Introduction to causal inference in healthcare

Javier González

August 7, 2022

Microsoft Research Cambridge
OxML summer school, 2022

"I checked it very thoroughly, said the computer, and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is."

Douglas Adams, The Hitchhiker's Guide to the Galaxy (1979)

Schedule of this tutorial

Regression, causality, statistical paradoxes and other fairy tales (2.5h)

- Develop the intuition to understand the main elements of casual reasoning.
- Provide a brief introduction to the different frameworks of causal reasoning and inference.
- Things we will discuss: causal models, DAGs, potential outcomes, experimental design, propensity scores, etc.

Causal inference and precision medicine (1h)

- Review the current efforts in casual inference for real world evidence in the healthcare domain.
- Discuss other research directions in causal inference for healthcare.

Regression, causality, statistical paradoxes and other fairy tales

Old friends, new friends

Linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Old friends, new friends

Linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Bayesian linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2), \;\; \beta \sim \mathcal{N}(0, \Sigma_p)$$

Old friends, new friends

Linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Bayesian linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2), \;\; \beta \sim \mathcal{N}(0, \Sigma_p)$$

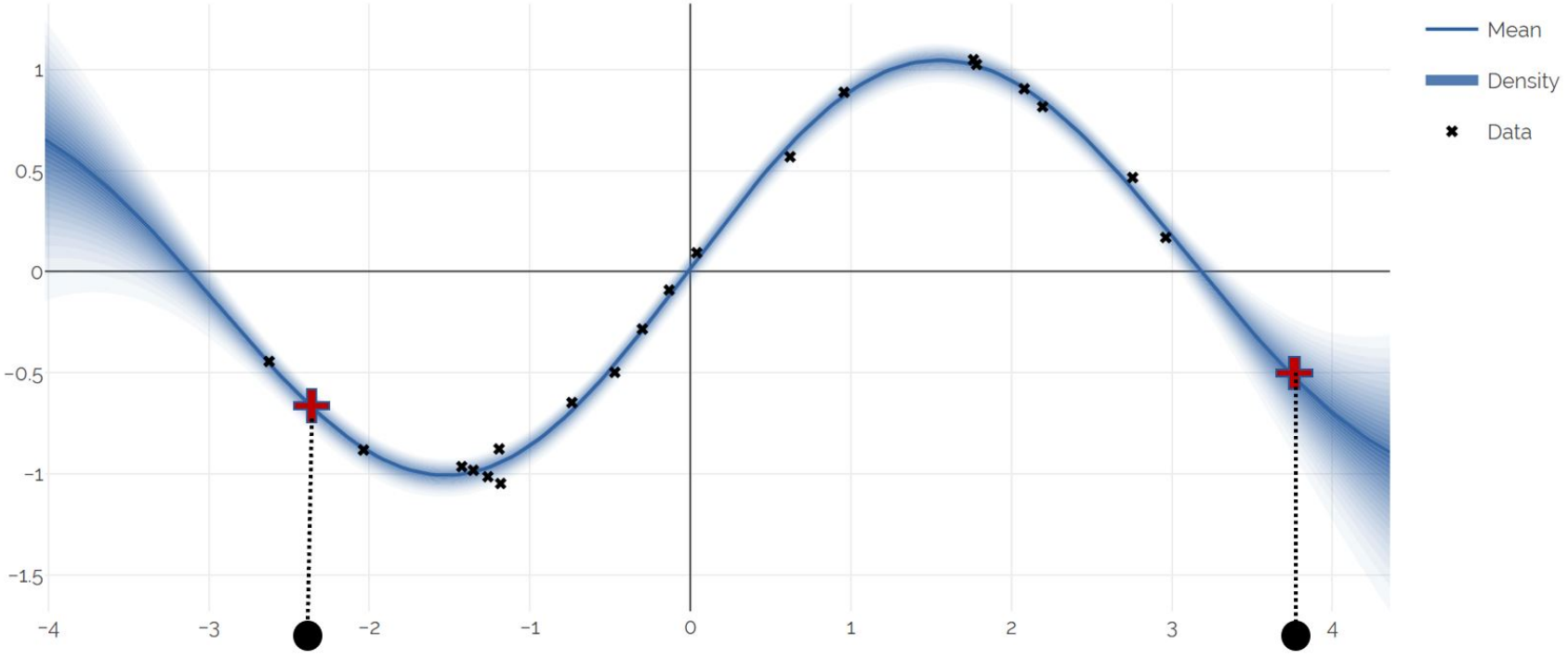
Bayesian non-linear regression (Gaussian process):

$$Y = f(X_1, \dots, X_p) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \quad f \sim \mathcal{GP}(0, K)$$

$$Y = \sum_{k}^{d_F} w_k \phi_k(X_1, \dots, X_p) + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2), \;\; w \sim \mathcal{N}(0, \Sigma_{d_F})$$

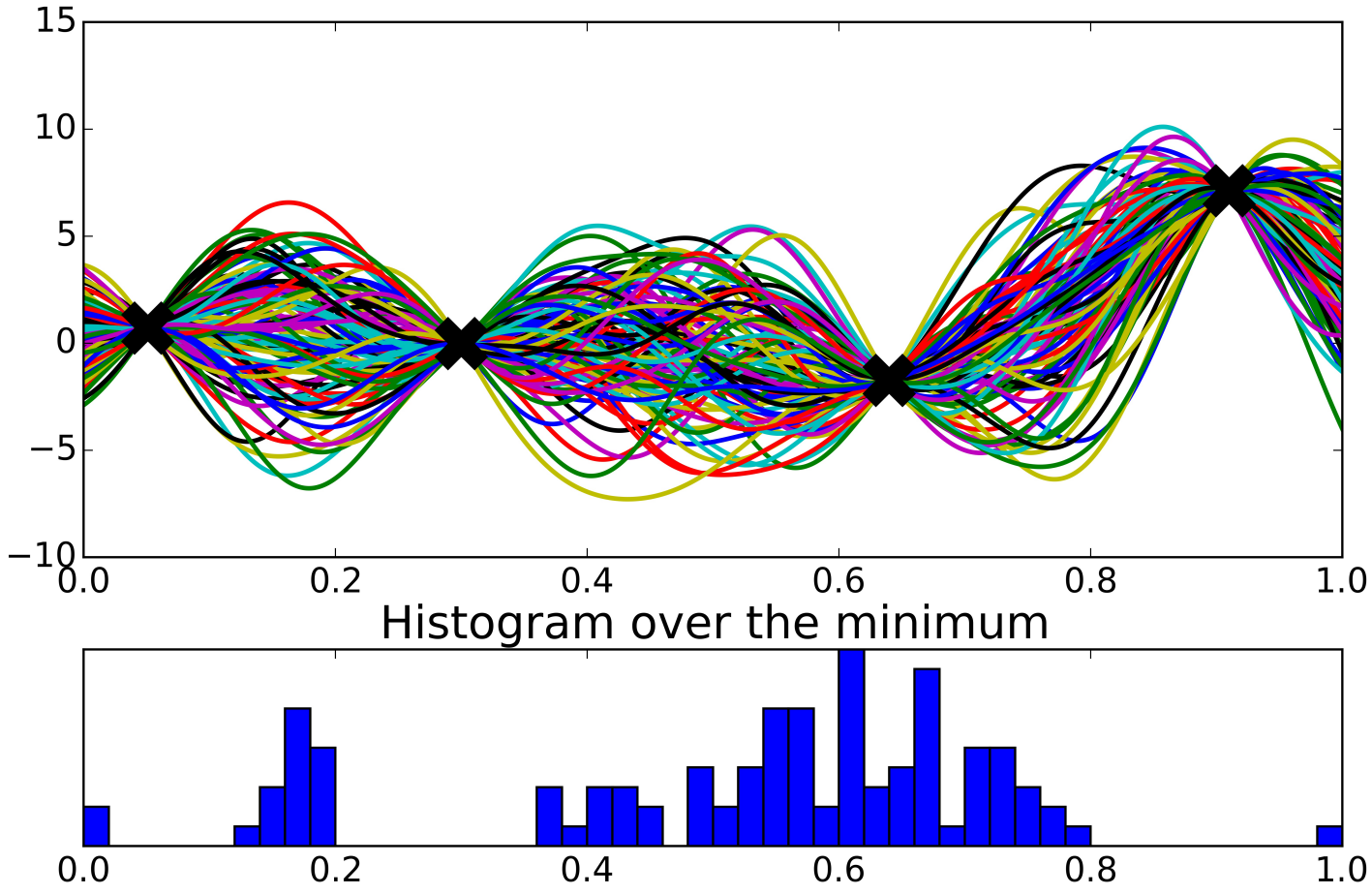
What can I do with a regression model?

1. I can make a predictions:



What can I do with a regression model?

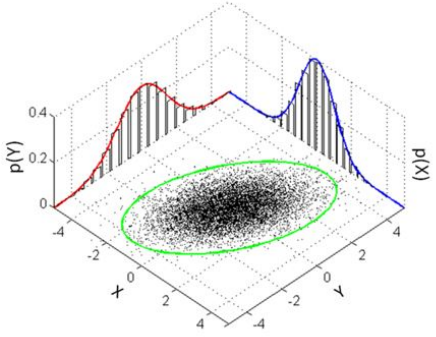
2. I can learn about about a property of $f(x)$.



What can I do with a regression model?

3. I can estimate a causal effect.

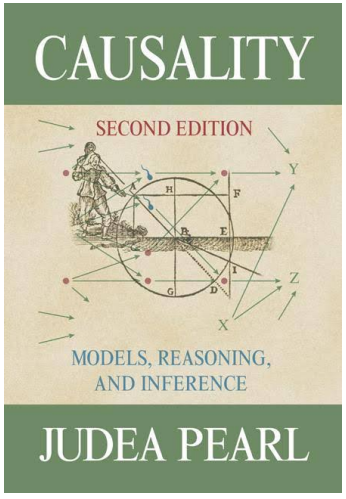
Observer: data + prior knowledge

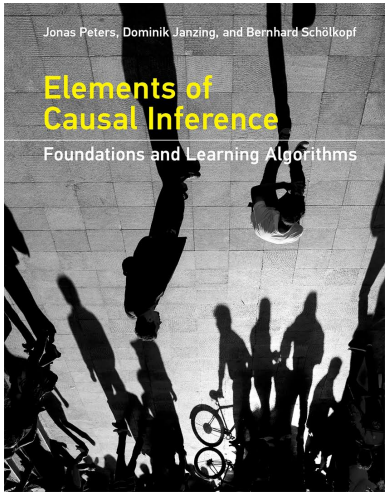


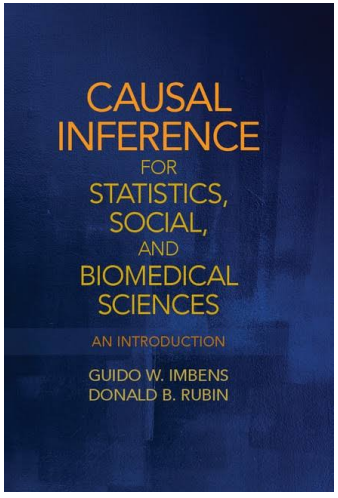
Data that emerge from observing the universe

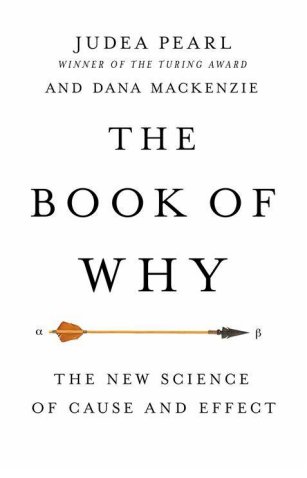
Laws of the universe

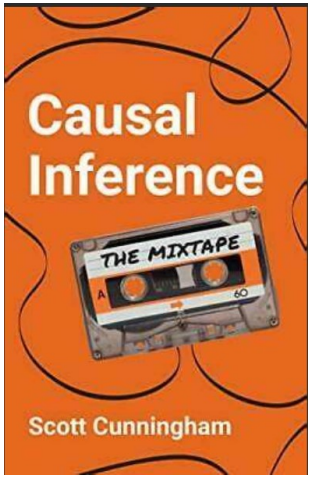
Main sources of this tutorial











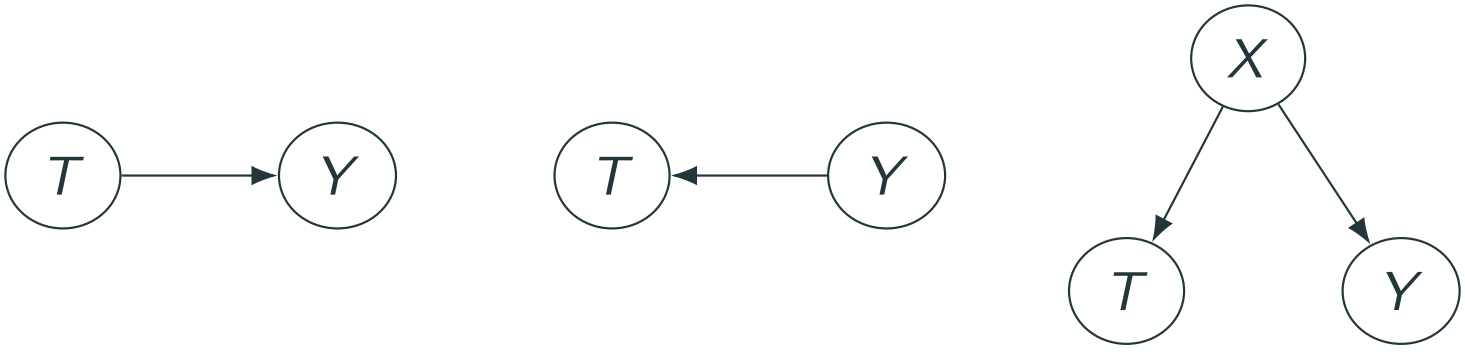
What is a causal effect and how can we represent it mathematically

Correlation is not causation...

Is someone really stealing the sun?

What can't I do If I perfectly know $\mathbb{P}(T, Y)$?

$\mathbb{P}(T, Y)$ helps to describe the world but not to understand how it works.



Reichenbach's Common Cause Principle

'If two events are correlated, then either there is a causal connection between them or there is a third event, a so called common cause, which brings about the correlation.'

Structural equation models

Question

How do we express mathematically that the dose of treatment (T) affects the recovery speed (Y)?

$$Y = \beta T + \epsilon_Y$$

- T stands for the dose of the treatment.
- Y stands for recovery speed.
- ϵ_Y stands for all factors that could affect Y when T is held constant.

Modelling the directionality of the causal effect

$Y = \beta T + \epsilon_Y$ does not properly express the causal relationship.

- Algebraic equations are symmetric.
- $T = (Y - \epsilon_Y)/\beta$, the recovery speed causes the dose (?!).
- We need a diagram to disambiguate the situation.

Modelling the directionality of the causal effect

$Y = \beta T + \epsilon_Y$ does not properly express the causal relationship.

- Algebraic equations are symmetric.
- $T = (Y - \epsilon_Y)/\beta$, the recovery speed causes the dose (?!).
- We need a diagram to disambiguate the situation.

Structural Causal model: $\mathbb{P}(T, Y)$ + causal graph G.

Full model for treatment and disease

$$T = \epsilon_T$$
$$Y = \beta T + \epsilon_Y$$

ϵ_t and ϵ_y are 'exogenous variables' (unobserved background factors).

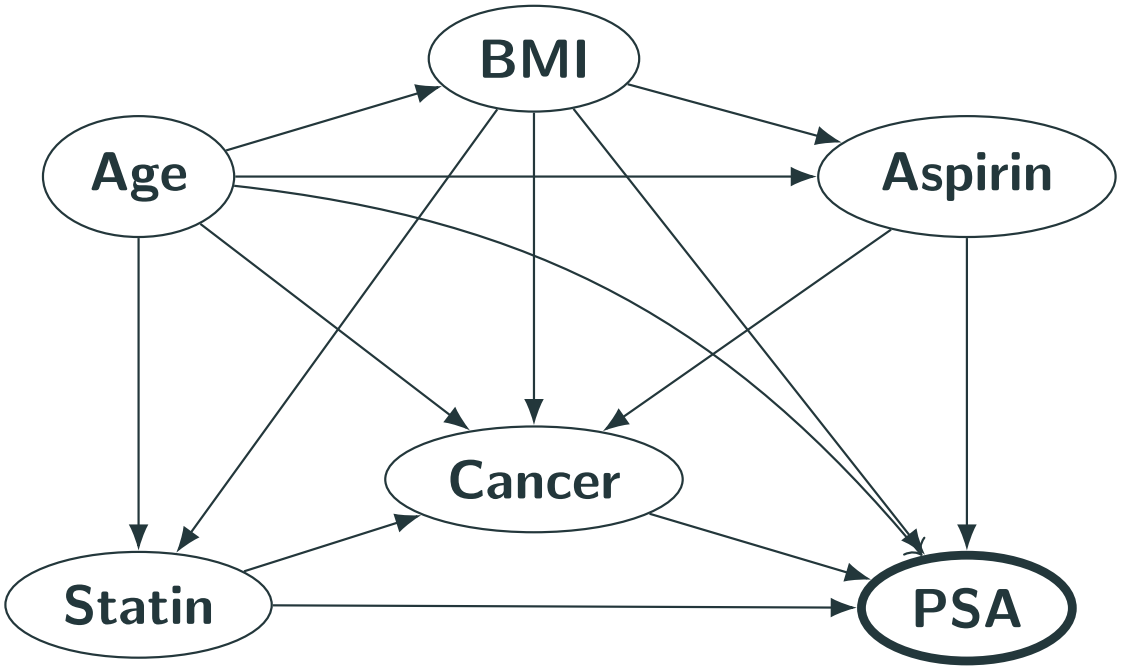
Principle of independent mechanisms

[Haavelmo 1943, Frisch 1948]: *'A structural relation not only explains the observed data, it captures a structure connecting the variables; related to autonomy and invariance.'*

Elements of causal inference book: *'The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each others. In the probabilistic case this means that the conditional distribution of each variable given its causes (mechanism) does not inform or influence the other conditional distributions. In the case of two variables this reduces to an independence between the cause distribution and the mechanism producing the effect distribution.'*

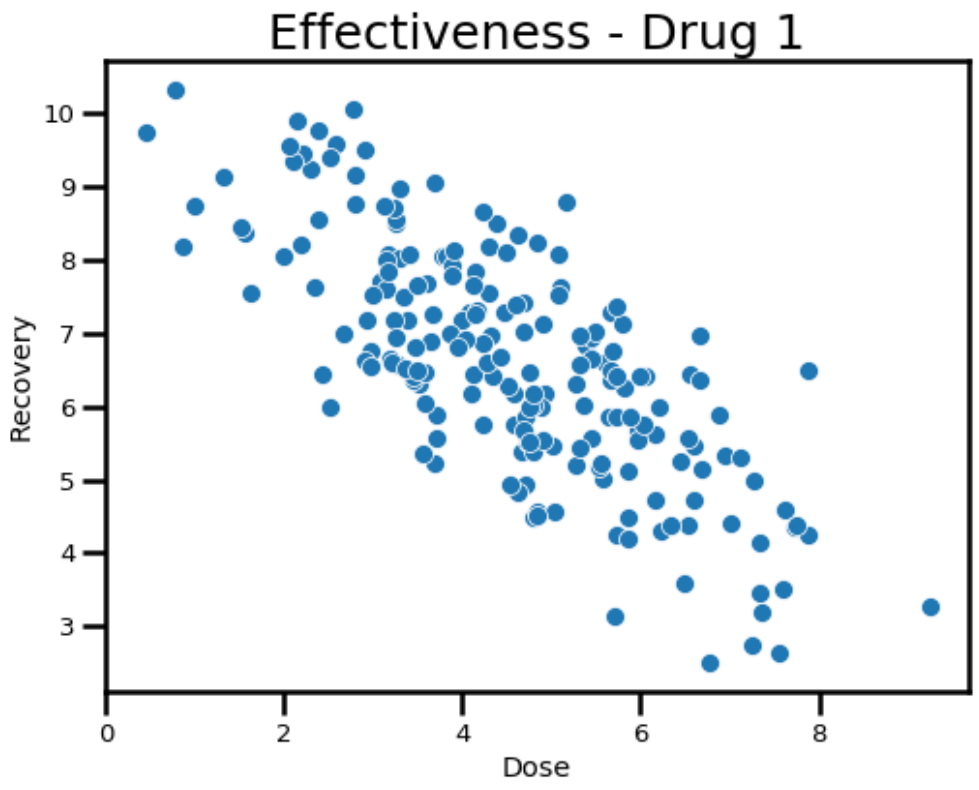
Principle of independent mechanisms

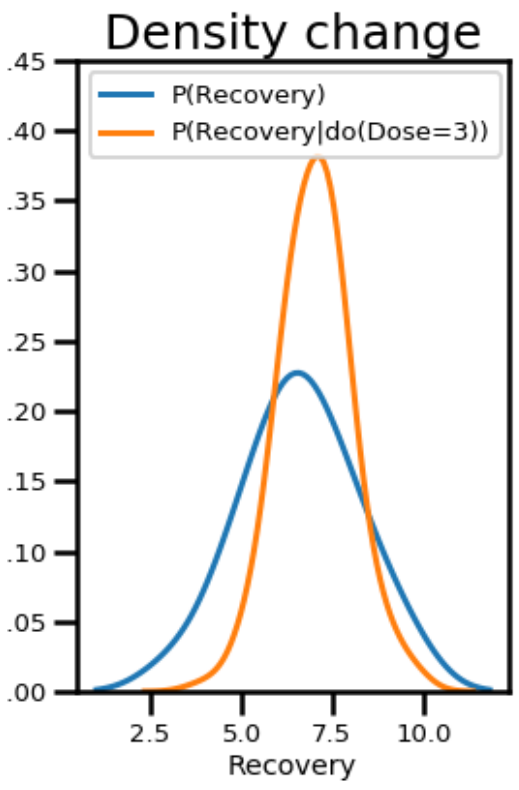
Intervening on one mechanism does not affect any other mechanism.



OK, but what is exactly a causal effect?

