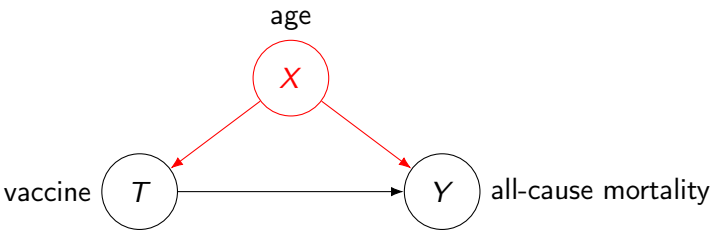
If a set of variables W satisfies the back-door criterion relative to (T, Y), then the average treatment effect of T on Y is given by the adjustment formula, i.e.

$$\mathbb{E}[Y^1] - \mathbb{E}[Y^0] = \mathbb{E}[\mathbb{E}[Y \mid W, T = 1]] - \mathbb{E}[\mathbb{E}[Y \mid W, T = 0]].$$

Example 1

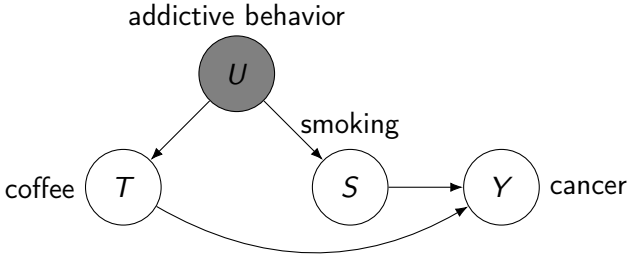


Example 1



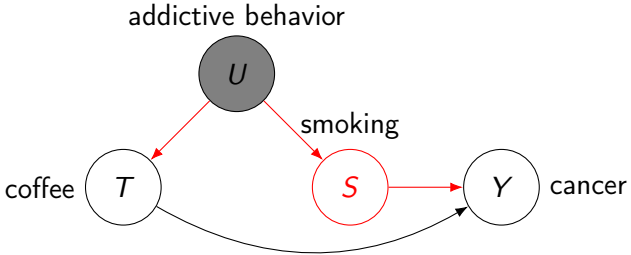
The back-door criterion tells us we should adjust for age!

Example 2



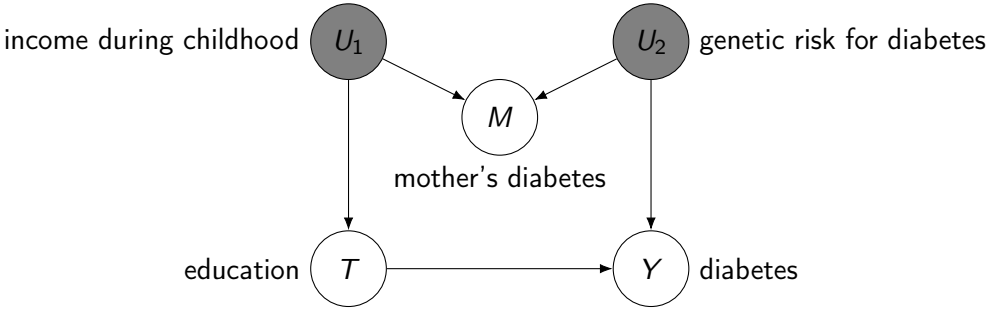
Addictive behavior is unmeasured.

Example 2

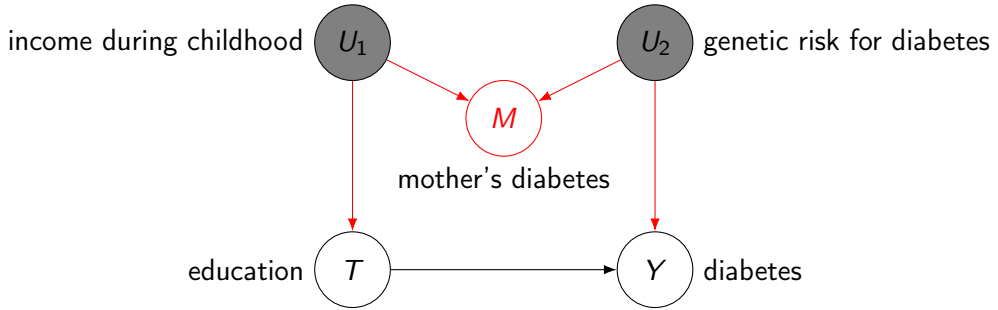


Addictive behavior is unmeasured. We should adjust for smoking even though it doesn't have a causal effect on drinking coffee!