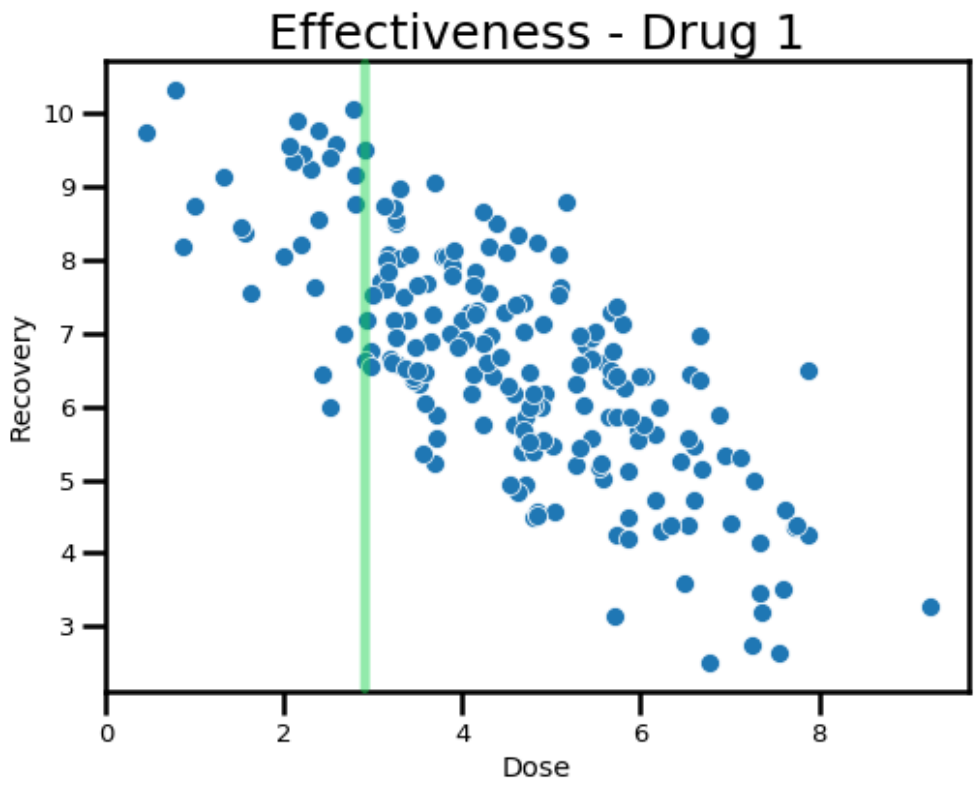
T causally affects Y if **intervening** on T changes the distribution of Y.

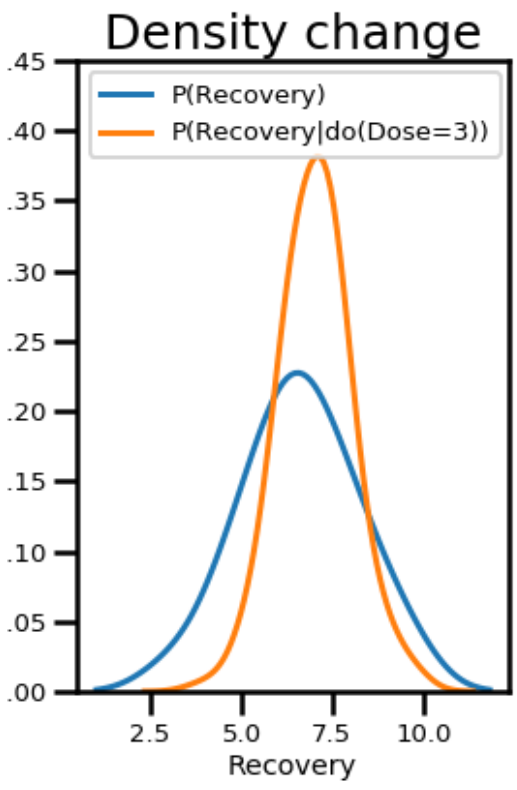




OK, but what is exactly a causal effect?

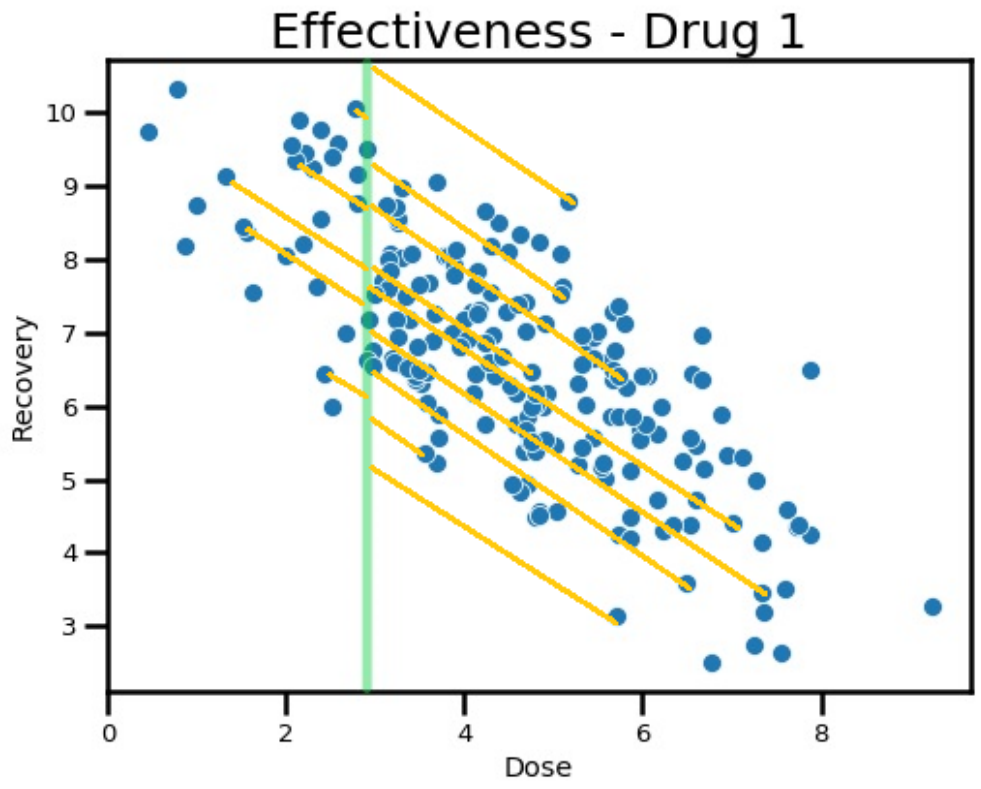
T causally affects Y if **intervening** on T changes the distribution of Y.

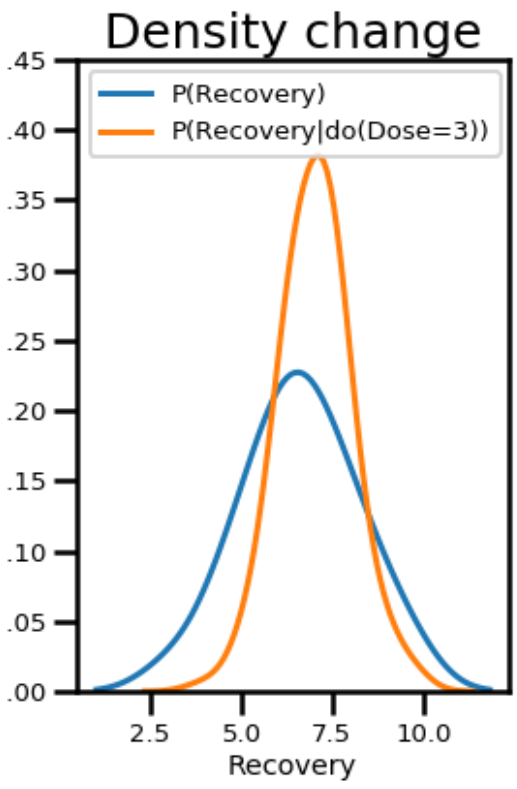




OK, but what is exactly a causal effect?

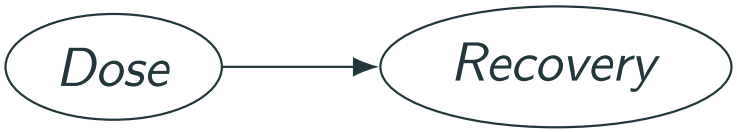
T causally affects Y if **intervening** on T changes the distribution of Y.





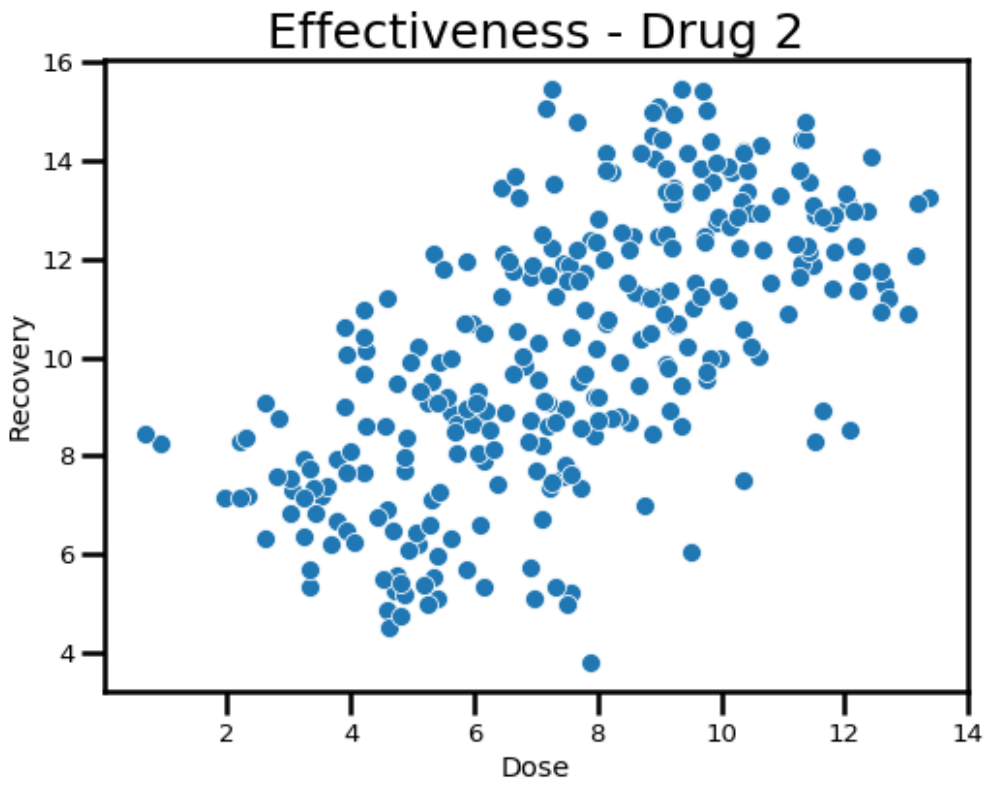
OK, but what is exactly a causal effect?

$$\mathbb{P}(recovery) \neq \mathbb{P}(Recovery|do(Dose = 3))$$



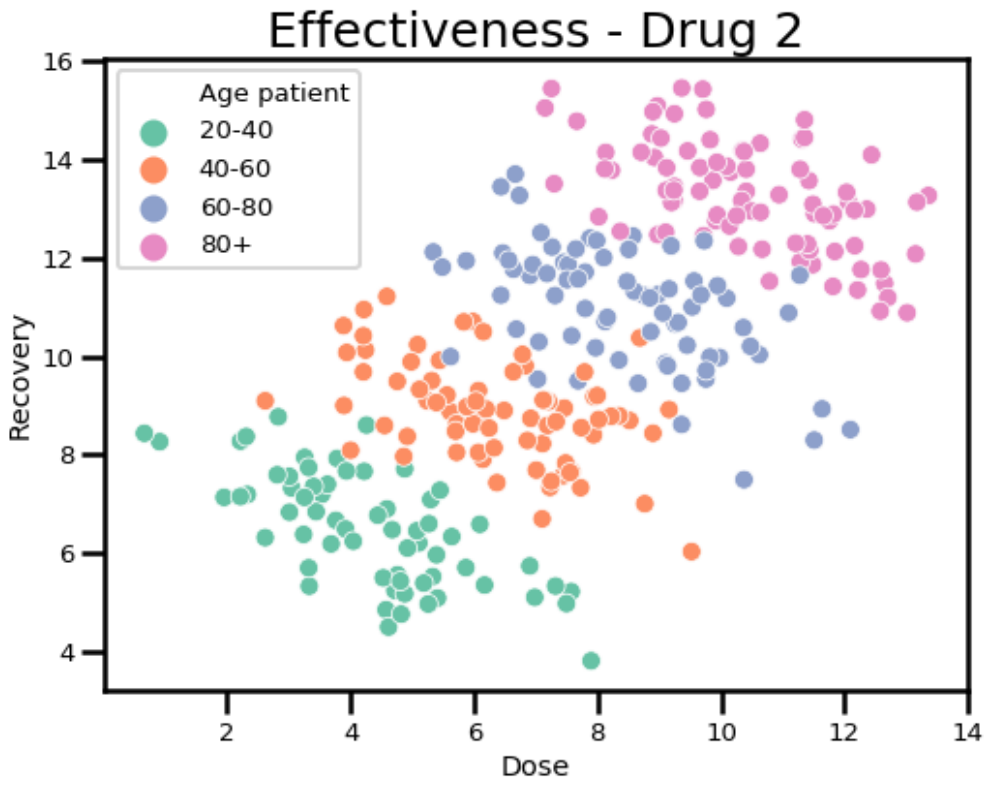
- A causal effect IS a 'physical' mechanisms.
- A causal effect IS NOT a property of the data.
- Intervening = change the laws of physics (not the data).
- *do* notation to represent an intervention (a change in the laws).
- In general $\mathbb{P}(Y|do(T = t)) \neq \mathbb{P}(Y|T = t)$

Another example - drug 2



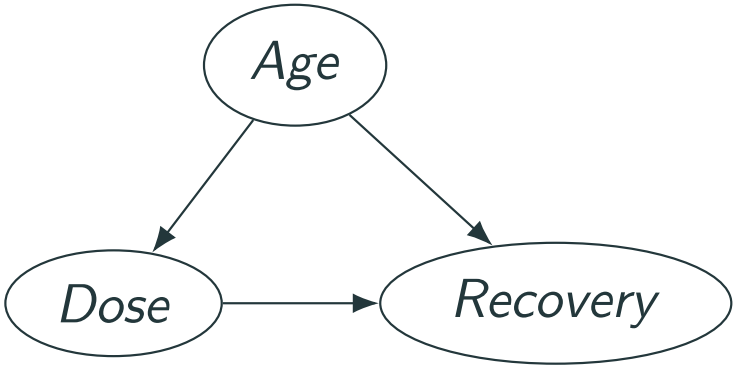
Increasing the dose in drug 2 seems to make patients to spend more time at the hospital (!!).

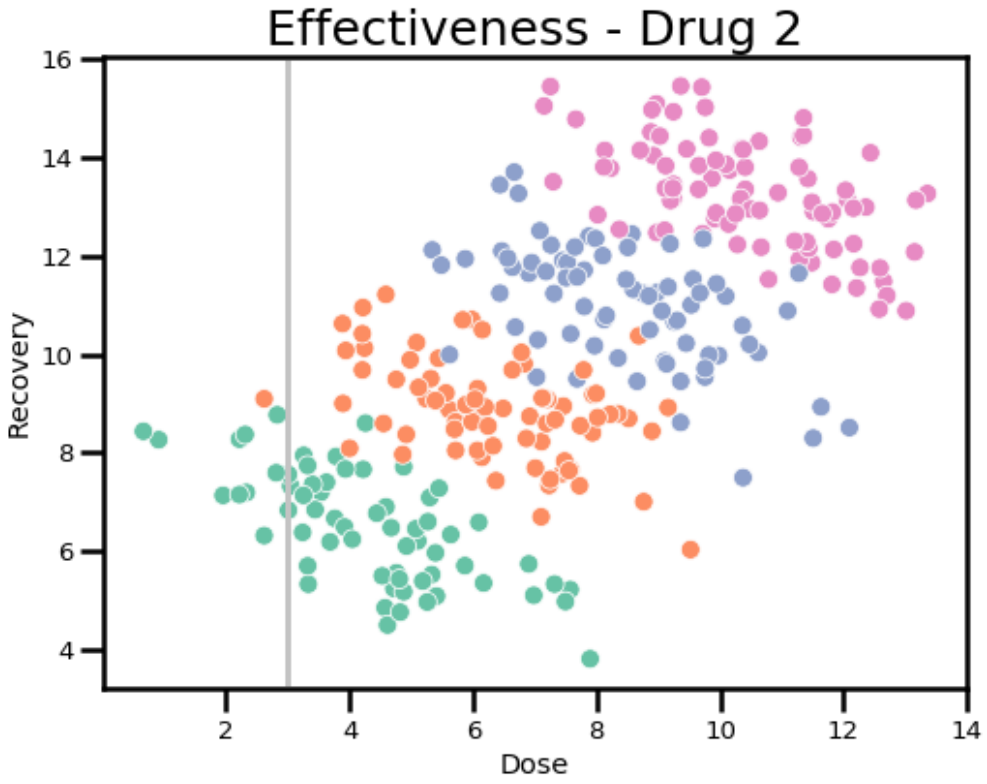
Days of recovery vs Dose - drug 2

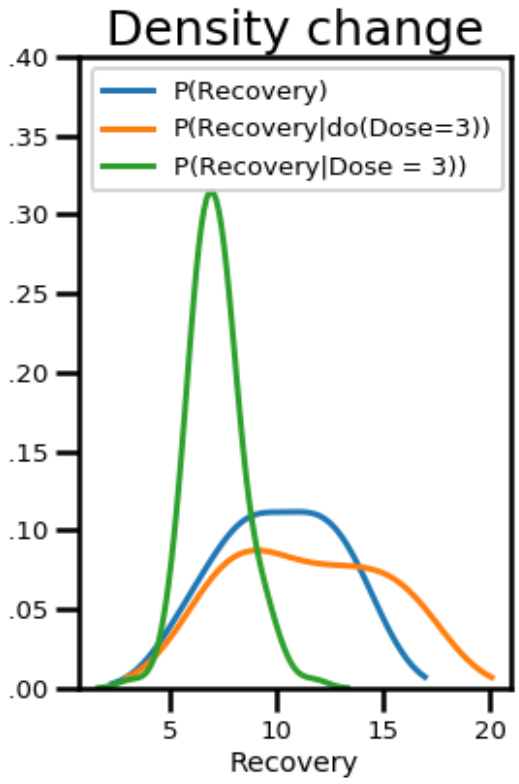


Age is a **confounder**. The drug is effective but older people suffer the disease more severely and require a larger dose.