Example 3: M-bias



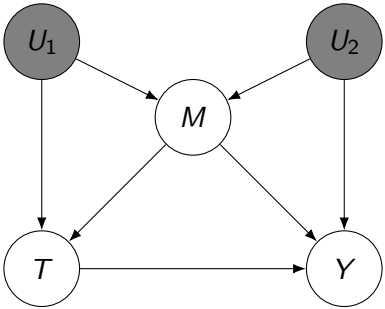
Example 3: M-bias



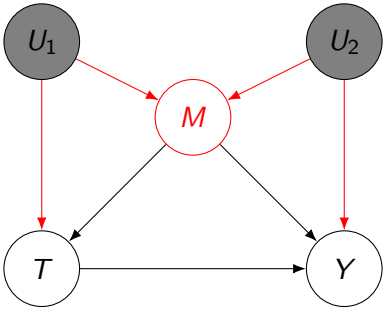
We shouldn't adjust for the mother's diabetes even though

- it's a pre-treatment variable
- it's associated with both T and Y.

Example 4: butterfly bias

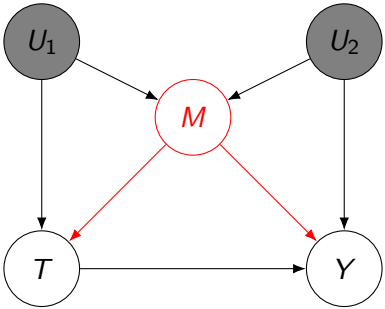


Example 4: butterfly bias



Similar to before, M is a collider on this path, which suggests we shouldn't adjust for it.

Example 4: butterfly bias



But at the same time, we need to adjust for M because it's a common cause of T and Y! In this case, the average treatment effect is unidentifiable.

Summary

- It is important to think about our problems from a causal perspective in order to guide decision making.
- The cornerstone of causal inference is the randomized experiment. But there are many questions that can only be answered with the help of non-randomized data.
- Causal analyses separate identification from estimation.
- Graphs are a very helpful way of representing our causal assumptions and enables convenient tools for identification such as the backdoor criterion.

Target trial emulation

Causal analysis of observational data is hard. This is not just because we require assumptions that lie outside of the data (e.g. SUTVA and unconfoundedness). Even if these assumptions hold, we must be careful not to introduce biases into the analysis by using the data inappropriately.

Recently, Hernan and Robins (2016) described the use of target trial emulation as a qualitative approach to ensure rigour in the causal analysis of observational data.

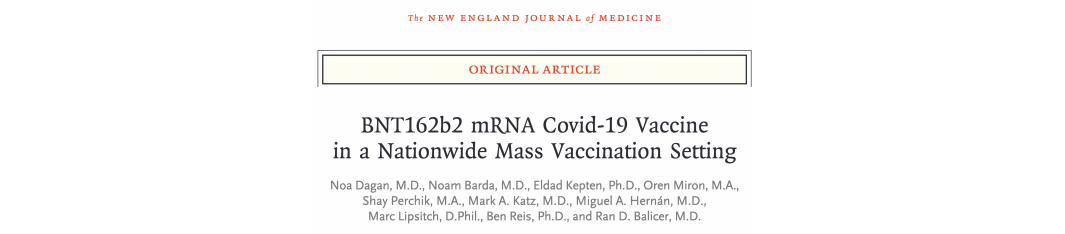
The idea of conceptualising an idealized randomized trial when designing observational studies dates back a long time. The recent innovation is in retrospective analysis of observation studies (that already exist) and the notio of a closest matching RCT

Target trial emulation

In Target Trial Emulation, the analyst is required to frame their causal questions by specifying the protocol of an explicit pragmatic (target) trial, i.e. design a matched target trial to the observational study, that they would have liked to carry out.

The analyst then follows the protocol of the target trial in the analysis and reporting of inference from the observational data.

Example: effectiveness of the Pfizer-BioNTech vaccine



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting

Noa Dagan, M.D., Noam Barda, M.D., Eldad Kepten, Ph.D., Oren Miron, M.A., Shay Perchik, M.A., Mark A. Katz, M.D., Miguel A. Hernán, M.D., Marc Lipsitch, D.Phil., Ben Reis, Ph.D., and Ran D. Balicer, M.D.

Target trial emulation was used to investigate the efficacy of the Pfizer-BioNTech COVID-19 vaccine using data from Israel's largest health care organisation.

The large sample size ($n \approx 600,000$) allowed the researchers to estimate vaccine effectiveness in subpopulations that previous randomized trials could not evaluate due to lack of power, e.g. individuals over 70.

Target trial protocol specification

Protocol component	Protocol specification
Eligibility criteria	Individuals aged ≥ 16 without a previous vaccination and documented positive PCR test, and has been a member of the healthcare organisation for the previous 12 months.
Treatment strategies	1. One dose of the vaccine at baseline and one dose 3 weeks later 2. No vaccination
Assignment	Individuals are randomly assigned to either strategy at baseline and are aware of their assigned strategy, i.e. no double-blind assignment.
Time zero & follow-up	Starts at assignment and ends at diagnosis of COVID-19 outcome, death, loss to follow-up, or administrative end of follow-up.
Outcomes	COVID-19 diagnosis, COVID-19 hospitalization, severe COVID-19 outcome, COVID-19 death.
Causal estimands	Per-protocol effect

First, the user must frame their causal questions by carefully specifying the protocol of a pragmatic randomized trial that they would have liked to carry out.

Target trial protocol emulation

Protocol component	Protocol emulation
Eligibility criteria	Same as for target protocol
Treatment strategies	Same as for target protocol
Assignment	Individuals are assumed to be randomly assigned to either strategy conditional on a set of measured confounders, e.g. age, sex, area of residence, prior infection, comorbidities etc.
Time zero & follow-up	For vaccinated, eligible individuals: time zero is the day of vaccination. For unvaccinated, eligible individuals: time zero is the first day of eligibility.
Outcomes	Same as for target protocol
Causal estimands	Same as for target protocol

The analysis of the observational data must then emulate the protocol as closely as possible.

Target trial workflow

Target trial emulation provides a qualitative workflow for causal analysis

This ensures that key variables and analysis objectives are pre-sepcified, improving reproducibilty and transparency of the analysis

We highlight that generative models can be used within target trial emulation to provide a complete, self-contained, quantitative causal framework without the need for counterfactuals.

'Causal inference can be classified into two distinct classes of problems: predicting effects of interventions and reasoning about counterfactuals. - Judea Pearl

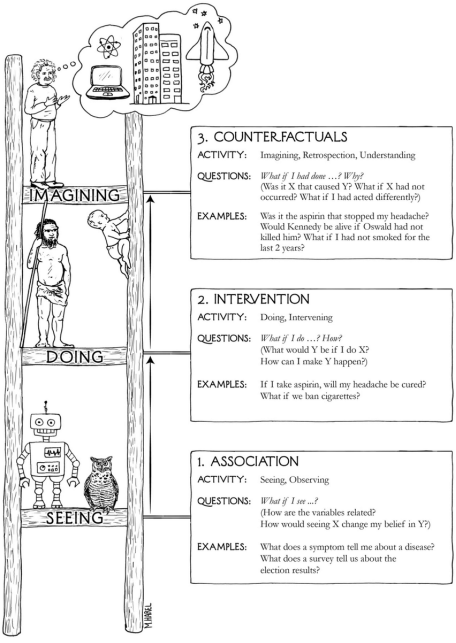
Most medical applications are of the first type, and we offer a bespoke framework that avoids any additional redundant machinery.

Pearl's causal ladder

The Three Layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in$Y$?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Figure 1: The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.



- Potential outcomes
- Target trial predictive framework
- Probability

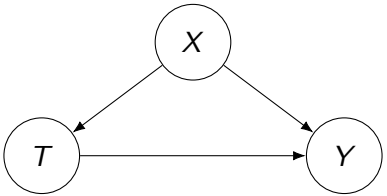
Set-up

We will consider the simplest (canonical) set-up for illustration.