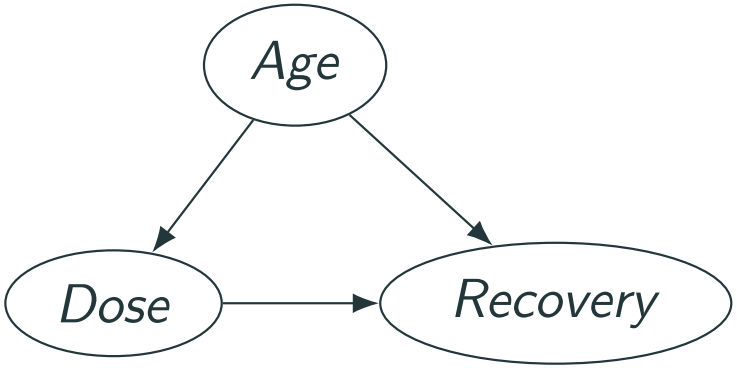
Days of recovery vs. Dose - drug 2

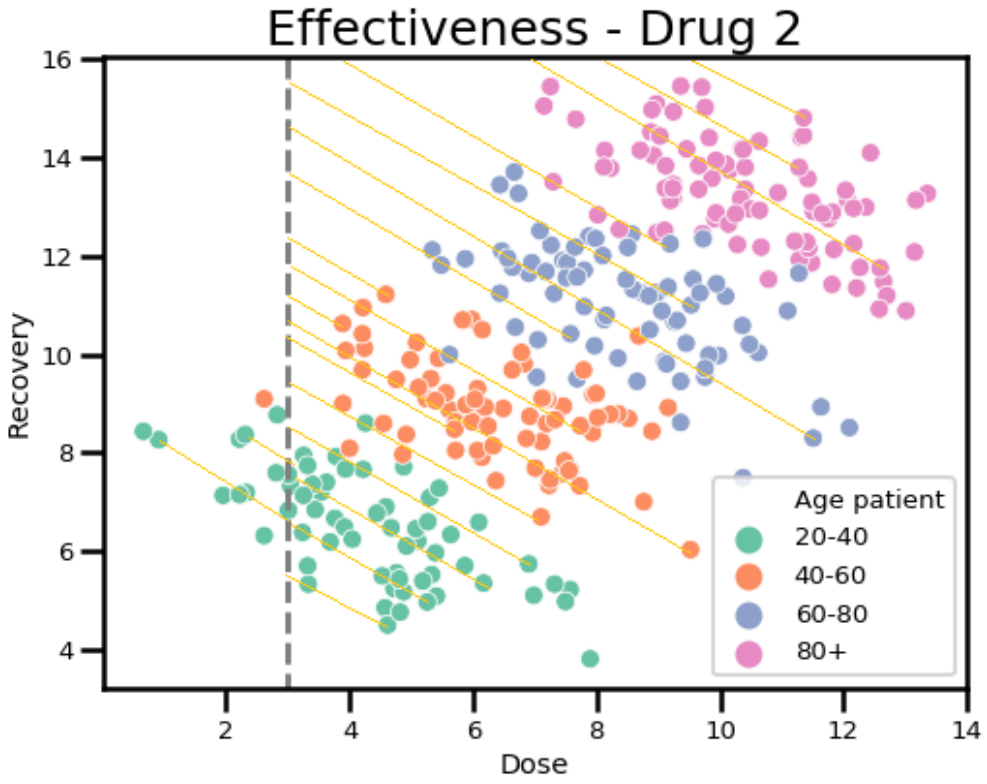


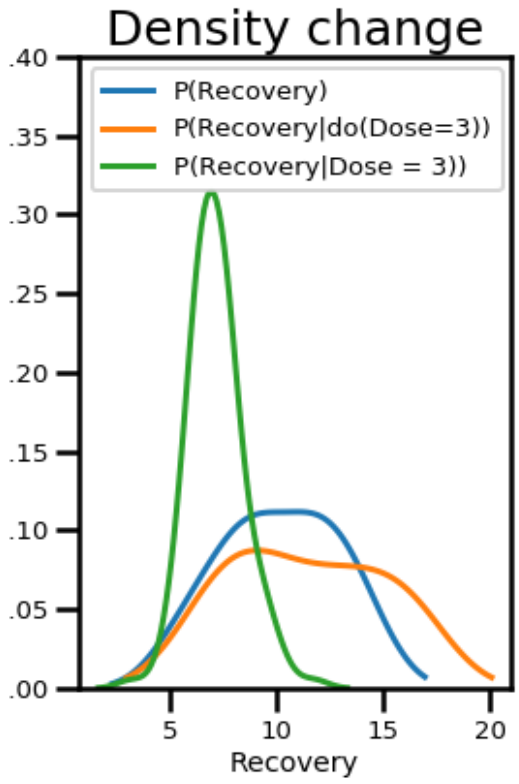




Days of recovery vs. Dose - drug 2

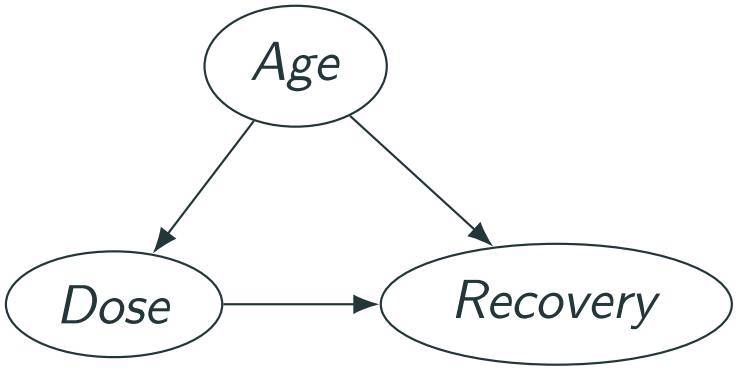




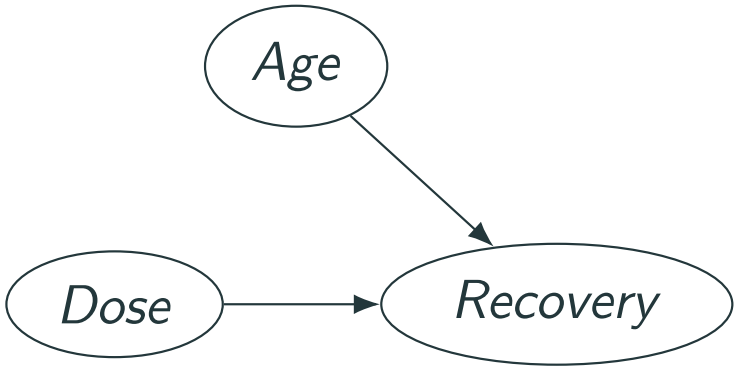


Representing interventions

Observed universe



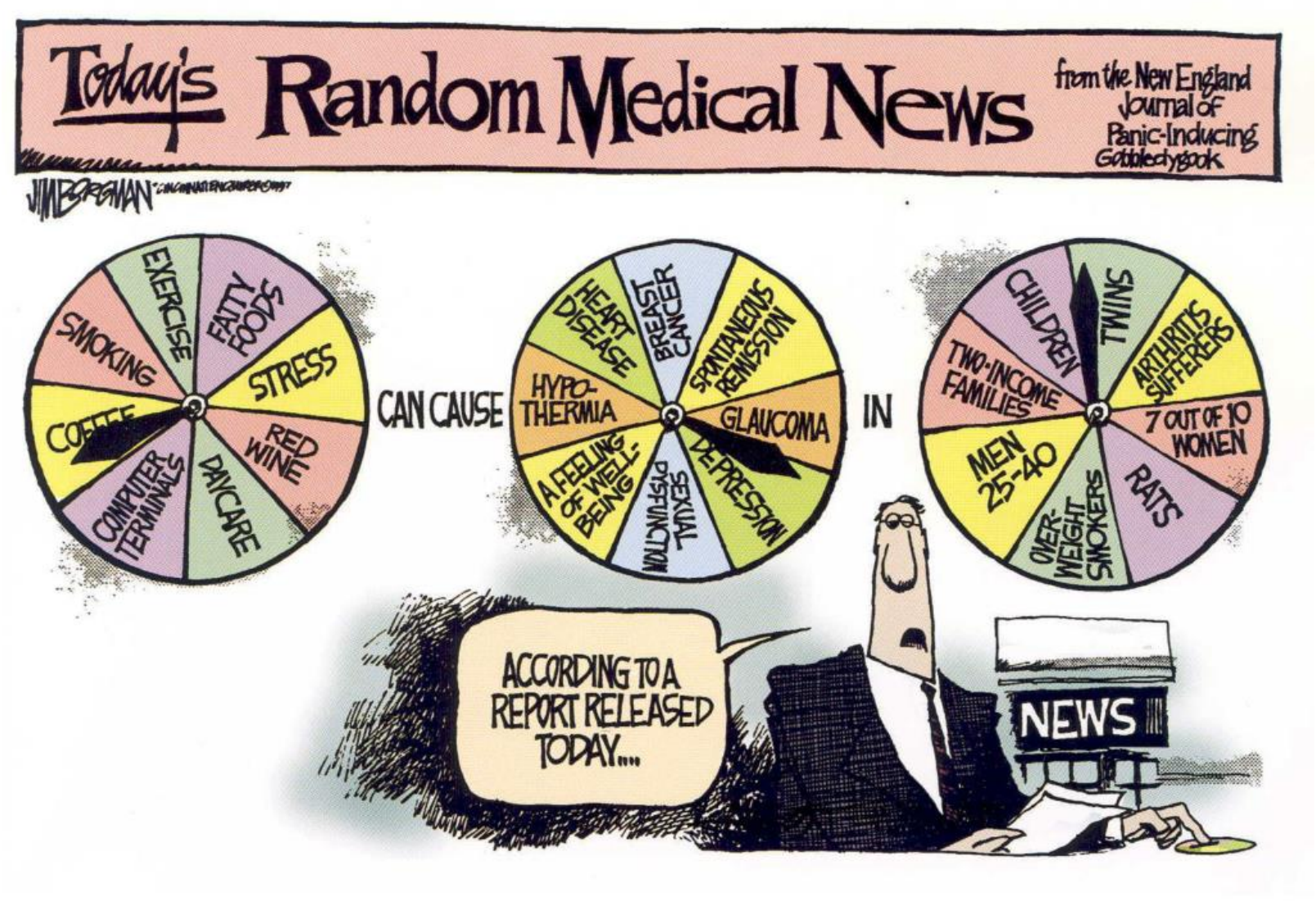
Intervened universe



$$
\begin{aligned}
age &= \epsilon_{age} \\
dose &= f_{dose}(age) + \epsilon_{dose} \\
rec. &= f_{rec.}(age, dose) + \epsilon_{rec.}
\end{aligned}
\qquad
\begin{aligned}
age &= \epsilon_{age} \\
dose &= d \\
rec. &= f_{rec}(age, dose = d) + \epsilon_{rec.}
\end{aligned}
$$

Change the laws of the universe $\rightarrow$ intervention $\rightarrow$ $do(Dose = d)$.

Correlation is not causation ... but is very easy to forget!



Source: Borgman, J (1997). The Cincinnati Enquirer. King Features Syndicate.

Simpson's paradox

Simpson's paradox

'A trend that appears in several different groups of data may disappear or reverse when these groups are combined.'

Example with medical data

Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy

C R CHARIG, D R WEBB, S R PAYNE, J E A WICKHAM

Abstract

This study was designed to compare different methods of treating renal calculi in order to establish which was the most cost effective and successful. Of 1052 patients with renal calculi, 350 underwent open surgery, 350 percutaneous nephrolithotomy, 328 extracorporeal shockwave lithotripsy (ESWL), and 24 both percutaneous nephrolithotomy and ESWL. Treatment was defined as successful if stones were eliminated or reduced to less than 2 mm after three months. Success was achieved in 273 (78%) patients after open surgery, 289 (83%) after percutaneous nephrolithotomy, 301 (92%) after ESWL, and 15 (62%) after percutaneous nephrolithotomy and ESWL. Comparative total costs to the NHS were estimated as £3500 for open surgery, £1861 for percutaneous nephrolithotomy, £1789 for ESWL, and £3210 for both ESWL and nephrolithotomy. ESWL caused no blood loss and little morbidity and is the cheapest and quickest way of returning patients to normal life.

kidney stones to be broken up in situ by using focused shock waves generated by an ultrashort, high tension underwater electrical discharge; it obviates the need for invasive surgery, although an anaesthetic is still required.[1] In all, 750 treatments have now been performed at the London Stone Centre and over 80 000 have been performed world wide (Dornier Systems, West Germany).

With any new and capitally expensive treatment three questions need to be answered: What is the difference in mortality? What is the difference in morbidity? Which mode of treatment is the most cost effective? We tried to define the current place of extracorporeal shockwave lithotripsy (ESWL) in the management of renal calculi by comparing 350 cases of open stone removal, 350 cases of percutaneous nephrolithotomy, and 352 cases of ESWL. All patients were treated by the same team of surgeons under the direct supervision of one consultant. Patients in all groups were unselected and treated consecutively over 14 years (open surgery from 1972 to 1980, percutaneous nephrolithotomy from 1980 to 1985, and ESWL in 1985).

Kidney stones

Success recovery rates of two treatments for kidney stones:

Treatment A	**Treatment B**
78% (273/350)	**83% (289/350)**

Which treatment is better?

Kidney stones

Success recovery rates of two treatments for kidney stones:

Treatment A	**Treatment B**
78% (273/350)	**83% (289/350)**

Which treatment is better?

Treatment B

Kidney stones

Success recovery rates of two treatments for kidney stones:

Treatment A	**Treatment B**
78% (273/350)	**83% (289/350)**

Which treatment is better?

Treatment B

Wait, are we sure? let's have a look to the data again....

Confounders

When the less effective treatment (B) is applied more frequently to less severe cases, it can appear to be a more effective treatment.

	Treatment A	Treatment B
Small stones	**93% (81/87)**	87% (234/270)
Large stones	**73% (192/263)**	69% (55/80)
Total	78% (273/350)	**83% (289/350)**

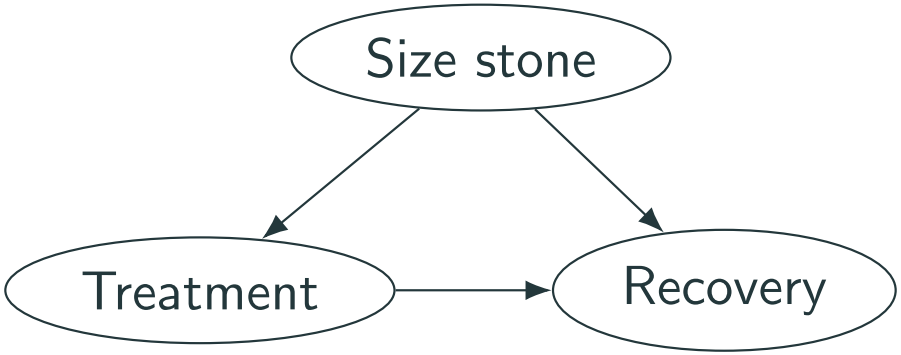
The size of the stone is a confounder.

Confounders

When the less effective treatment (B) is applied more frequently to less severe cases, it can appear to be a more effective treatment.

	Treatment A	Treatment B
Small stones	**93% (81/87)**	87% (234/270)
Large stones	**73% (192/263)**	69% (55/80)
Total	78% (273/350)	**83% (289/350)**

The size of the stone is a confounder.



Solution

Weighting the effect of each treatment by the number of cases.

	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Total	78% (273/350)	83% (289/350)

Solution

Weighting the effect of each treatment by the number of cases.

	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Total	78% (273/350)	83% (289/350)

$$\begin{aligned}\mathbb{P}(\mathit{Recover}|\mathit{do}(T = A)) &= \mathbb{P}(\mathit{small})\mathbb{P}(\mathit{Recover}|\mathit{small}, A) \\ &+ \mathbb{P}(\mathit{big})\mathbb{P}(\mathit{Recover}|\mathit{big}, A) \\ &= \mathbf{0.8325}\end{aligned}$$

Solution

Weighting the effect of each treatment by the number of cases.

	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Total	78% (273/350)	83% (289/350)

$$
\begin{aligned}
\mathbb{P}(\mathit{Recover}|\mathit{do}(T=A)) &= \mathbb{P}(\mathit{small})\mathbb{P}(\mathit{Recover}|\mathit{small},A) \\
&+ \mathbb{P}(\mathit{big})\mathbb{P}(\mathit{Recover}|\mathit{big},A) \\
&= \mathbf{0.8325} \\
\mathbb{P}(\mathit{Recover}|\mathit{do}(T=B)) &= \mathbb{P}(\mathit{small})\mathbb{P}(\mathit{Recover}|\mathit{small},B) \\
&+ \mathbb{P}(\mathit{big})\mathbb{P}(\mathit{Recover}|\mathit{big},B) \\
&= \mathbf{0.7788}
\end{aligned}
$$

Treatment A is indeed better.

Berkson's paradox

Berkson's paradox

'Two independent events A and B may become dependent when conditioning on a common effect (collider)'.

Bias in research studies

J Chron Dis Vol. 32, pp. 51 to 63
Pergamon Press Ltd 1979. Printed in Great Britain

BIAS IN ANALYTIC RESEARCH

DAVID L. SACKETT

INTRODUCTION

CASE-CONTROL studies are highly attractive. They can be executed quickly and at low cost, even when the disorders of interest are rare. Furthermore, the execution of pilot case-control studies is becoming automated; strategies have been devised for the 'computer scanning' of large files of hospital admission diagnoses and prior drug exposures, with more detailed analyses carried out in the same data set on an *ad hoc* basis [1]. As evidence of their growing popularity, when one original article was randomly selected from each issue of **The New England Journal of Medicine, The Lancet,** and the **Journal of the American Medical Association** for the years, 1956, 1966 and 1976, the proportion reporting case-control analytic studies increased fourfold over these two decades (2–8%) whereas the proportion reporting cohort analytic studies fell by half (30–15%); incidentally, a general trend toward fewer study subjects but more study authors was also

Berkson's paradox

We know that the is no causal effect between the two diseases:

$$\mathbb{P}(Bone|do(Respiratory = Yes)) = \mathbb{P}(Bone)$$

	General population		
	Bone disease		
Respiratory disease	Yes	No	%Yes
Yes	17	207	**8.4%**
No	184	2376	**7.7%**

Berkson's paradox

	General population			Hospitalizations last 6 months		
	Bone disease			Bone disease		
Respiratory disease	Yes	No	%Yes	Yes	No	%Yes
Yes	17	207	**7.6%**	5	15	**25%**
No	184	2376	**7.2%**	18	219	**7.6%**

Patients going to hospital are less healthy in general and suffer more diseases.

Berkson's paradox

- The respiratory and bone diseases are independent.
- But they are conditionally dependent given hospitalization.

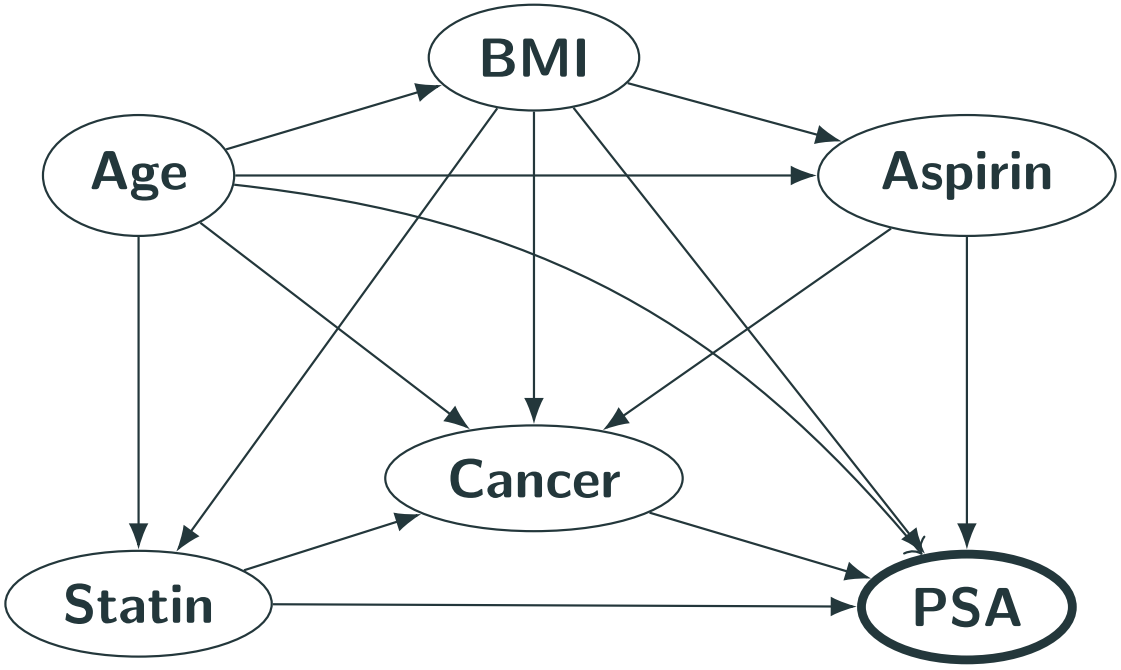
Adjusting by hospitalization is wrong!

$$\mathbb{P}(Bone|do(Re. = Yes)) = \mathbb{P}(Bone) \neq \int \mathbb{P}(Bone|Re. = Yes, Hos.)\mathbb{P}(Hos.)$$

Do-calculus

Do-calculus (Judea Pearl)

- Rules useful to select the minimal set of confounders X in DAGs.



General approach to identify a causal effect

General adjustment formula

If X is a **admissible adjustment set (confounders)** then:

$$\mathbb{P}(Y|do(T=t)) = \sum_X \mathbb{P}(Y|T=t, X=x)\mathbb{P}(X=x)$$

$$\mathbb{P}(Y|do(T=t)) = \int \mathbb{P}(Y|T=t, X)\mathbb{P}(X)dX$$

- 'Mapping' from 'observations' to 'interventions'.
- We should only need to control by X, nothing else (remember Berkson's paradox).

D-separation

A set X of nodes is said to block a path between T and Y if either:

- The path contains at least one arrow-emitting node that is in X or
- The path contains at least one collider that is outside X and has no descendant in X.

D-separation

If X blocks all paths from T to Y , it is said to "d-separate T and Y," and then, T and Y are independent given X, written $T \perp\!\!\!\perp Y|X$

Examples

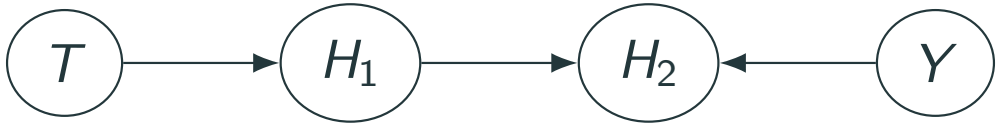
Example 1:

Admissible sets: $\{H_1\}$ and $\{H_2\}$.

$$T \perp\!\!\!\perp Y | H_1$$

$$T \perp\!\!\!\perp Y | H_2$$

Example 2:



Admissible sets: $\{\emptyset\}$

$$T \perp\!\!\!\perp Y$$

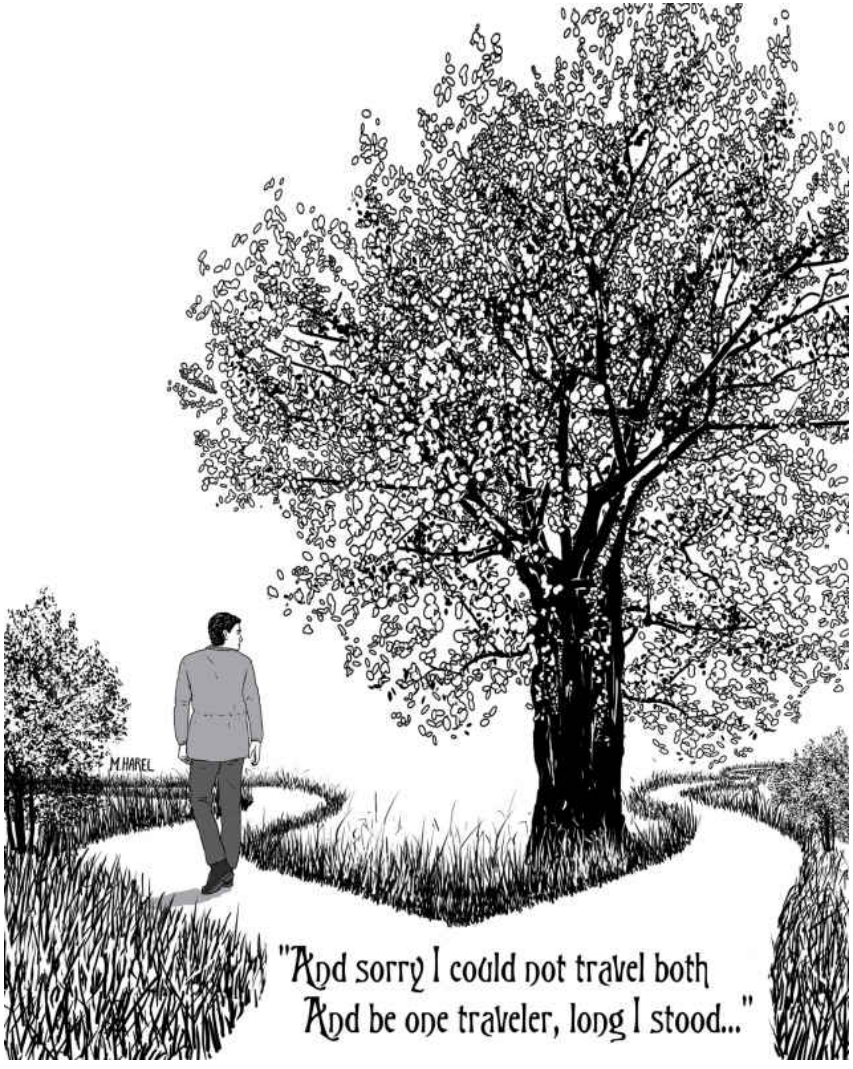
(remember, conditioning on colliders is a bad idea...)

Counterfactuals and potential outcomes framework

Counterfactuals

Counterfactuals.

We cannot travel both roads



Source: Drawing by Maayan Harel.

Potential outcomes framework (Rubin-Neyman)

Each unit X_k has two potential outcomes of a treatment $T_k = \{0, 1\}$:

- $Y_k(0)$ if the unit k is in the control group.
- $Y_k(1)$ if the unit k is in the treatment group.

Observed (factual) outcome (given T_k is observed):

$$Y_k = T_k Y_k(1) + (1 - T_k) Y_k(0)$$

Not observed (counterfactual) outcome:

$$Y_k^* = (1 - T_k) Y_k(1) + T_k Y_k(0)$$

Causality as a missing value problem

Causal inference as a missing value problem.

Patient	Age (X)	Treated (T)	$Y_k(0)$	$Y_k(1)$	Y_i	Individual effect
1	50	0	3	nan	3	?
2	45	0	4	nan	4	?
3	57	0	2.4	nan	2.4	?
4	20	1	nan	7	7	?
5	38	1	nan	8	8	?
6	36	1	nan	9	9	?

Group Mean difference

$$
\begin{aligned}
GMD &:= \mathbb{E}[Y(1)|T=1] - \mathbb{E}[Y(1)|T=0] = \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] \\
&\approx \frac{1}{n_1}\sum_{i=1}^{n_1}[y_i|t_i=1] - \frac{1}{n_0}\sum_{j=1}^{n_0}[y_i|t_i=0]
\end{aligned}
$$

Causal quantities of interest

Individual treatment effect, ITE:

$$\tau_k := Y_k(1) - Y_k(0)$$

Conditional averaged treatment effect, CATE

$$\tau(x) := Y(1) - Y(0)|X = x$$

Averaged treatment effect, ATE

$$\tau := \mathbb{E}[\tau(X)] = \mathbb{E}[Y(1) - Y(0)|X]$$

Selection bias (confounding again)

- $ATCa := \mathbb{E}[Y(1) - Y(0)|T = 1]$, ATE for the cases
- $ATCo := \mathbb{E}[Y(1) - Y(0)|T = 0]$ ATE for the controls.

Then:

$$\underbrace{\mathbb{E}[Y(1)|T = 1] - \mathbb{E}[Y(0)|T = 0]}_{\text{Groups Means Difference}} = \underbrace{\mathbb{E}[Y(1) - Y(0)]}_{\text{Average Treatment Effect}} + \underbrace{\mathbb{E}[Y(0)|T = 1] - \mathbb{E}[Y(0)|T = 0]}_{\text{Selection Bias}} + \underbrace{(1 - \beta)(ATCa - ATCo)}_{\text{Heterogeneous Effect Bias}}$$

Selection bias (confounding again)

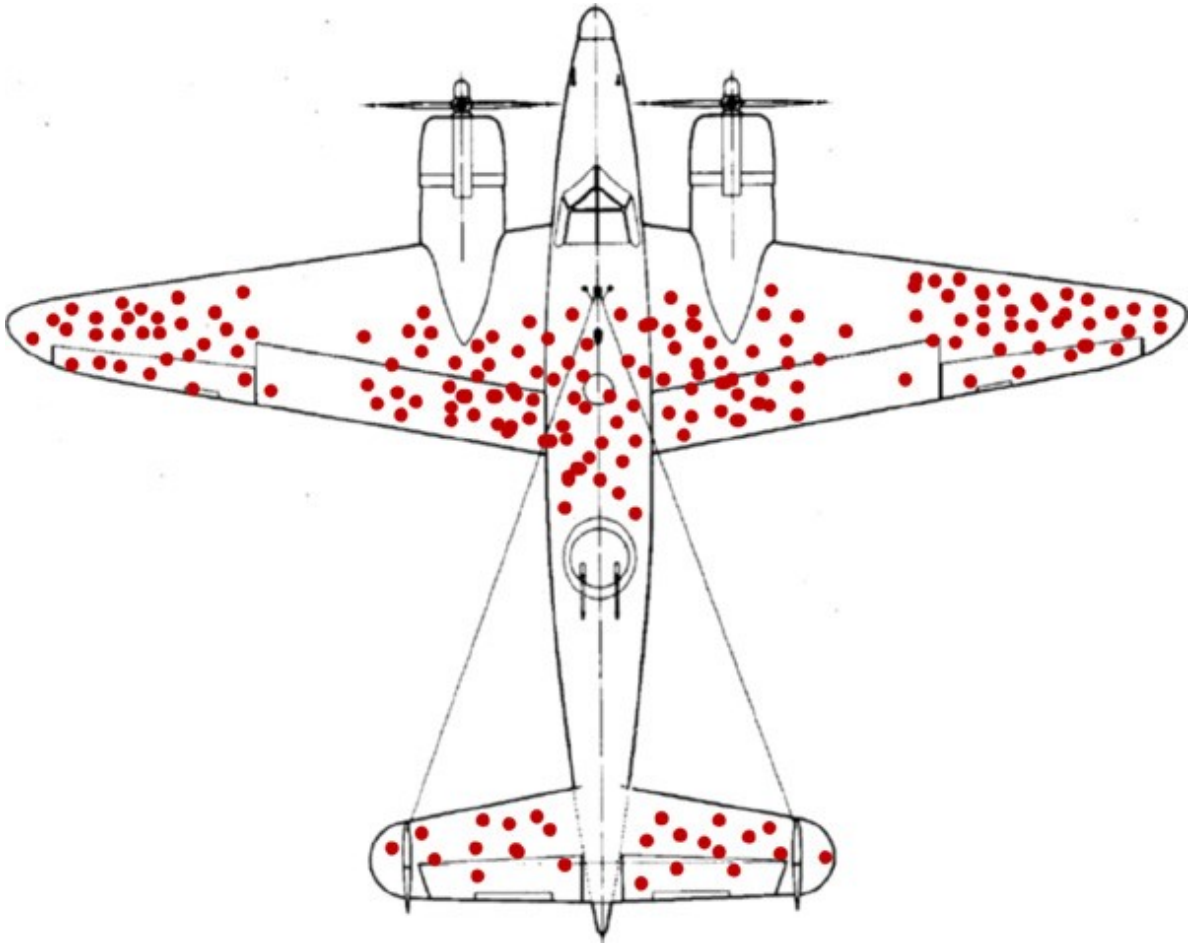
The term

$$\mathbb{E}[Y(0)|T=1] - \mathbb{E}[Y(0)|T=0]$$

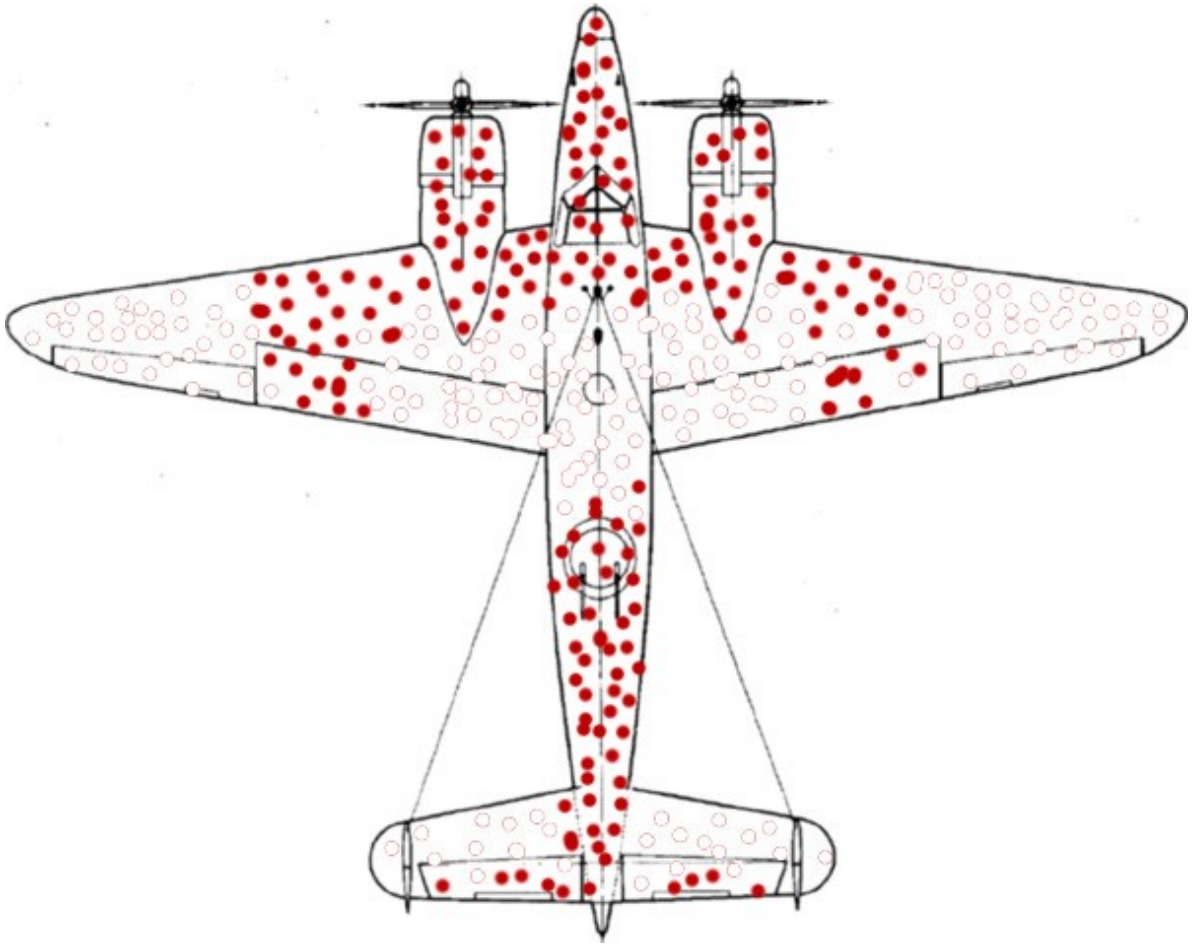
captures the differences between the groups 'before anyone was treated'.

- For the cases this is a counterfactual (never directly estimable).
- If both groups are equal before the treatment this term is zero.
- Called 'selection' because the differences between the groups are caused by a third factor that 'selects' the subjects to be cases or controls (not at random).

Selection bias (confounding again)



Key remark 1: selection bias



Assumptions in the potential outcomes framework

Consistency

If $T_k = t_k$ then $Y_k(t_k) = y_k$ for every individual k and treatment T.

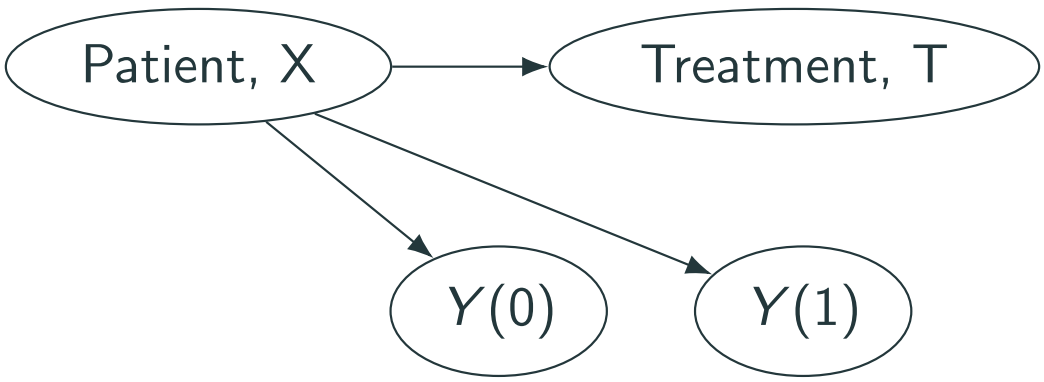
Assumptions in the potential outcomes framework

Unconfoundedness

X: *age, BMI, clinical condition.*

T: *Cancer treatment or placebo.*

$Y(0)$, $Y(1)$: *response under placebo and treatment.*



$$(Y(0), Y(1)) \perp\!\!\!\perp T|X$$

- X contains all the information about the assignment mechanism.
- Not unmeasured confounders.

Assumptions in the potential outcomes framework

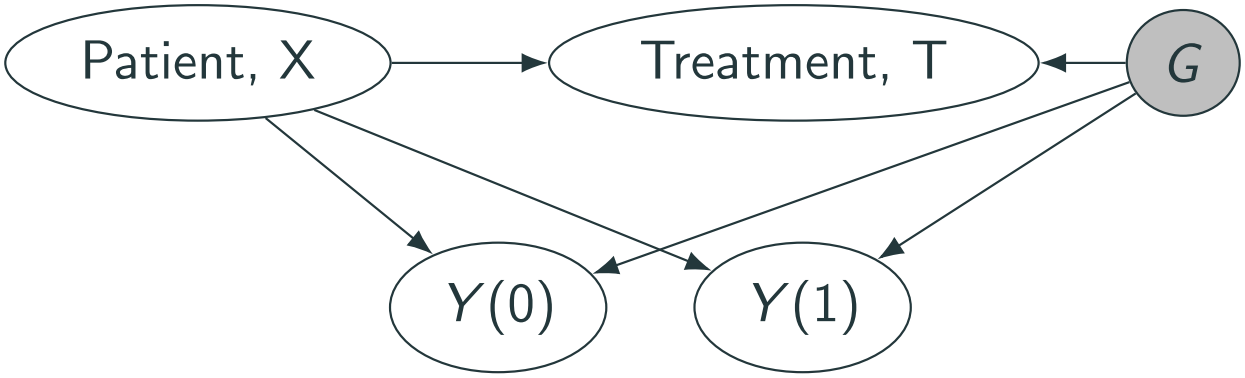
Unconfoundedness

X: *age, BMI, clinical condition.*

T: *Cancer treatment or placebo.*

$Y(0)$, $Y(1)$: *response under placebo and treatment.*

G: *unobserved genetic background.*



$$(Y(0), Y(1)) \not\perp\!\!\!\perp T|X$$

Assumptions in the potential outcomes framework

Positivity and common support

X: *age, BMI, clinical condition.*

T: *Cancer treatment or placebo.*

$Y(0)$, $Y(1)$: *response under placebo and treatment.*

$$1 > \mathbb{P}(T = t|X = x) > 0, \forall t, x$$

Non-zero probability of assignment for all individuals in the population.

Some connection between approaches

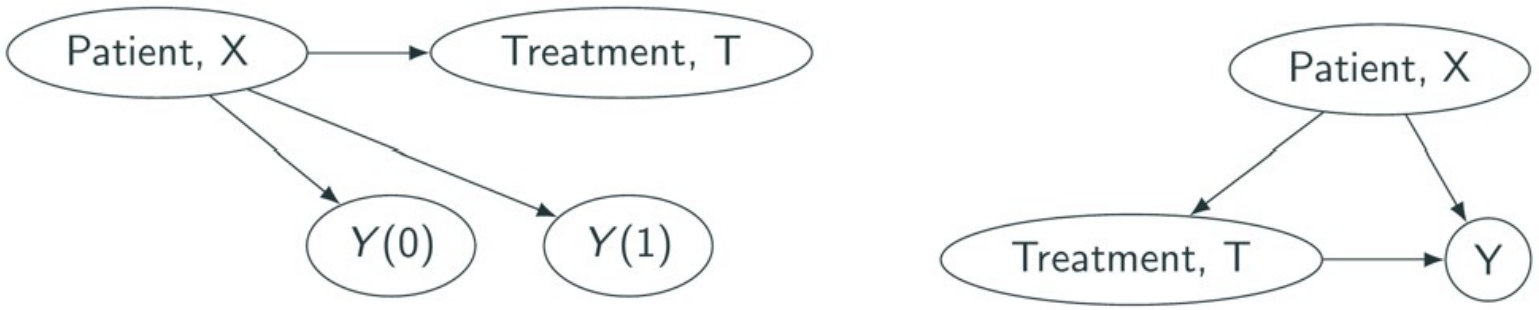
Key connection 1

Counterfactuals can be defined starting with a causal graph. In the potential outcomes framework they are the core building block.

- M_t is modified version of a causal model M, with the equation(s) of T replaced by $T = t$.
- Counterfactual $Y_t(\epsilon) := \Delta Y_{M_t}(\epsilon)$, where ϵ is a realization of the exogenous variables (one specific individual).
- The counterfactual in causal model M is defined as the solution for Y in the "surgically modified" submodel M_t.

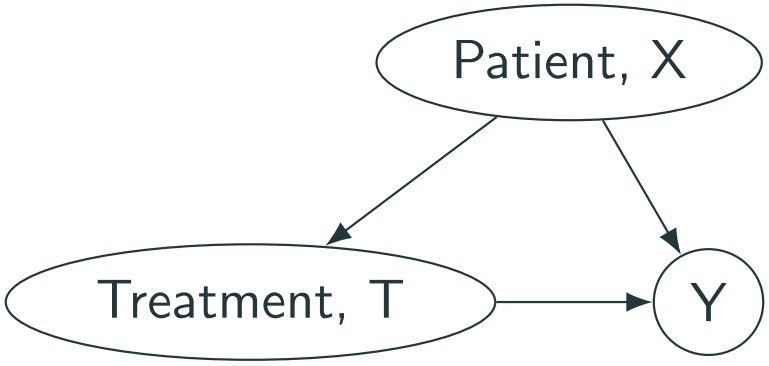
Key connection 2

The unconfoundedness assumption that $(Y(0), Y(1)) \not\perp\!\!\!\perp T|X$ is equivalent to say that X blocks all the backdoor paths between Y and T.



Key connection 3

The average of the potential outcomes $Y(0)$ or $Y(1)$ (factuals and counterfactuals) for the entire populations are the casual effect of the two treatments.



$$
\begin{aligned}
\tau &:= \mathbb{E}[\tau(X)] \\
&= \mathbb{E}[\mathbb{E}[Y(1) - Y(0)|X]] \\
&= \mathbb{E}[\mathbb{E}[Y(1)|X]] - \mathbb{E}[\mathbb{E}[Y(0)|X]] \\
&= \mathbb{E}[\mathbb{E}[Y, T = 1|X]] - \mathbb{E}[\mathbb{E}[Y|X, T = 0]] \\
&= \mathbb{E}[Y|do(T = 1)] - \mathbb{E}[Y|do(T = 0)]
\end{aligned}
$$

How to compute a causal effect...when experimentation is possible

How to compute a causal effect in practice?

Let's first travel back to the earliest 1920s... and have a cup of tea.

Ronald Fisher

Muriel Bristol Roach

Now, a nice story

...He knew she liked milk with tea, so he poured some milk into a cup and added the tea to it. And the trouble started. Bristol refused the cup.
–I won't drink that–, she declared. Fisher, looked at her surprised.
–Why?–
–You poured the milk into the cup first, I never drink tea unless the milk goes in second–.
Fisher thought the debate was nonsense. Following the laws Thermodynamic, mixing A with B was the same as mixing B with A. The final temperature and relative proportions would be identical.
–I am sure that the order doesn't matter.– Fisher replied.
–It does!– she insisted. *–I can taste the difference between tea brewed each way–*.
–That's impossible.– He replied.

Source: Science History Institute. Ronald Fisher, a Bad Cup of Tea, and the Birth of Modern Statistics.

So, Fisher proposed an experiment

- Prepared 8 **randomly** ordered cups of tea.
- 4 prepared by first pouring the tea, 4 by first pouring the milk.
- Bristol had to select 4 cups prepared by one method.

Null hypothesis: order doesn't matter. *Casual question* (although Fisher never calls it 'causality'):

Results of the experiment

Fisher prepared 8 cups of tea and the experiment started.
–Milk first, milk first, tea first, milk first...”– Said Bristol while tasting the cups one after the other.
Fisher couldn't believe it, Muriel got all 4 the cups right. Shocked, he tried to understand what just happened.
–She is right!–, Fisher said.
–There are 70 possible combinations to select the cups. The probability of getting the 8 cups right by chance is only $1/70 = 0.014$*”–*

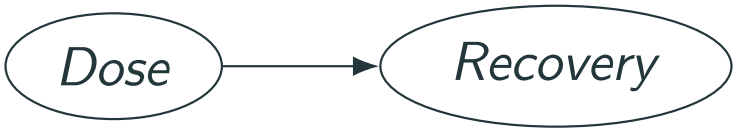
Success count	Combinations of selection	Number of Combinations
0	oooo	1 × 1 = 1
1	ooox, ooxo, oxoo, xooo	4 × 4 = 16
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	6 × 6 = 36
3	oxxx, xoxx, xxox, xxxo	4 × 4 = 16
4	xxxx	1 × 1 = 1
	Total	70

source: Wikipedia

What just happened?

- We have *statistically* demonstrated the existence of causal effect.
- Given the current evidence (data) the probability that the differences found by Bristol in the cups are not real is 0.014 (p-value).
- Randomization is the key. Let's see why.

Hypothesis testing



We have two groups of patients (placebo and treatment):

$$H_0 : \mathbb{E}[Y_0] = \mathbb{E}[Y_1]$$

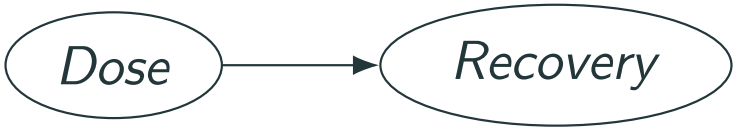
$$H_1 : \mathbb{E}[Y_0] \neq \mathbb{E}[Y_1]$$

- Consider all possible random assignments of patients to the groups.
- We observe one of those randomizations.

Average treatment effect: $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$

Is this effect causal? YES!

Hypothesis testing



We have two groups of patients (placebo and treatment):

$$H_0 : \mathbb{E}[Y_0] = \mathbb{E}[Y_1]$$

$$H_1 : \mathbb{E}[Y_0] \neq \mathbb{E}[Y_1]$$

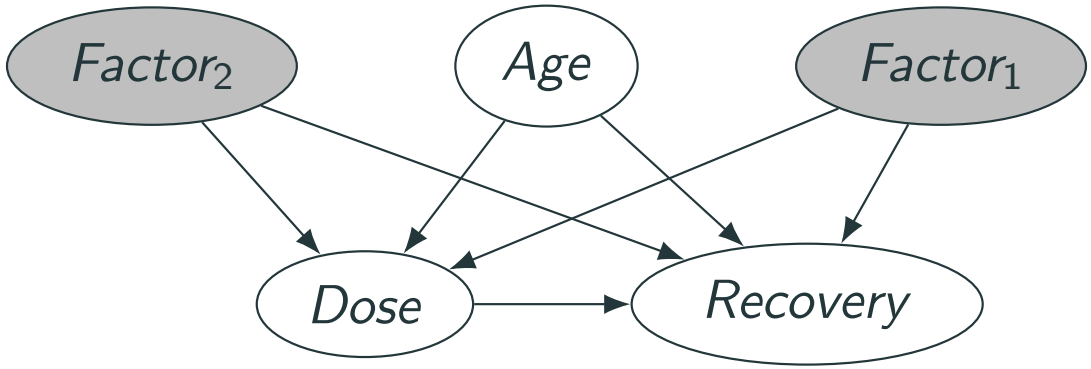
- Consider all possible random assignments of patients to the groups.
- We observe one of those randomizations.