We have data $Z_i = (X_i, T_i, Y_i)$ $(i = 1, \ldots, n)$, recorded on n independent units from an observational study where

- X_i is a vector of pre-treatment covariates
- Y_i is an outcome variable of interest
- T_i is a binary treatment indicator, $T_i \in \{0, 1\}$

We have potential confounding and we wish to infer something about the causal treatment effect.



The idea

- We treat causal inference as a missing data problem, where the missing data is from a closest matching, hypothetical, population-scale randomized controlled trial (RCT), matched to the observational study
- We use a predictive generative model (joint probability distribution) to then simulate participants and outcomes from the population RCT, conditional on information from the observational study
- Following which any scientific (refutable) causal quantity of interest can be read off from the synthetic RCT data, without the need to introduce counterfactuals
- Repeated simulation of the population-scale RCT quantifies uncertainty in the population causal effect

Foundations

In the paper we develop an axiomatic causal framework without the need for counterfactuals – building on decision-theoretic causal inference (e.g. Dawid 2021) and target trial emulation.

A regime indicator F_T indexes the joint distribution of observables under different regimes:

- $P(Y, T, X \mid F_T = \mathscr{O})$: observational regime
- $P(Y, T, X \mid F_T = \mathscr{E})$: closest matching experimental (randomized) regime

We'd like data $Z^{\mathscr{E}} = (Y^{\mathscr{E}}, X^{\mathscr{E}}, T^{\mathscr{E}})$ but we have $Z^{\mathscr{O}} = (Y^{\mathscr{O}}, X^{\mathscr{O}}, T^{\mathscr{O}})$.

Solution: build a joint predictive (generative) model conditioned on what you know

$$P(Z^{\mathscr{E}} \mid z^{\mathscr{O}}_{i=1:n})$$

Simulate $Z^{\mathscr{E}}_{n+1:N}$ for very large N for the missing population on the target RCT and pick off any (observable) causal quantity of interest.

Causal inference as a missing data problem

Recall that causal inference can be framed as a missing data problem in the Rubin causal model.

Unit	Y^0	Y^1	T	X
1	?	Y_1^1	1	X_1
2	Y_2^0	?	0	X_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	?	Y_n^1	1	X_n

Only one potential outcome per unit is ever observed, the other being counterfactual on assignment of the treatment to a unit.

Target trial prediction – data table

Unit	Y	T	X	F_T
1	Y_1	1	X_1	$\mathscr{O}$
2	Y_2	0	X_2	$\mathscr{O}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	Y_n	1	X_n	$\mathscr{O}$
$n+1$	?	?	?	$\mathscr{E}$
$n+2$	?	?	?	$\mathscr{E}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

By assumption, once we have $Z_{n+1:N}$, we can exactly recover the target parameter $\theta(Z_{n+1:N})$, e.g. the population average treatment effect is the Neyman difference-in-means estimator

$$\theta^{ATE}(Z_{n+1:\infty}) = \lim_{N\to\infty} \left[\frac{\sum_{k=n+1}^{N} Y_k 1(T_k = 1)}{\sum_{k=n+1}^{N} 1(T_k = 1)} - \frac{\sum_{k=n+1}^{N} Y_k 1(T_k = 0)}{\sum_{k=n+1}^{N} 1(T_k = 0)} \right].$$

Prequential factorization

We require a joint predictive distribution $p(z_{n+1:N} \mid z_{1:n}, F_T = \mathscr{E})$.

Given a joint predictive we can then impute the missing data conditioned on the observational data $Z_{1:n}$.

But specifying a joint predictive generative model is challenging. To make this task more manageable, it is helpful to use the chain rule to decompose the joint into a product of conditional factors

$$p(z_{n+1:N} \mid z_{1:n}, F_T = \mathscr{E}) = \prod_{i=n+1}^{N} p(z_i \mid z_{1:i-1}, F_T = \mathscr{E}).$$

Predictive resampling

We use the shorthand $p_k(z) = p(z \mid z_{1:k}, F_T = \mathscr{E})$.

$$
\begin{aligned}
\text{Draw } Z_{n+1} &\sim p_n, \text{ then update the predictive with } Z_{1:n+1} \\
\text{Draw } Z_{n+2} \mid Z_{n+1} &\sim p_{n+1}, \text{ then update the predictive with } Z_{1:n+2} \\
\text{Draw } Z_{n+3} \mid Z_{n+2} &\sim p_{n+2}, \text{ then update the predictive with } Z_{1:n+3} \\
&\vdots
\end{aligned}
$$

This computational scheme—introduced by Fong et al. (2022)—is called predictive resampling.