Average treatment effect: $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$

Is this effect causal? YES!

Why? Because of the randomization.

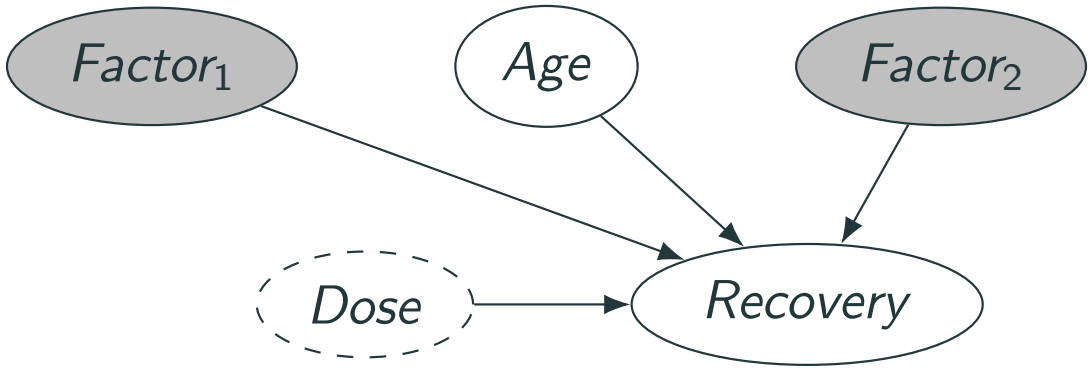
Randomized control trials (RCT)



- What if we don't know the causal graph?
- Or we have unobserved confounders ($Factor_1$ and $Factor_2$)?

We cannot adjust by $Factor_1$ and $Factor_2$ when computing the effect of $Dose$ on $Recovery$.

Randomized control trials (RCT)



- Randomize dose: assign random levels independent of characteristics.
- This 'kills' all the incoming arrows from confounders.
- On the new data the adjustment set is the empty set.

With RCTs, no matter how complex the world is:

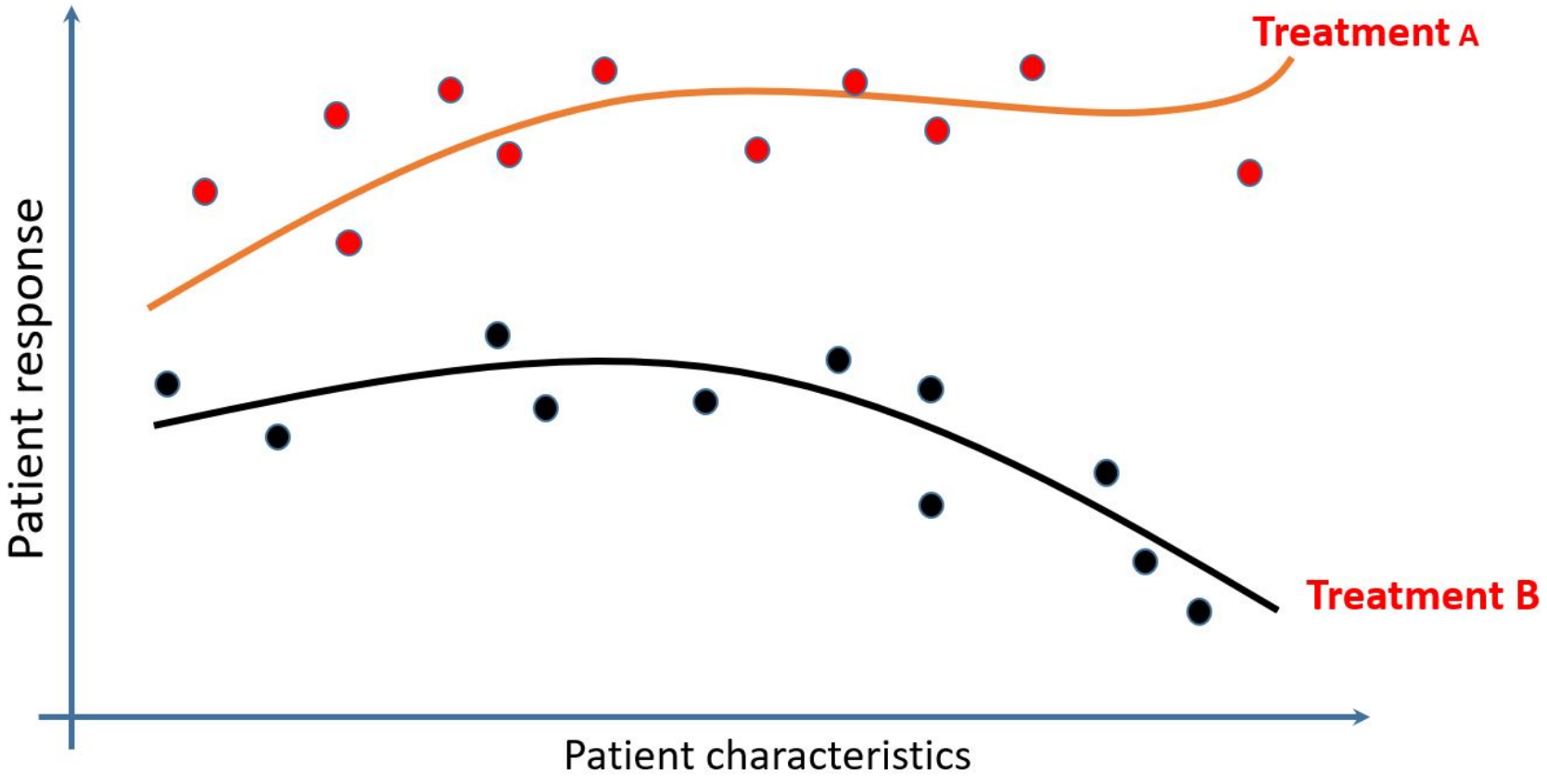
$$\mathbb{P}(Recovery|do(Dose = d)) = \mathbb{P}(Recovery|Dose = d)$$

How to compute causal effect...when experimentation is NOT possible

Fundamental problem of causal inference

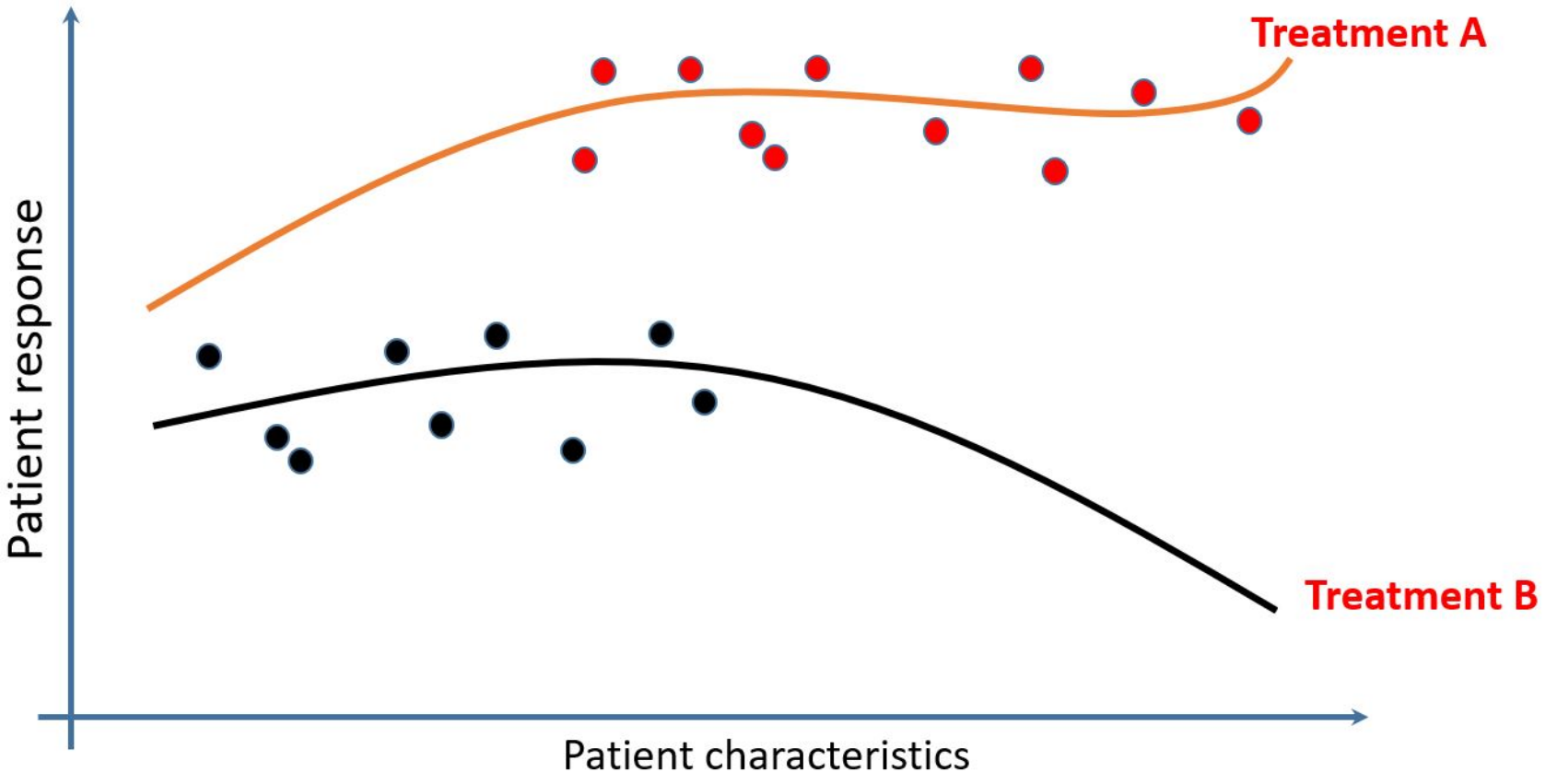
- In most cases we cannot run a RCT or an experiment.
- We have observational data and we only observe one of the two outcomes.
- We have confounders: the value of X affects the probability of assignment.

Example



The patient characteristics don't affect the probability of assignment (like in RCT).

Example

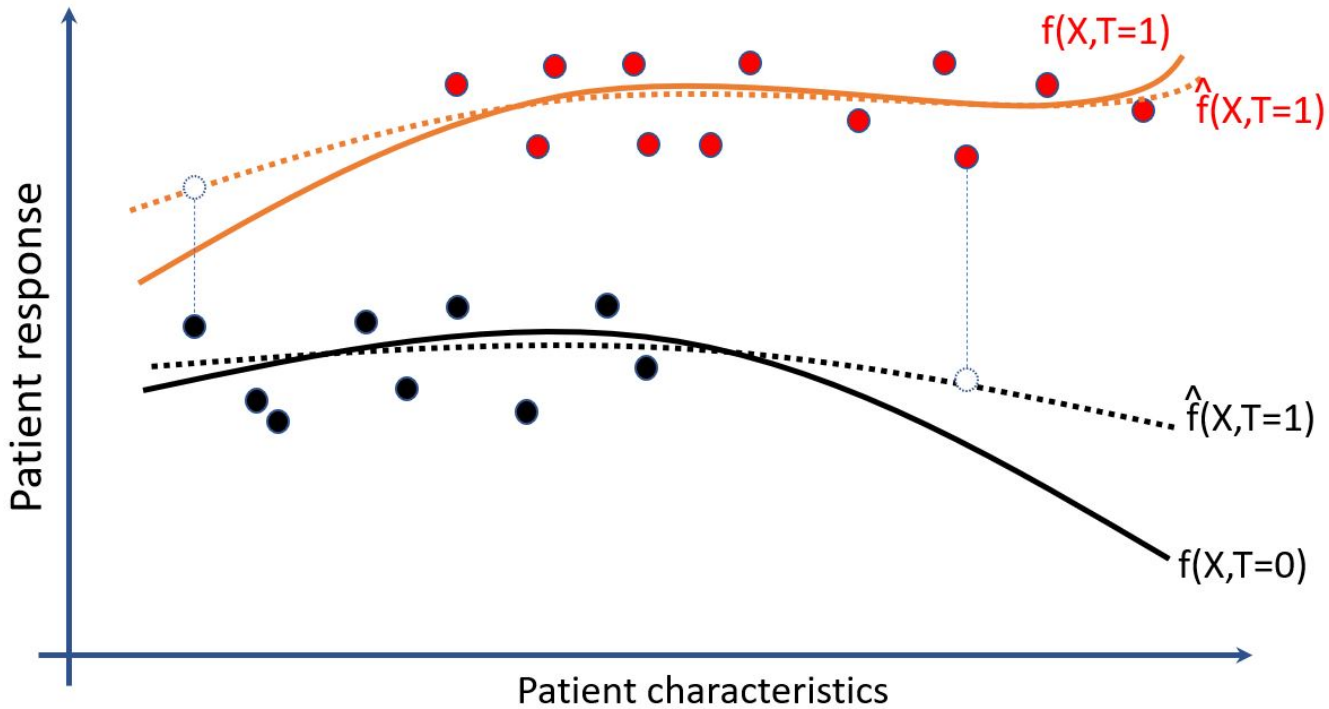


The patient characteristics affect the probability of assignment (confounding).

Old friends, new friends.... and covariate adjustment

Covariate Adjustment: regression to compute causal effects

- Model the relationship between treatments, response and patients
- Treats the causal problem as supervised learning problem.
- Can be used to estimate both ITE, CATE and ATE.



From regression to causation: average treatment effect

T: Treatment

X: Confounders

Y: Response

Let's compute $ATE := \mathbb{E}[Y|do(T = t_1)] - \mathbb{E}[Y|do(T = t_2)]$.

Step 1: Identification.

Find and observe all confounders X or substitute confounders.

From regression to causation: average treatment effect

Step 2: Estimation.

Build a model that predicts the response Y using T, X.

Linear regression: $\mathbb{E}[Y|T,X] = w_0 + \tau T + wX$

Gaussian process: $\mathbb{E}[Y|T,X] = \mu(T,X)$

Let $\hat{\mu}(\cdot)$ be the estimated posterior mean of a Gaussian process.

From regression to causation: average treatment effect

Step 3: Marginalization

$$\text{Approximate } \mathbb{E}_X[\mathbb{E}[Y|T=1,X]] - \mathbb{E}_X[\mathbb{E}[Y|T=0,X]]$$

For a sample $\{t_k, x_k, y_k\}_{k=1}^n$ compute

$$\hat{ATE} = \frac{1}{n}\sum_{k=1}^{n} \hat{\mu}(T=1, X=x_k) - \frac{1}{n}\sum_{k=1}^{n} \hat{\mu}(T=0, X=x_k)$$

Fun fact

If you are using a linear regression model where

$$\mathbb{E}[Y|T, X] = w_0 + \tau T + wX$$

then:

- $\mathbb{E}[Y|do(T = t_1)] = \tau t_1$
- $\frac{\partial \mathbb{E}[Y|do(T=t)]}{\partial t} = \tau$

Linear models are pretty useful to compute causal effects!

More fun facts

Under the same models and assuptions:

$$\hat{ITE}_k = \hat{\mu}(T = 1, X = x_k) - \hat{\mu}(T = 0, X = x_k)$$

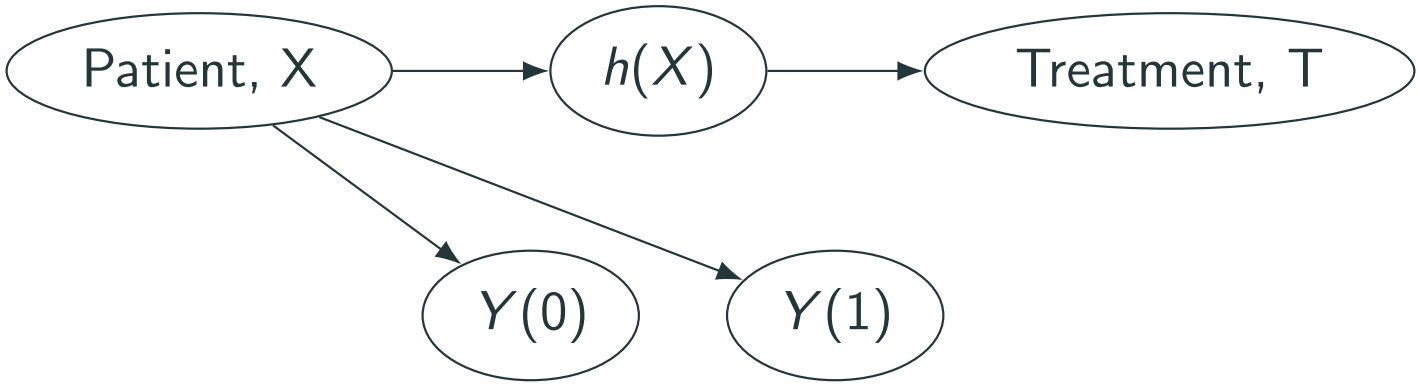
$$\hat{CATE}(x) = \hat{\mu}(T = 1, X = x) - \hat{\mu}(T = 0, X = x)$$

Note: the values of ITE and CATE in regions with low data may be very imprecise and have very high variance.

Old friends, new friends.... and propensity scores

Propensity score

Balancing score: function $h(X)$ such that $X \perp\!\!\!\perp T | h(x)$.



Propensity score: probability of assignment to treatment.

$$e(x) := \mathbb{P}(T = 1 | X = x)$$

Propensity score

$$e(x) := \mathbb{P}(T = 1|X = x)$$

- The propensity $e(x)$ score is a balancing score.
- Coarsest balancing score function (takes a multidimensional object, x_k and transforms it into one dimension).
- Can be estimated with any supervised learning method.
- $(Y(0), Y(1)) \perp\!\!\!\perp T|e(X)$.
- Controlling by X or $e(X)$ is equivalent.

Propensity score re-weighting

In RCTs we now that:

$$\mathbb{P}(T = 1|X = x) = \mathbb{P}(T = 0|X = x) = 0.5$$

Under confounding variables, however:

$$\mathbb{P}(T = 1|X = x) \neq \mathbb{P}(T = 0|X = x) \neq 0.5$$

Idea of reweighing: find $w_0(x)$ and $w_1(x)$ such that:

$$\mathbb{P}(T = 1|X = x) \cdot w_1(x) \approx \mathbb{P}(T = 0|X = x) \cdot w_0(x) \approx 0.5$$

Propensity score re-weighting

Sample with n elements (x_k, t_k, y_k) ($n/2$ treated).

Step 1: Use ML to estimate $\mathbb{P}(T = 1|X = x)$

Step 2:

$$\hat{ATE} = \frac{1}{n} \sum_{k \in G_1} \frac{y_k}{\hat{\mathbb{P}}(T = 1|X = x_k)} - \frac{1}{n} \sum_{k \in G_0} \frac{y_k}{\hat{\mathbb{P}}(T = 0|X = x_k)}$$

Propensity score re-weighting

Sample with n elements (x_k, t_k, y_k) ($n/2$ treated).

Step 1: Use ML to estimate $\mathbb{P}(T = 1|X = x)$

Step 2:

$$\hat{ATE} = \frac{1}{n} \sum_{k \in G_1} \frac{y_k}{\hat{\mathbb{P}}(T = 1|X = x_k)} - \frac{1}{n} \sum_{k \in G_0} \frac{y_k}{\hat{\mathbb{P}}(T = 0|X = x_k)}$$

Reminder: In RCT we have $\mathbb{P}(T = 1|X = x) = 0.5$ and therefore:

$$\hat{ATE} = \frac{1}{n} \sum_{k \in G_1} \frac{y_k}{0.5} - \frac{1}{n} \sum_{k \in G_0} \frac{y_k}{0.5}$$

Propensity score re-weighting

Sample with n elements (x_k, t_k, y_k) ($n/2$ treated).