Predictive causal structure

We enforce each p_k to factorize as

$$p_k(y, t, x) = p_k(y \mid t, x) p_k(t) p_k(x),$$

to reflect the dependence structure of the target trial. We can then decompose each step of the predictive resampling as follows.

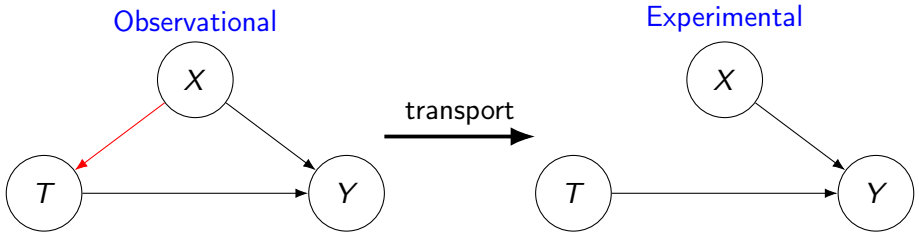
Conditional on $Z_{1:k-1}$:

1. Sample X_k from $p_{k-1}(x)$ *(Predict the pre-treatment covariates for the k-th unit)* – for example using a Bootstrap (eCDF) to draw a datum to copy
2. Sample T_k from Ber(0.5) *(Assign a randomized treatment value)*
3. Sample Y_k from $p_{k-1}(y \mid T_k, X_k)$ *(Predict the outcome given the administered treatment and pre-treatment covariates)*
4. Update the predictive to $p_k \mid X_k, T_k, Y_k$

Building a predictive from observational data

A conspicuous problem is the absence of target trial data with which to build our first-step predictive $p_n(z) = p(z \mid z_{1:n}, F_T = \mathscr{E})$.

To resolve this, we instead elicit a first-step predictive for the observational regime $p(z \mid z_{1:n}, F_T = \mathscr{O})$ and transport it into the experimental regime using an axiomatic framework.



Transportability

Suppose we have specified a first-step predictive $p(z \mid z_{1:n}, \mathscr{O})$. We will first assume that the population characteristics do not change from the observational to the experimental regimes:

$$p(x \mid z_{1:n}, \mathscr{E}) = p(x \mid z_{1:n}, \mathscr{O}).$$

Israel vaccine example: in the observational data, X includes age, sex, area of residence, prior infection etc. We would require our target population (e.g. the population of Israel) to share the same distribution on these variables. This can be relaxed if we have additional information on our target population (e.g. census data) to construct a different predictive.

Transportability

We also assume stability in the outcome regression model

$$p(y \mid z_{1:n}, t, x, \mathscr{E}) = p(y \mid z_{1:n}, t, x, \mathscr{O}).$$

Israel vaccine example: the outcome Y is COVID-19 hospitalization and the treatment T is two doses of the Pfizer-BioNTech vaccine. So we assume that if we know the individual's T and X, their probability of becoming hospitalized from COVID-19 does not depend on how the treatment was assigned, i.e. under intervention or observation.

The g-formula density

Finally, we assume that treatment allocation is fully randomized in our target trial:

$$p(t = 1 \mid z_{1:n}, \mathscr{E}) = 0.5.$$

Our first-step predictive for the experimental regime is now given by the g-formula density

$$p(y, t, x \mid z_{1:n}, \mathscr{E}) = p(y \mid z_{1:n}, t, x, \mathscr{O}) p(t \mid \mathscr{E}) p(x \mid z_{1:n}, \mathscr{O}).$$

The only change from $\mathscr{O}$ to $\mathscr{E}$ is replacing the observational treatment prediction $p(t \mid z_{1:n}, \mathscr{O})$ with the fully randomized assignment $p(t \mid \mathscr{E})$.

Application

We applied our methodology to study the effect of maternal smoking cessation during pregnancy on birthweight. Our dataset is an excerpt of the data studied in Almond (QJE 2005). The dataset contains information on singleton births in Pennsylvania between 1989 and 1991.

We define the treatment variable T to take the value 1 if the individual is assigned to the smoking cessation group, and 0 if they are instructed to continue smoking.

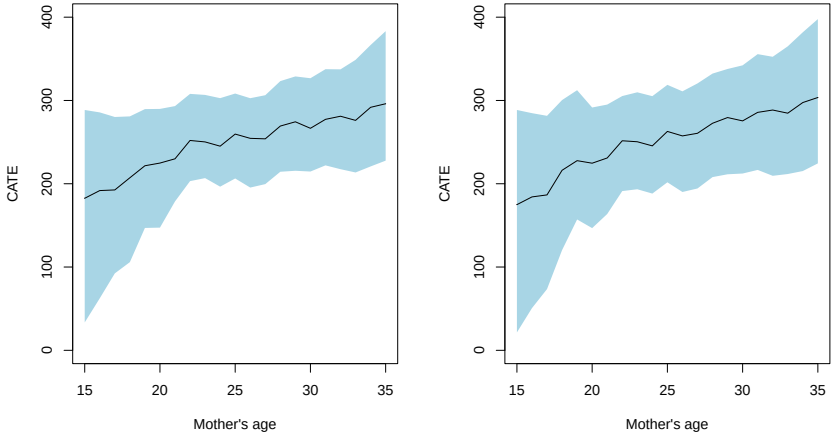


Figure: The posterior means and pointwise 95% intervals for the conditional average treatment effects given age (y-axis scale is in grams): BART and the Bayesian bootstrap (left); BART augmented with the clever covariate and the Bayesian bootstrap (right).

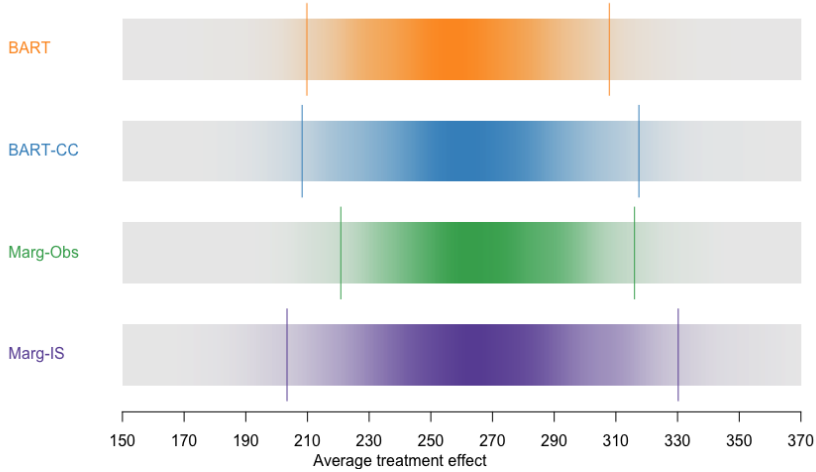


Figure: Posterior distributions for the average treatment effect (x-axis scale is in grams). The darkness of the strips is proportional to the posterior density, with the central 95% credible regions indicated.

Conclusions

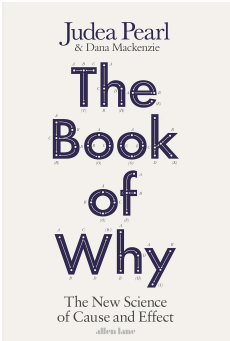
- Target trial emulation (Hernan and Robins 2016) provides a powerful qualitative workflow for causal analysis of observational studies
- By augmenting the target trial protocol with a generative model we can simulate outcomes from the target trial, conditional on observational data, and pick off any (scientific) causal effects of interest
- The framework highlights that counterfactuals are not needed for causal inference on scientific (falsifiable) hypothesis – as a population-scale RCT is as good as it gets
- The usual causal assumptions map to interpretable and intuitive assumptions on transportability of predictive models across populations and conditions

References

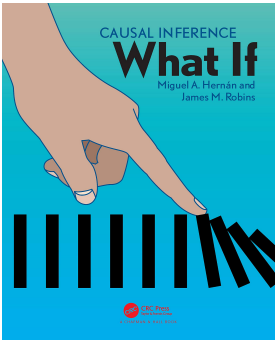
Preprint: Yiu, A., Fong, E., Walker, S.G. and Holmes, C. *"Causal predictive inference and target trial emulation."* https://arxiv.org/abs/2207.12479

- Dagan, N. et al. (2021) BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting. *The New England Journal of Medicine,* 384:1412-1423
- Dawid, A.P. (2021) Decision-theoretic foundations for statistical causality. *Journal of Causal Inference,* 9:39–77
- Fong, E., Holmes, C. and Walker, S.G. Martingale posterior distributions (with discussion). *JRSS-Series B,* 2022.
- Hernán, M. A., and Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8), 758-764.