Step 1: Use ML to estimate $\mathbb{P}(T = 1|X = x)$

Step 2:

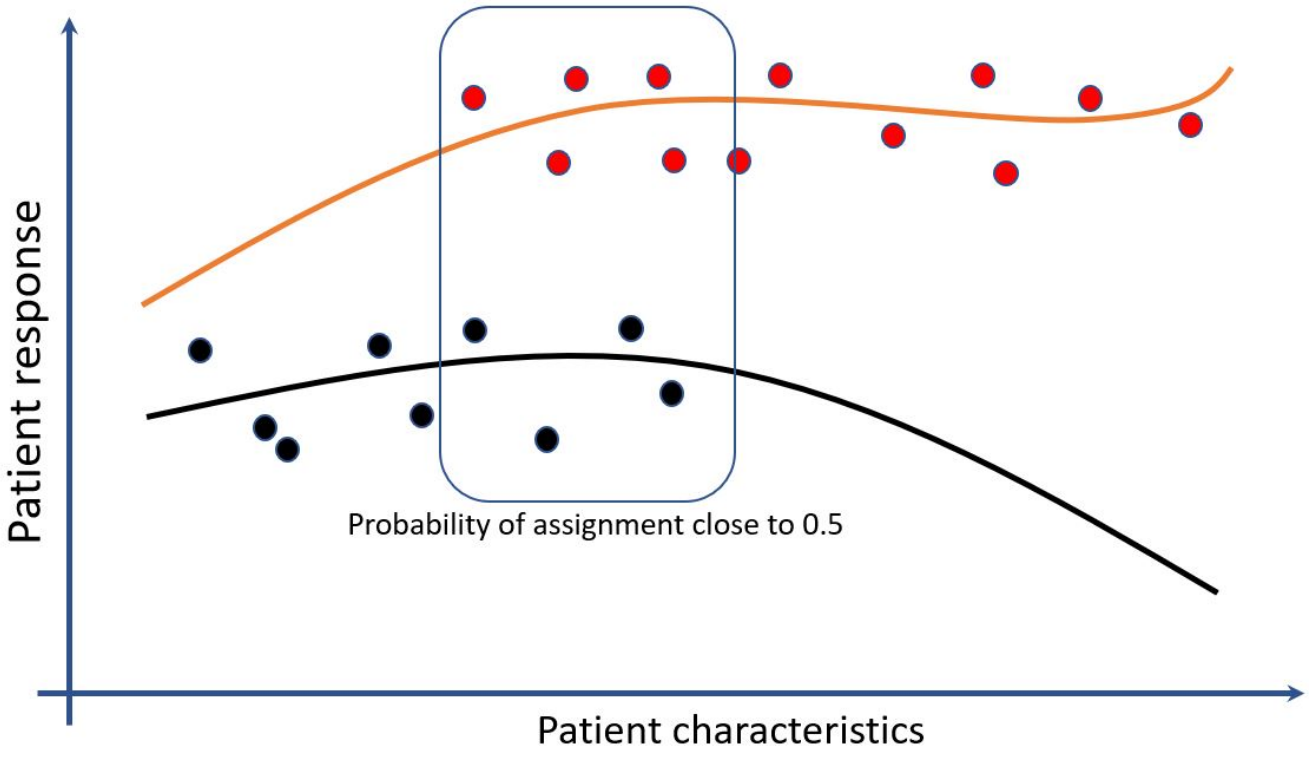
$$\hat{ATE} = \frac{1}{n} \sum_{k \in G_1} \frac{y_k}{\hat{\mathbb{P}}(T = 1|X = x_k)} - \frac{1}{n} \sum_{k \in G_0} \frac{y_k}{\hat{\mathbb{P}}(T = 0|X = x_k)}$$

Reminder: In RCT we have $\mathbb{P}(T = 1|X = x) = 0.5$ and therefore:

$$\begin{aligned} \hat{ATE} &= \frac{1}{n} \sum_{k \in G_1} \frac{y_k}{0.5} - \frac{1}{n} \sum_{k \in G_0} \frac{y_k}{0.5} \\ &= \frac{2}{n} \sum_{k \in G_1} y_k - \frac{2}{n} \sum_{k \in G_0} y_k \end{aligned}$$

Further intuition of the propensity score

- The harder the assignment $\mathbb{P}(T|X) \approx 0.5$ the higher the weight.
- Leverage those points that look more like a random assignment.



Propensity score re-weighting and the backdoor adjustment

Again, all can be derived from the back-door adjustment:

$$ATE = \mathbb{E}[Y|do(T=1)] - \mathbb{E}[Y|do(T=0)]$$

Propensity score re-weighting and the backdoor adjustment

Again, all can be derived from the back-door adjustment:

$$\begin{aligned} ATE &= \mathbb{E}[Y|do(T=1)] - \mathbb{E}[Y|do(T=0)] \\ &= \int Y\mathbb{P}(Y|X,T=1)\mathbb{P}(X)dX - \int Y\mathbb{P}(Y|X,T=0)\mathbb{P}(X)dX \end{aligned}$$

Propensity score re-weighting and the backdoor adjustment

Again, all can be derived from the back-door adjustment:

$$
\begin{aligned}
ATE &= \mathbb{E}[Y|do(T=1)] - \mathbb{E}[Y|do(T=0)] \\
&= \int Y\mathbb{P}(Y|X,T=1)\mathbb{P}(X)dX - \int Y\mathbb{P}(Y|X,T=0)\mathbb{P}(X)dX \\
&= \int Y\frac{\mathbb{P}(Y,X|T=1)}{\mathbb{P}(T=1|X)} - \int Y\frac{\mathbb{P}(Y,X|T=0)}{\mathbb{P}(T=0|X)}
\end{aligned}
$$

Propensity score re-weighting and the backdoor adjustment

Again, all can be derived from the back-door adjustment:

$$
\begin{aligned}
ATE &= \mathbb{E}[Y|do(T=1)] - \mathbb{E}[Y|do(T=0)] \\
&= \int Y\mathbb{P}(Y|X,T=1)\mathbb{P}(X)dX - \int Y\mathbb{P}(Y|X,T=0)\mathbb{P}(X)dX \\
&= \int Y\frac{\mathbb{P}(Y,X|T=1)}{\mathbb{P}(T=1|X)} - \int Y\frac{\mathbb{P}(Y,X|T=0)}{\mathbb{P}(T=0|X)} \\
&\approx \frac{1}{n}\sum_{k\in G_1}\frac{y_k}{\mathbb{P}(T=1|X=x_k)} - \frac{1}{n}\sum_{k\in G_0}\frac{y_k}{\mathbb{P}(T=0|X=x_k)}
\end{aligned}
$$

We use that

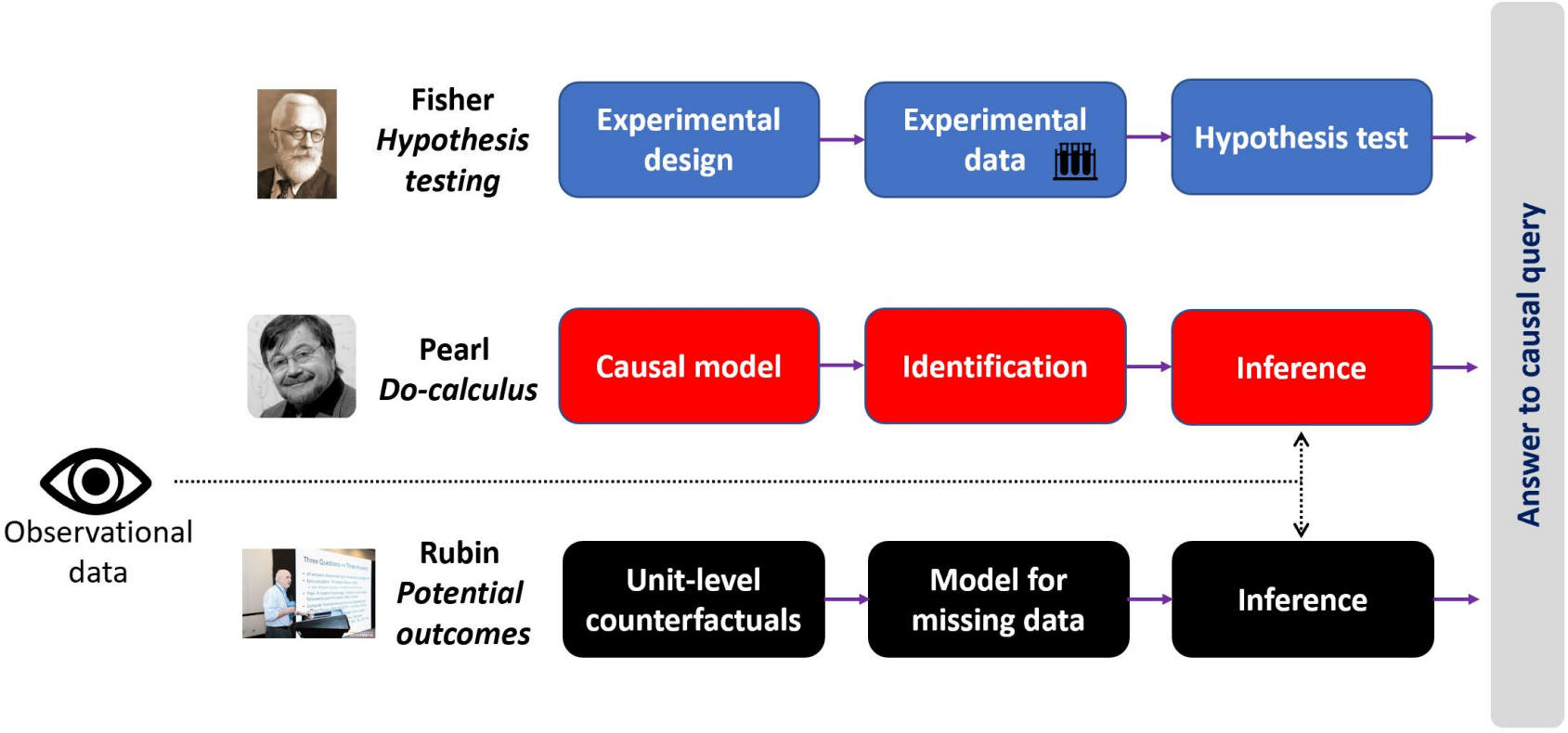
$$
\mathbb{P}(Y|X,T) = \frac{\mathbb{P}(Y,X,T)}{\mathbb{P}(T|X)\mathbb{P}(X)}
$$

Further comments about the propensity score

- Can be used as a metric for matching (propensity score matching).
- Can be used to directly control the causal effect in the back-door.
- Transforms the problem of computing a high dimensional integral, into learning a high dimensional mapping.
- Same idea as importance sampling in machine learning.

Summary of approaches to causal inference

Approaches to causality



Time for some statistical fairy tales...

Statistical fairy tale 1

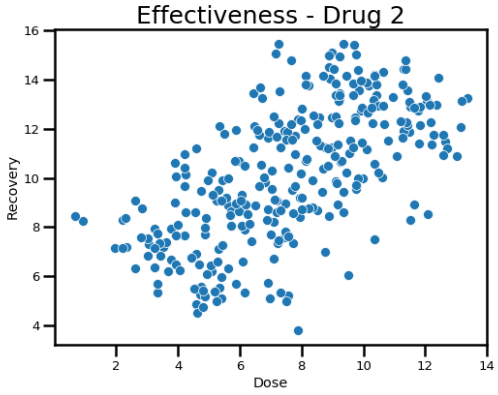
'To estimate an effect all I need is more data points'

Statistical fairy tale 1

'To estimate an effect all I need is more data points'

False!!

Identification and estimation are orthogonal steps.



Statistical fairy tale 2

'To estimate an effect it is fine if I just add all the observed variables to the model'

Statistical fairy tale 2

'To estimate an effect it is fine if I just add all the observed variables to the model'

False!!

Using colliders as confounders may introduce dependencies where they don't exist in the real mechanism.

Statistical fairy tale 3

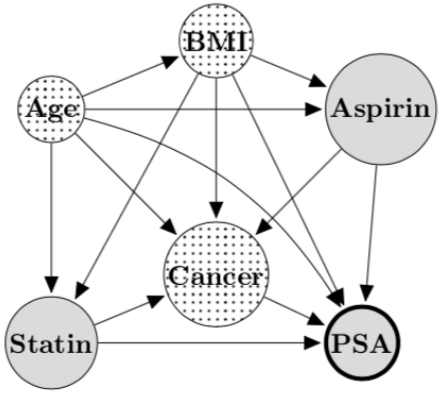
'I can do hypothesis-free causal inference'

Statistical fairy tale 3

'I can do hypothesis-free causal inference'

False!!

Causal inference ALWAYS involve making causal (and modelling) assumptions. These can be made explicit using causal graphs.



Statistical fairy tale 4

'All the validation I need to do, I can do it with my dataset.

Statistical fairy tale 4

'All the validation I need to do, I can do it with my dataset.

False!!

It is usually VERY hard to know if there are unobserved confounders. In those cases, external validation is needed (an experiment).

Unknown unknowns

Statistical fairy tale 5

'A p-value < 0.05 *always tells me something about how the world works'*

Statistical fairy tale 5

'A p-value < 0.05 *always tells me something about how the world works'*

False!! P-values are *sample size* statements. Are only 'causal' if we work with experimental data. Only valid to *falsify* the null hypothesis

Still a debate?

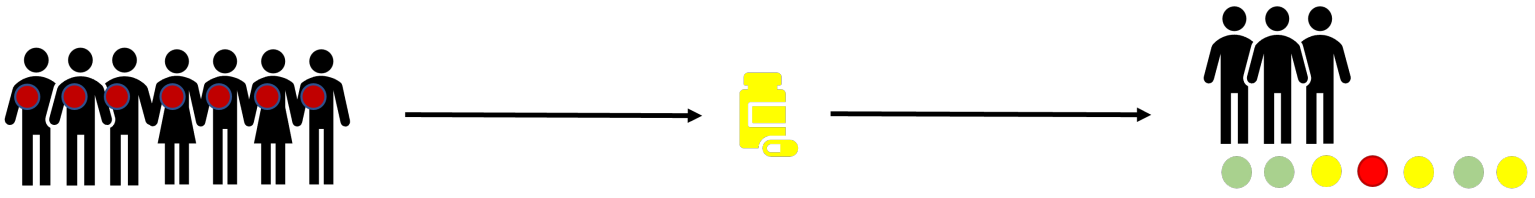
Questions?

Causal inference and precision medicine

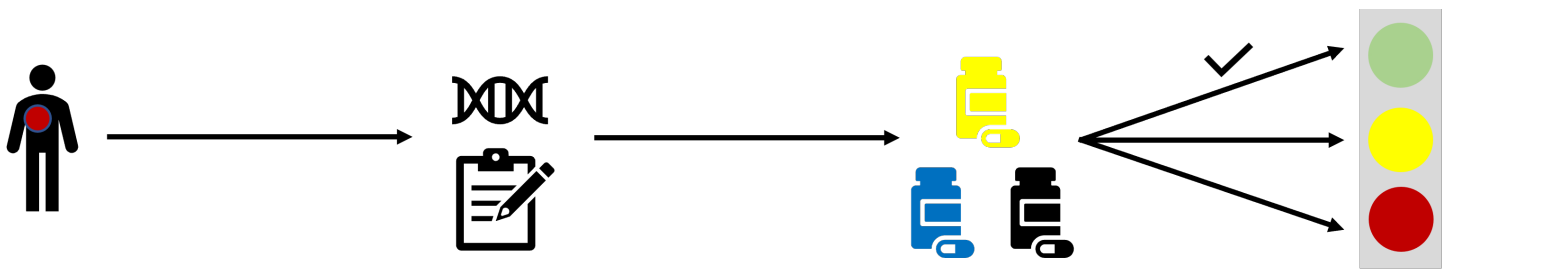
The dream of a better personalized care

The 21st Century Cures Act (CCA) placed additional focus on the use of big real-world data to support decision making and precision medicine

One drug-fit-them-all-model

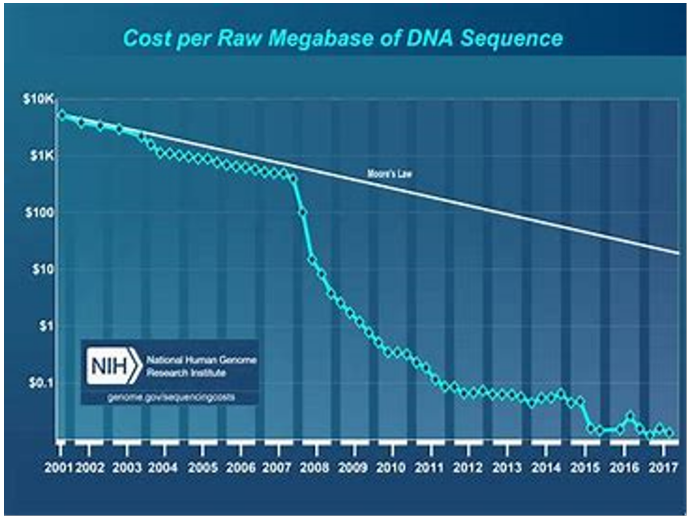


Precision medicine model

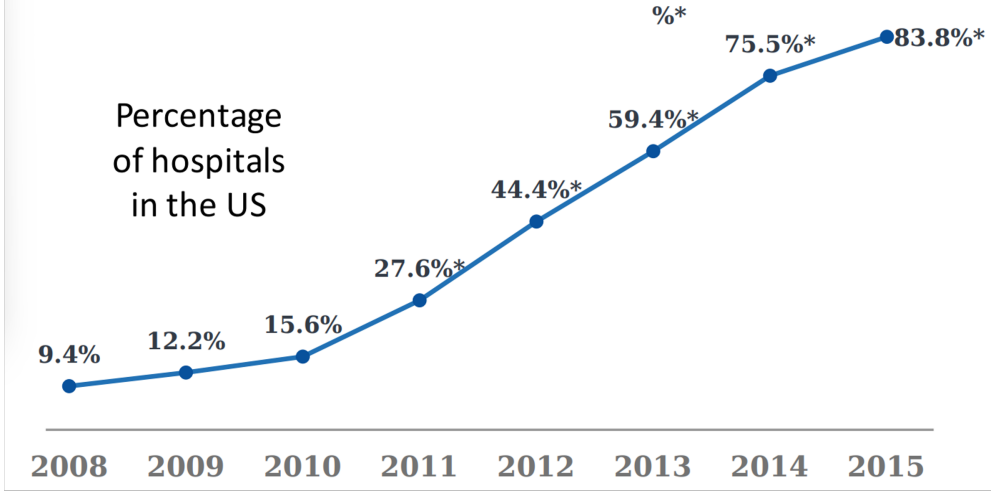


Why precision medicine now?

1. Cheap and automatic data collection at multiple levels



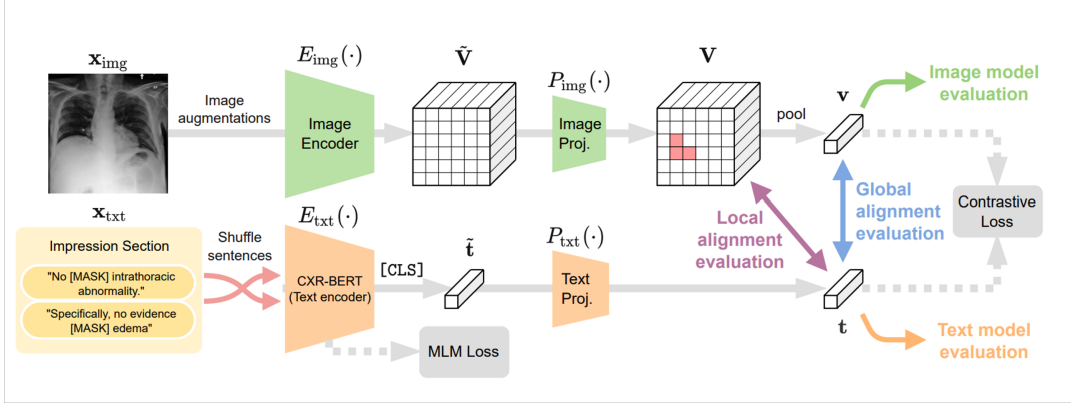
Cheaper to sequence the human genome

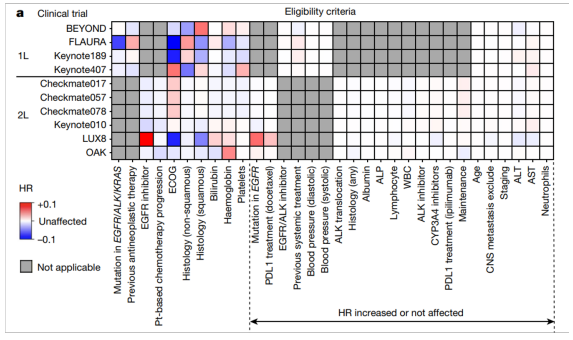


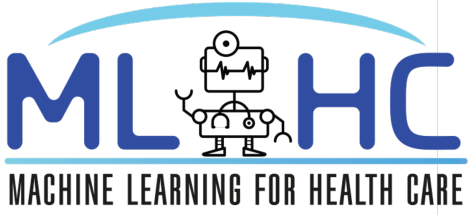
Automatic data collection like medical records, etc.

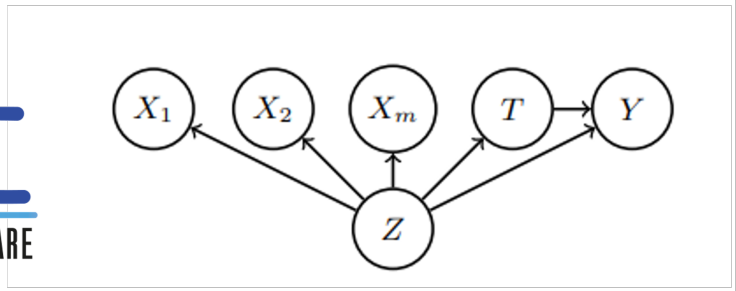
Why precision medicine now?

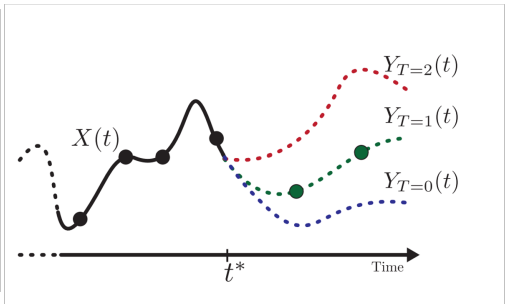
2. Recent advances in machine learning, specialized research workshops





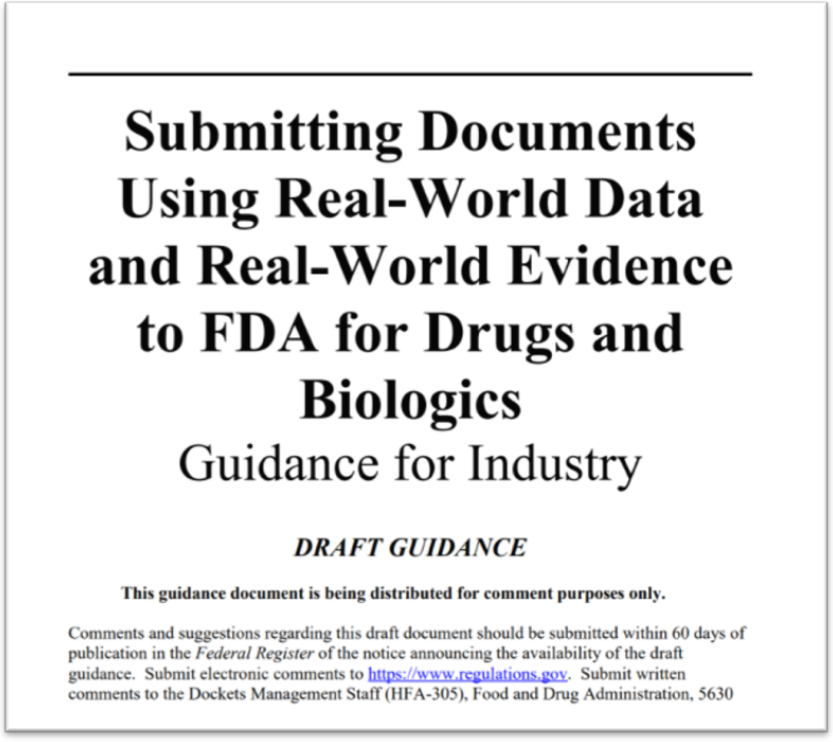






Key opportunity

Use real evidence (RWE) generated from read word data (RWD) to augment traditional randomized control trials (RCTs) to accelerate and personalize new treatments.



Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics

Guidance for Industry

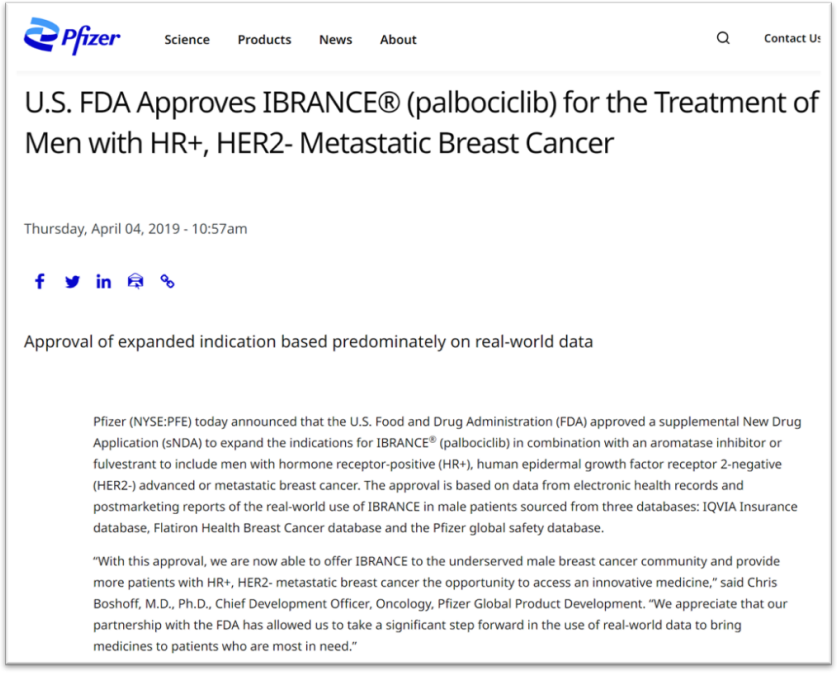
DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630

"RWD and RWE can and should be included if design and analysis is done appropriately."

Recent success stories in the use of Real-world evidence (RWE)



Pfizer Science Products News About Contact Us

U.S. FDA Approves IBRANCE® (palbociclib) for the Treatment of Men with HR+, HER2- Metastatic Breast Cancer

Thursday, April 04, 2019 - 10:57am

Approval of expanded indication based predominately on real-world data

Pfizer (NYSE:PFE) today announced that the U.S. Food and Drug Administration (FDA) approved a supplemental New Drug Application (sNDA) to expand the indications for IBRANCE® (palbociclib) in combination with an aromatase inhibitor or fulvestrant to include men with hormone receptor-positive (HR+), human epidermal growth factor receptor 2-negative (HER2-) advanced or metastatic breast cancer. The approval is based on data from electronic health records and postmarketing reports of the real-world use of IBRANCE in male patients sourced from three databases: IQVIA Insurance database, Flatiron Health Breast Cancer database and the Pfizer global safety database.

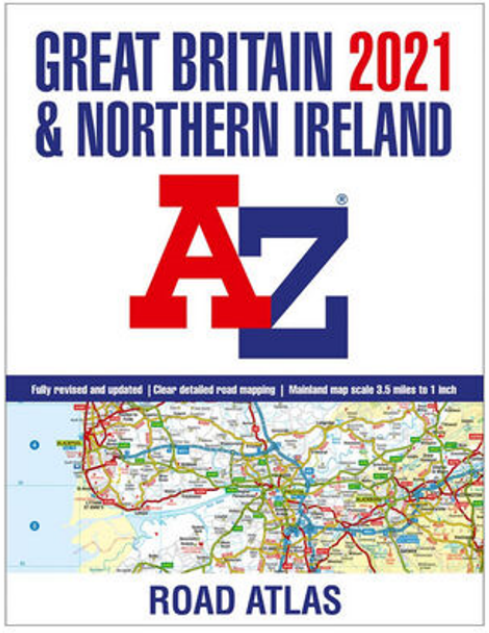
"With this approval, we are now able to offer IBRANCE to the underserved male breast cancer community and provide more patients with HR+, HER2- metastatic breast cancer the opportunity to access an innovative medicine," said Chris Boshoff, M.D., Ph.D., Chief Development Officer, Oncology, Pfizer Global Product Development. "We appreciate that our partnership with the FDA has allowed us to take a significant step forward in the use of real-world data to bring medicines to patients who are most in need."

Pfizer used data from electronic health records (Flatiron Health breast cancer) and post market data (collected from medical practice) to get approval for a new drug for male breast cancer patients.

One map-fit-them-all travelling

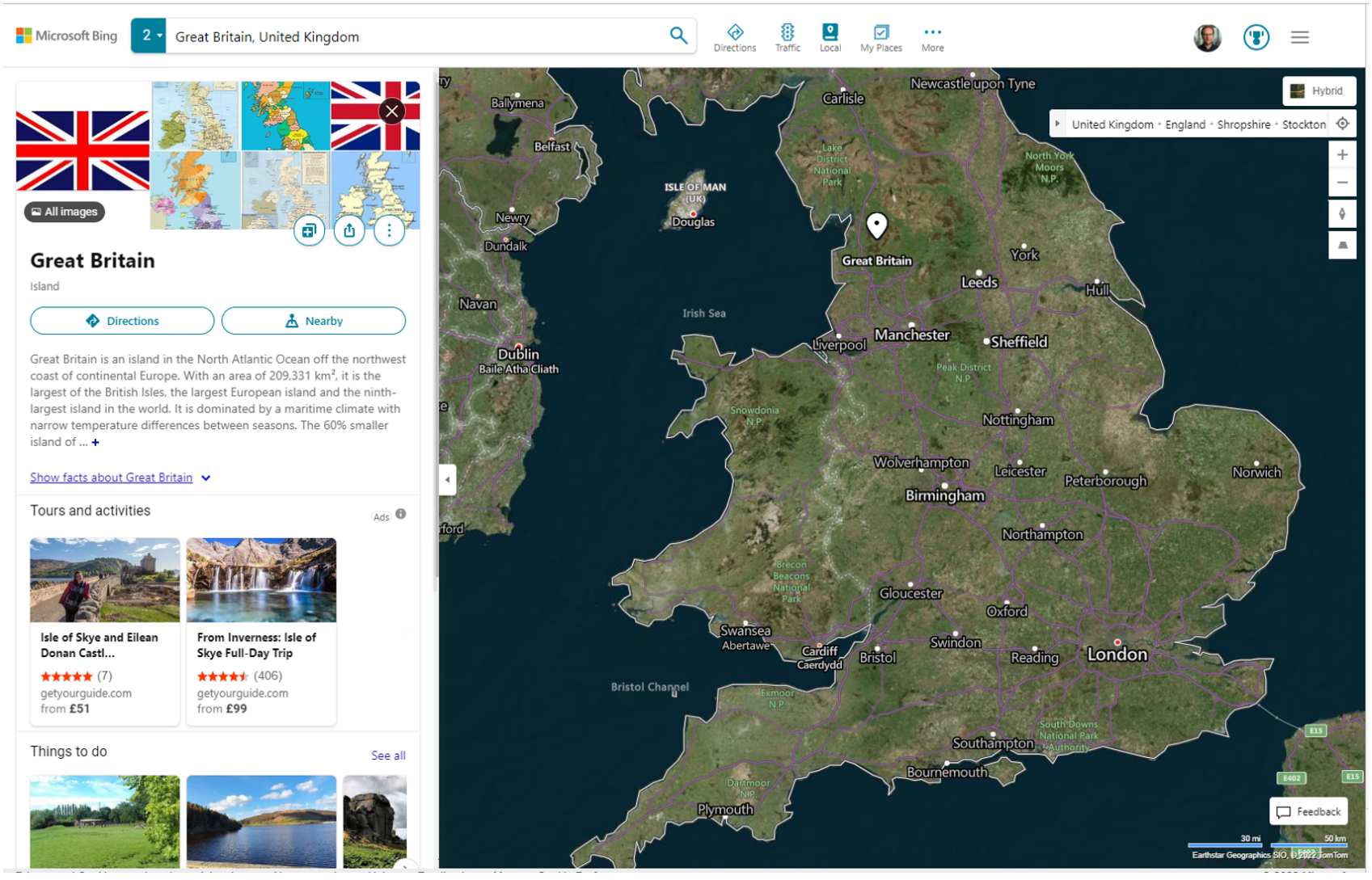
Fist Atlas of Britain: Anglia Regnum by Gerard Mercator (1595)

Source: https://www.historyhit.com/

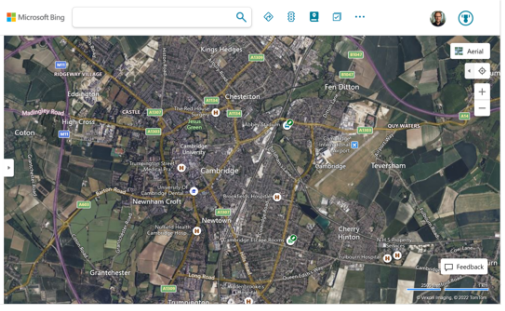


'Modern' Atlas (2022)

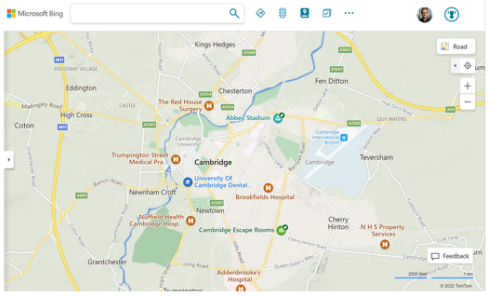
Precision travelling: Bing maps



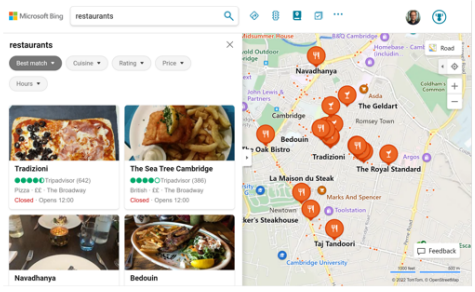
Bing maps: combining data sources of multiple granularity



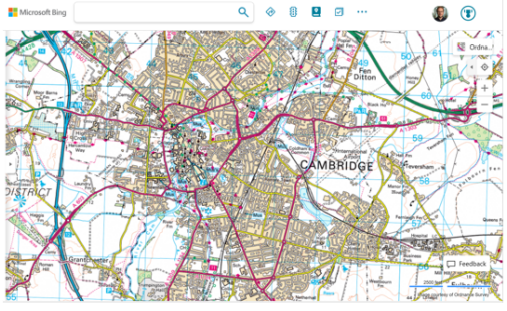
Natural maps



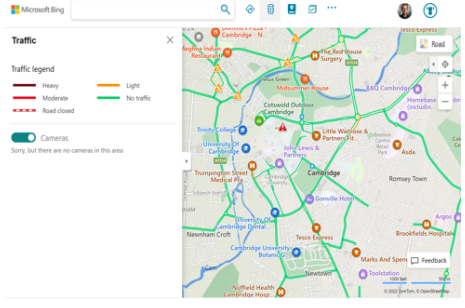
Information about shops



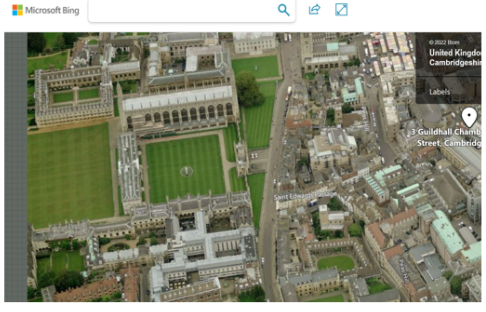
Product reviews



Political maps



Roads, traffic information



Details of streets and buildings

Low

Granularity level

High

Which is the Bing maps of precision medicine?

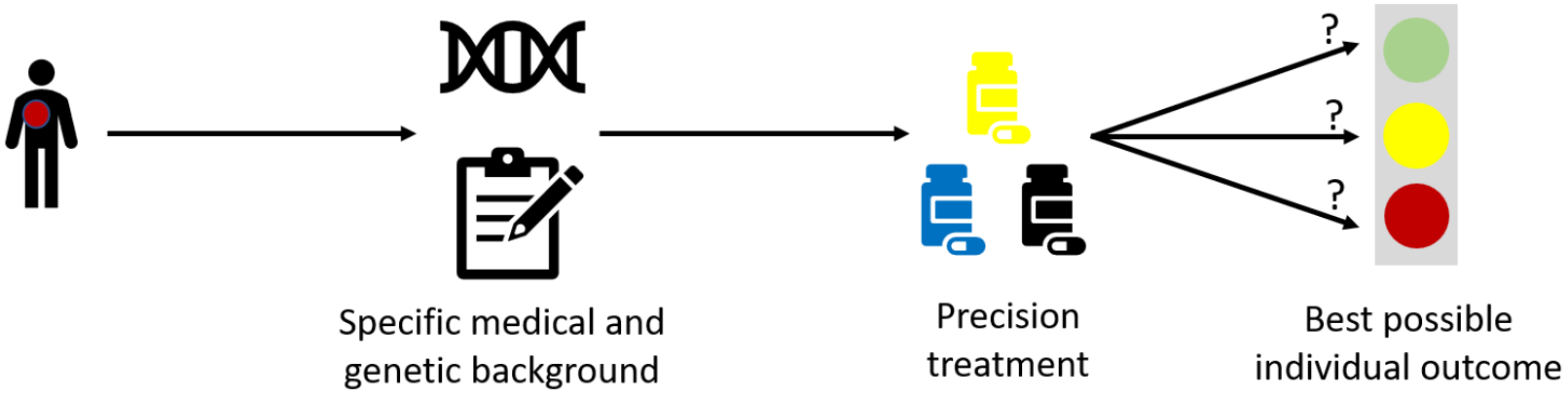
'Information commons' for precision medicine that uses all available data (modalities and types of data, including RCTs) to predict the effect of interventions.

Answers this tool could answer

- How do we expect a patient to respond to a specific treatment?
- How the probability of survival of patient will change if we change the treatment?
- Can an approved drug be repurposed?
- Are any of the eligibility criteria of an RCT redundant?
- Can we improve fairness in the design of new trials?

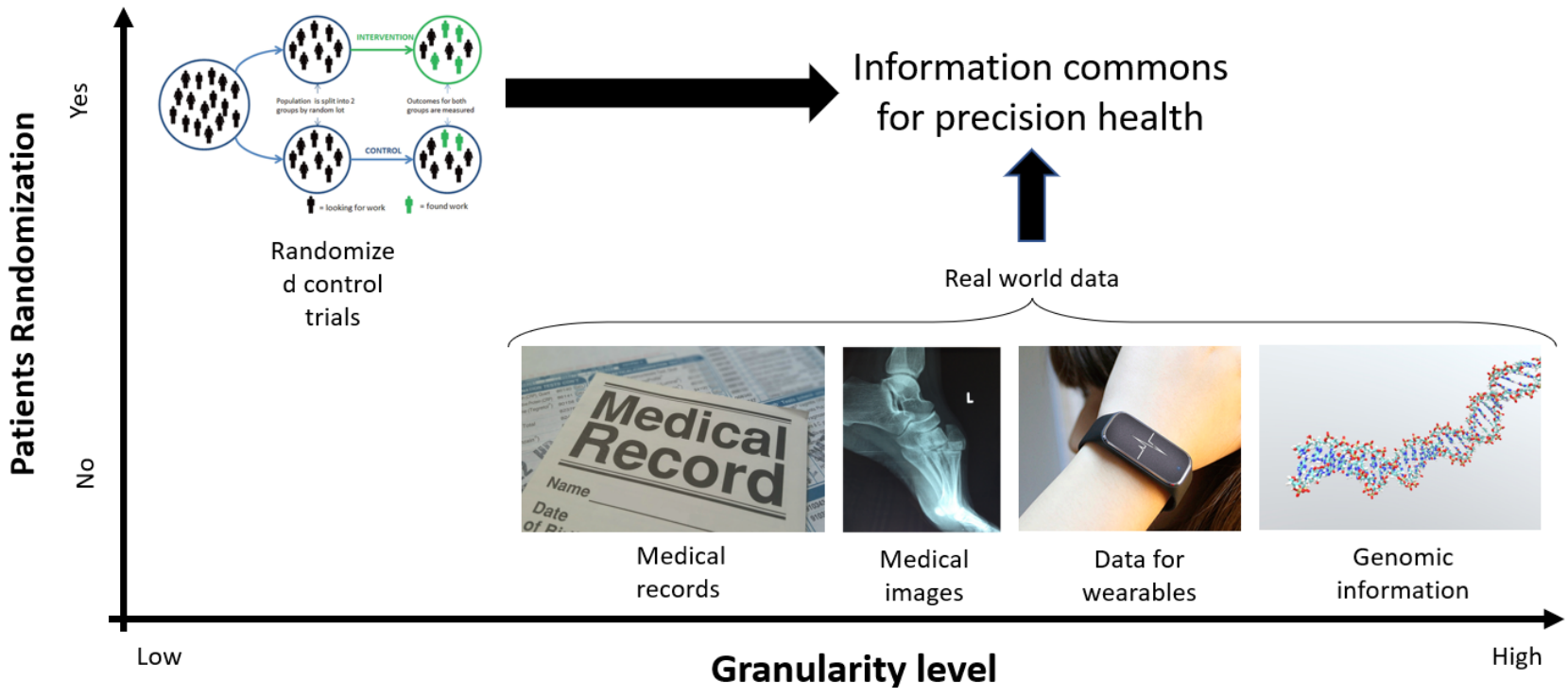
A causal view

Precision medicine questions are usually causal questions:



- Capture a medical/biological mechanism and go beyond correlations.
- Be robust/invariant to environmental conditions.
- Quantify uncertainty (so we know when to trust predictions).

Available sources of information in healthcare

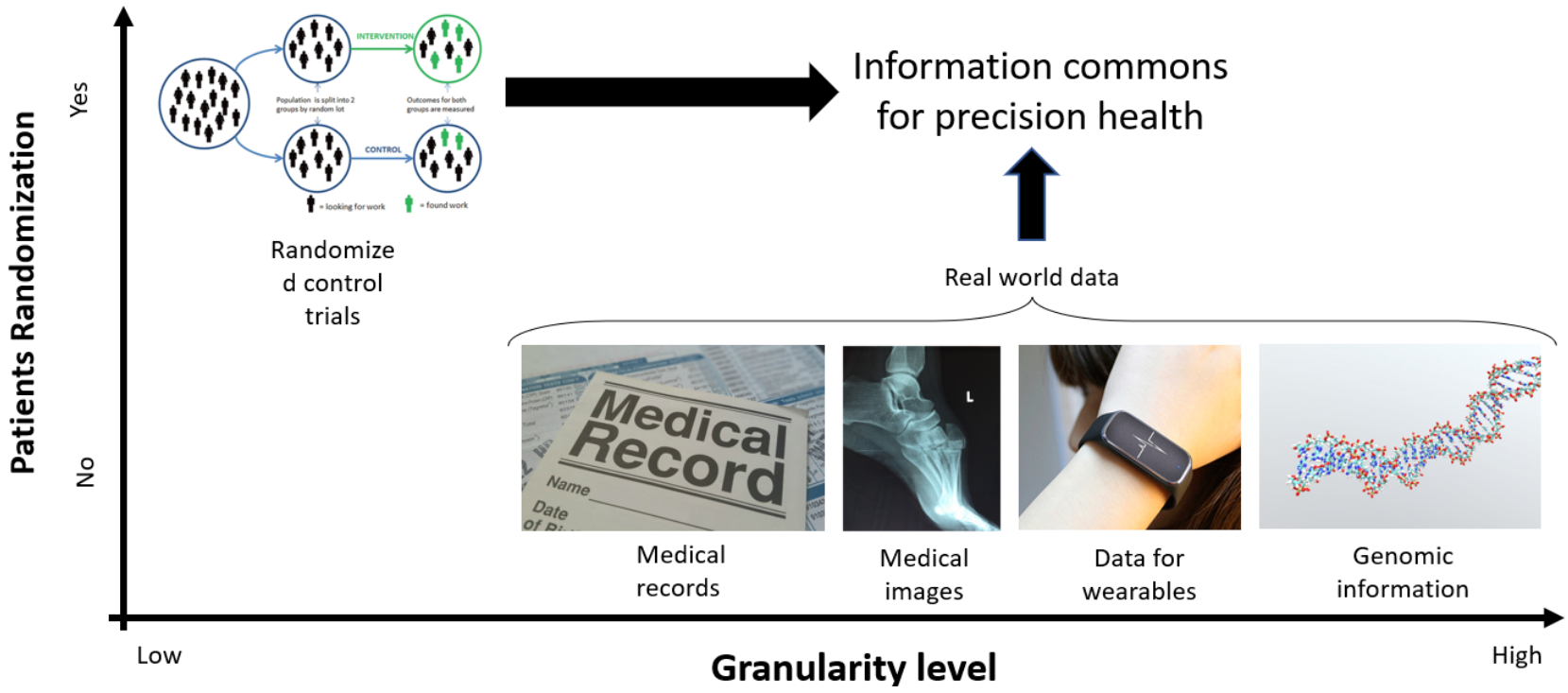


Available data to capture the effect of an intervention

RCT data		Real world data
✓	Randomization	✗
✗	Broad enrolment	✓
✗	Representativeness	✓
✓	Data quality	✗
SMALL	Sample size	LARGE
LARGE	Economic cost	SMALL
LARGE	Time cost	SMALL (LONG FOLLOW UP)
GOLD STANDARD	Regulatory validity	LOW, BUT INCREASING IN RELEVANCE

Available sources of information in healthcare

How to combine these data sources into one a consistent model able to answer precision medicine questions?



Casual data fusion

Well stablished theory to combine observational and interventional data.