Thank you for listening!



Pearl (2018)



Hernán and Robins (2021)

- "The Book of Why" (2018) - Judea Pearl and Dana Mackenzie: popular science-type book with lots of history and fun examples.
- "Causal Inference: What If" (2021) - Miguel Hernán and James Robins: accessible textbook that provides a balanced overview of the different approaches to causal inference.

Further reading: papers

- Rubin, D.R. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *Annals of Statistics*, 6(1), 34-58. The first rigorous formulation of the Rubin Causal Model. Frames causal inference as a mising data problem and advocates Bayesian inference for estimating causal effects.
- Holland, P.W. (1986). Statistics with Causal Inference. *JASA*, 86, 945-960. Overview of the RCM and the Fundamental Problem of Causal Inference.
- Dawid, A.P. (2000). Causal Inference without Counterfactuals (with discussion). *JASA*, 95, 407-424. Arguments against the use of counterfactuals for causal inference, with discussion from prominent advocates of counterfactuals.
- Observational Studies. Volume 8, Issue 2, 2022. Interviews with the most influential causal researchers: Judea Pearl, Don Rubin, James Heckmann, and Jamie Robins.
- Hernán, M.A., Wang, W. and Leaf, D.E. (2022). Target Trial Emulation: A Framework for Causal Inference From Observational Data *JAMA*, 328, 2446-2447. Recent overview of target trial emulation.

Practical materials

Link to practical materials: https://github.com/andrew-yiu/IIAS

Joint modelling

We have advocated using a joint predictive distribution $p(z_{n+1:N} \mid z_{1:n})$ to predict the hypothetical target trial data. The emphasis here is to have dependence between the data points.

One might question the merit of having dependence in our predictive structure if the data are believed to be independent. The point is that we are (partially) using probability to quantify epistemic uncertainty; the dependence structure allows us to update our predictive as we acquire more data in order to iteratively improve our predictive accuracy for the future.

Exchangeability

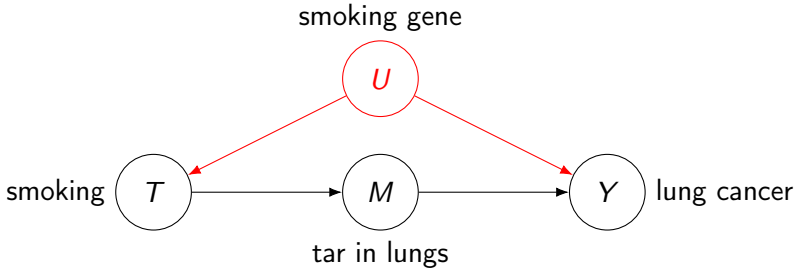
Even if the data are believed to be i.i.d., the underlying data-generating mechanism is unknown, so we can model the data as being conditionally i.i.d. given an unknown parameter P. By averaging over a distribution on P (representing our epistemic uncertainty on P c.f. a prior in Bayesian inference), the resulting distribution on the variables is exchangeable, i.e.

$$Pr(X_1 = x_1, \ldots, X_k = x_k) = Pr(X_{\sigma(1)} = x_1, \ldots, X_{\sigma(k)} = x_k)$$

for all k and any permutation σ of $\{1, \ldots, k\}$. In words, the joint distribution is invariant under reorderings.

Under very general conditions, de Finetti's theorem tells us that the converse is also true; any exchangeable probability distribution can be derived from a conditionally i.i.d. model averaged over a mixing distribution on P.

The front-door criterion



There is no back-door adjustment set, but the treatment effect is identifiable:

$$\mathbb{E}[Y^t] = \sum_{m,t'} \mathbb{E}[Y \mid M = m, T = t']p(M = m \mid T = t)p(T = t').$$