PNAS PNAS PNAS

Causal inference and the data-fusion problem

Elias Bareinboim[a,b,1] and Judea Pearl[a]

[a]Department of Computer Science, University of California, Los Angeles, CA 90095; and [b]Department of Computer Science, Purdue University, West Lafayette, IN 47907

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved March 15, 2016 (received for review June 29, 2015)

We review concepts, principles, and tools that unify current approaches to causal analysis and attend to new challenges presented by big data. In particular, we address the problem of data fusion—piecing together multiple datasets collected under heterogeneous conditions (i.e., different populations, regimes, and sampling methods) to obtain valid answers to queries of interest. The availability of multiple heterogeneous datasets presents new opportunities to big data analysts, because the knowledge that can be acquired from combined data would not be possible from any individual source alone. However, the biases that emerge in heterogeneous environments require new analytical tools. Some of these biases, including confounding, sampling selection, and cross-population biases, have been addressed in isolation, largely in restricted parametric models. We here present a general, nonparametric framework for handling these biases and, ultimately, a theoretical solution to the problem of data fusion in causal inference tasks.

causal inference | counterfactuals | external validity | selection bias | transportability

Assume that the information available to us comes from an observational study, in which X, Y, Z, and W are measured, and samples are selected at random. We ask for conditions under which the query Q can be inferred from the information available, which takes the form $P(y,x,z,w)$, where Z and W are sets of observed covariates. This represents the standard task of policy evaluation, where controlling for confounding bias is the major issue (Fig. 1, task *1*).

Consider now Fig. 1, task *2*, in which the goal is again to estimate the effect of the intervention $do(X=x)$ but the data available to the investigator were collected in an experimental study in which variable Z, more accessible to manipulation than X, is randomized. [Instrumental variables (4) are special cases of this task.] The general question in this scenario is under what conditions can randomization of variable Z be used to infer how the population would react to interventions over X. Formally, our problem is to infer $P(Y=y|do(X=x))$ from $P(y,x,w|do(Z=z))$. A nonparametric solution to these two problems is presented in *Policy Evaluation and the Problem of Confounding*.

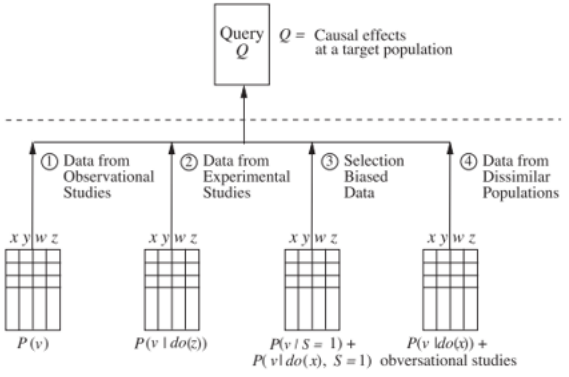


Fig. 1. Prototypical generalization tasks where the goal is, for example, to estimate a causal effect in a target population (*Top*). Let $V=\{X, Y, Z, W\}$. There are different designs (*Bottom*) showing that data come from nonidealized conditions, specifically: (*1*) from the same population under an observational regime, $P(v)$; (*2*) from the same population under an experimental regime when Z is randomized, $P(v|do(z))$; (*3*) from the same population under sampling selection bias, $P(v|S=1)$ or $P(v|do(x), S=1)$; and (*4*) from a different population that is submitted to an experimental regime when X is randomized, $P(v|do(x), S=s)$, and observational studies in the target population.

Where to start? Check the consistency of the datasets

In Bing, layers provide complementary information but need to be consistent

Q: Do RWD and RCTs provide consistent information?
A: RCT emulator, causal model built using RWD data that can replicate the results of a randomized trial.

Note: the goal is not to emulate a trial it is to externally validate the causal model.

Survival analysis in a nutshell

T: is a continuous variable with density function $f(t)$ that represents the **probability of an event** (death) occurring at time time t.

$S(t)$, **survival**: probability of being alive just before duration t.

$$S(t) = P(T \geq t) = \int_t^{\infty} f(x)dx$$

$\lambda(t)$, **Hazard**: instantaneous occurrence rate of occurrence of the event at time t, given that it didn't happen before

$$\lambda(t) = \lim_{t \to 0} \frac{P(t \geq T < t + dt | T \leq t)}{dt}, \quad \lambda(t) = \frac{f(t)}{S(t)}$$

Survival analysis

Proportional Hazards Cox model for a binary treatment $W = \{0, 1\}$ and covariates $X_1, \ldots, X_p$

$$\lambda(t|W, X) = \lambda_0(t) \exp\left(\beta_w W + \sum_{j=1}^{p} \beta_j X_j\right)$$

HR, Hazard ratio: relative chances of death with and without the treatment.

$$HR = \frac{\lambda(t|W = 1, X)}{\lambda(t|W = 0, X)} = \exp(\beta_w)$$

$$log(HR) = \beta_w$$

Note: the lower the HR the more effective is the treatment with respect to the control group.

Causal survival analysis

Compute the marginal HR of the treatment using a PH-Cox model

$$\lambda(t|W) = \lambda_0(t) \exp(\beta_w W)$$

where the goups are balanced with inverse propensity weighting

$$w_i = W_i + (1 - W_i) \left[\frac{e(X_i)}{1 - e(X_i)} \right]$$

$$e(X) = P(W|X)$$

Emulating an RCT with electronic medical records

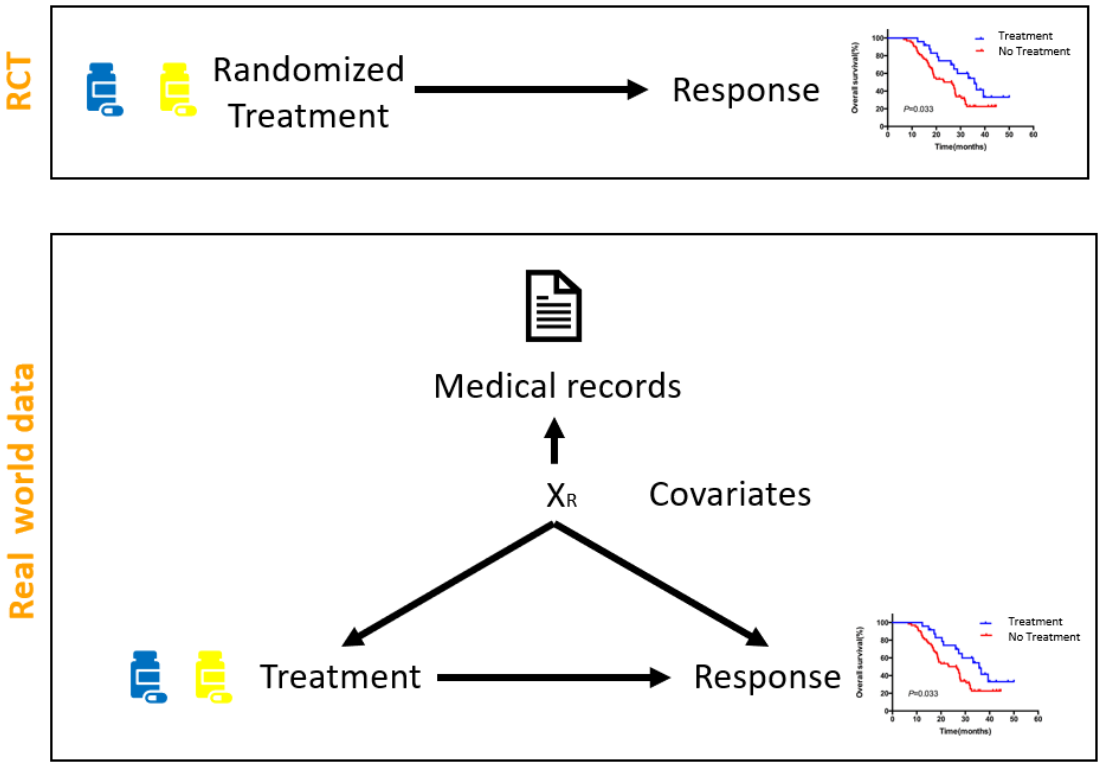
Dataset 1: Cases and controls are randomized in the trial

RCT-HR

?

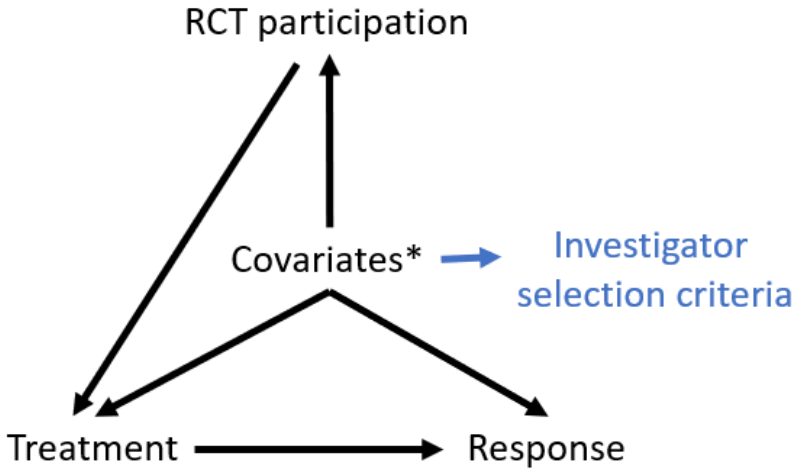
RWE-HR

Dataset 2: cases and control are observed toguether with the response and the electronic medical records of patients.

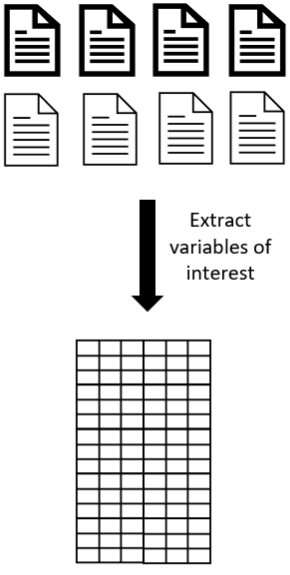


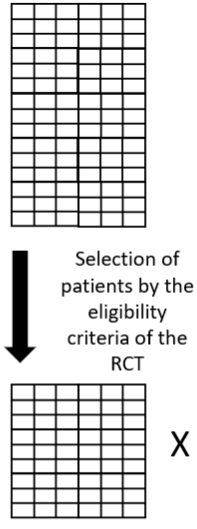
Assumptions for RCT emulation

1. Complete adherence to assigned treatment.
2. No loss-to-follow-up.
3. No measurement error.
4. *Covariates capture all relevant confounders and eligibility criteria.
5. Trial participation doesn't affect outcomes.
6. Positive prob. of participation in the trial and the treatment.
7. RCT participation = selection criteria
8. Same background populations.

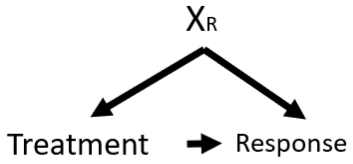


Baseline approach to emulate an RCT





- **Causal assumptions:**

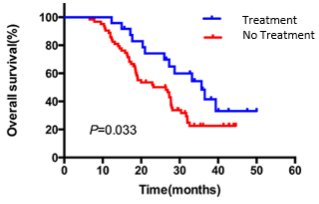


- **Cox model:**

$$h(t|X) = h_0(t)\exp\left(b_w W + \sum_{j=1}^{p} b_j X_j\right)$$

- **Inverse Propensity weighting**

$$w_i = W_i + (1 - W_i)\left[\frac{e(X_i)}{1 - e(X_i)}\right]$$



HR from medical records:
0.67 (0.60, 0.74)

Reported HR from RCT
0.68

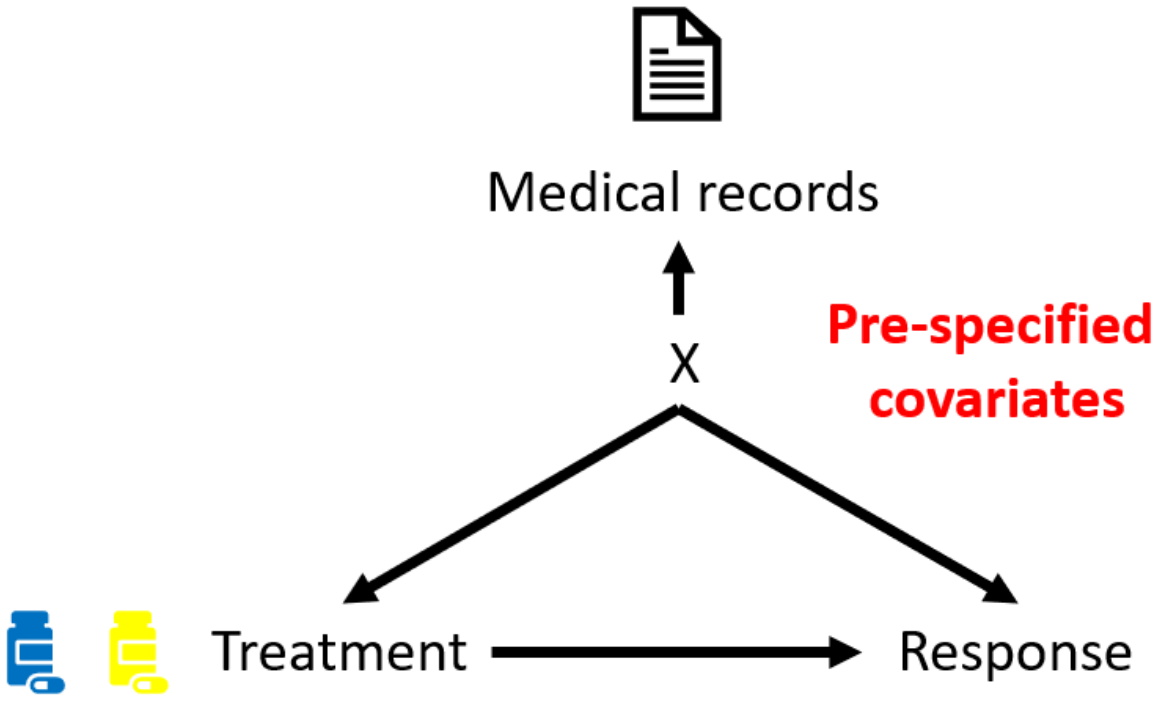
Step 1: structuring available data

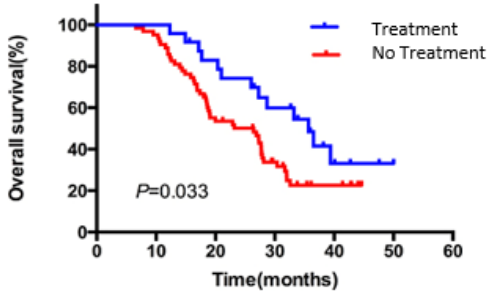
Step 2: matching patients with eligibility criteria

Step 3: building causal model of the outcome (HR)

Step 4: Querying and benchmarking the model ☺

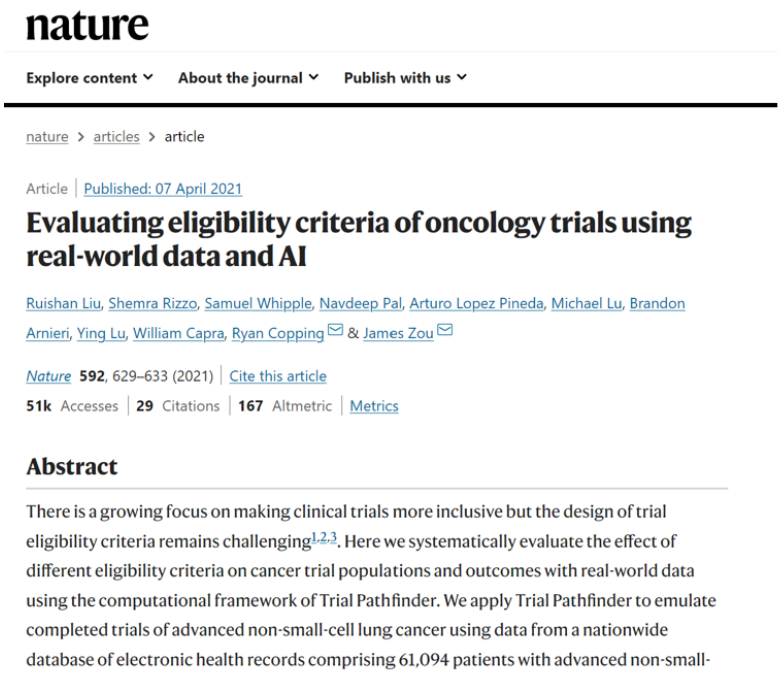
Is this possible?





Is this possible?

Yes, this is an active area of research.



nature

Explore content ▾ About the journal ▾ Publish with us ▾

nature › articles › article

Article | Published: 07 April 2021

Evaluating eligibility criteria of oncology trials using real-world data and AI

Ruishan Liu, Shemra Rizzo, Samuel Whipple, Navdeep Pal, Arturo Lopez Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William Capra, Ryan Copping ✉ & James Zou ✉

Nature **592**, 629–633 (2021) | Cite this article

51k Accesses | **29** Citations | **167** Altmetric | Metrics

Abstract

There is a growing focus on making clinical trials more inclusive but the design of trial eligibility criteria remains challenging[1,2,3]. Here we systematically evaluate the effect of different eligibility criteria on cancer trial populations and outcomes with real-world data using the computational framework of Trial Pathfinder. We apply Trial Pathfinder to emulate completed trials of advanced non-small-cell lung cancer using data from a nationwide database of electronic health records comprising 61,094 patients with advanced non-small-

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-28546-8 OPEN

Check for updates

Uncovering interpretable potential confounders in electronic medical records

Jiaming Zeng[1✉], Michael F. Gensheimer[2], Daniel L. Rubin[3], Susan Athey[4] & Ross D. Shachter[1]

Randomized clinical trials (RCT) are the gold standard for informing treatment decisions. Observational studies are often plagued by selection bias, and expert-selected covariates may insufficiently adjust for confounding. We explore how unstructured clinical text can be used to reduce selection bias and improve medical practice. We develop a framework based on natural language processing to uncover interpretable potential confounders from text. We validate our method by comparing the estimated hazard ratio (HR) with and without the confounders against established RCTs. We apply our method to four cohorts built from

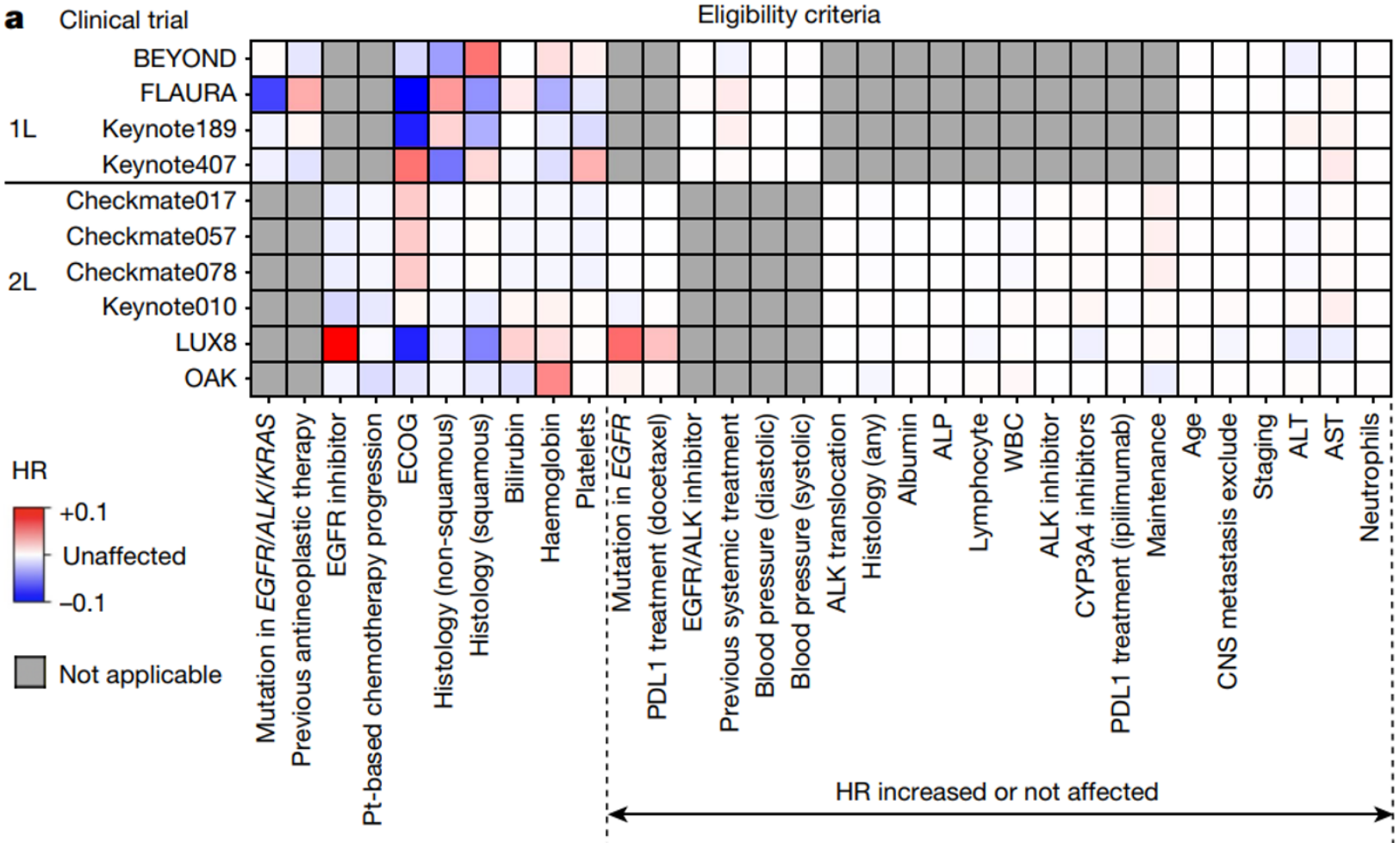
Results of Liu et al. 2021 using the Flatiron database

Trial Short Code	Published HR (95% CI)
FLAURA	0.63 (0.45, 0.88)
LUX8	0.81 (0.69, 0.95)
Checkmate017	0.59 (0.44, 0.79)
Checkmate057	0.73 (0.59, 0.89)
Checkmate078	0.68 (0.52, 0.90)
Keynote010	0.71 (0.58, 0.88) 0.61 (0.49, 0.75)
Keynote189	0.49 (0.38, 0.64)
Keynote407	0.64 (0.49, 0.85)
BEYOND	0.68 (0.50, 0.93)
OAK	0.73 (0.62, 0.87)

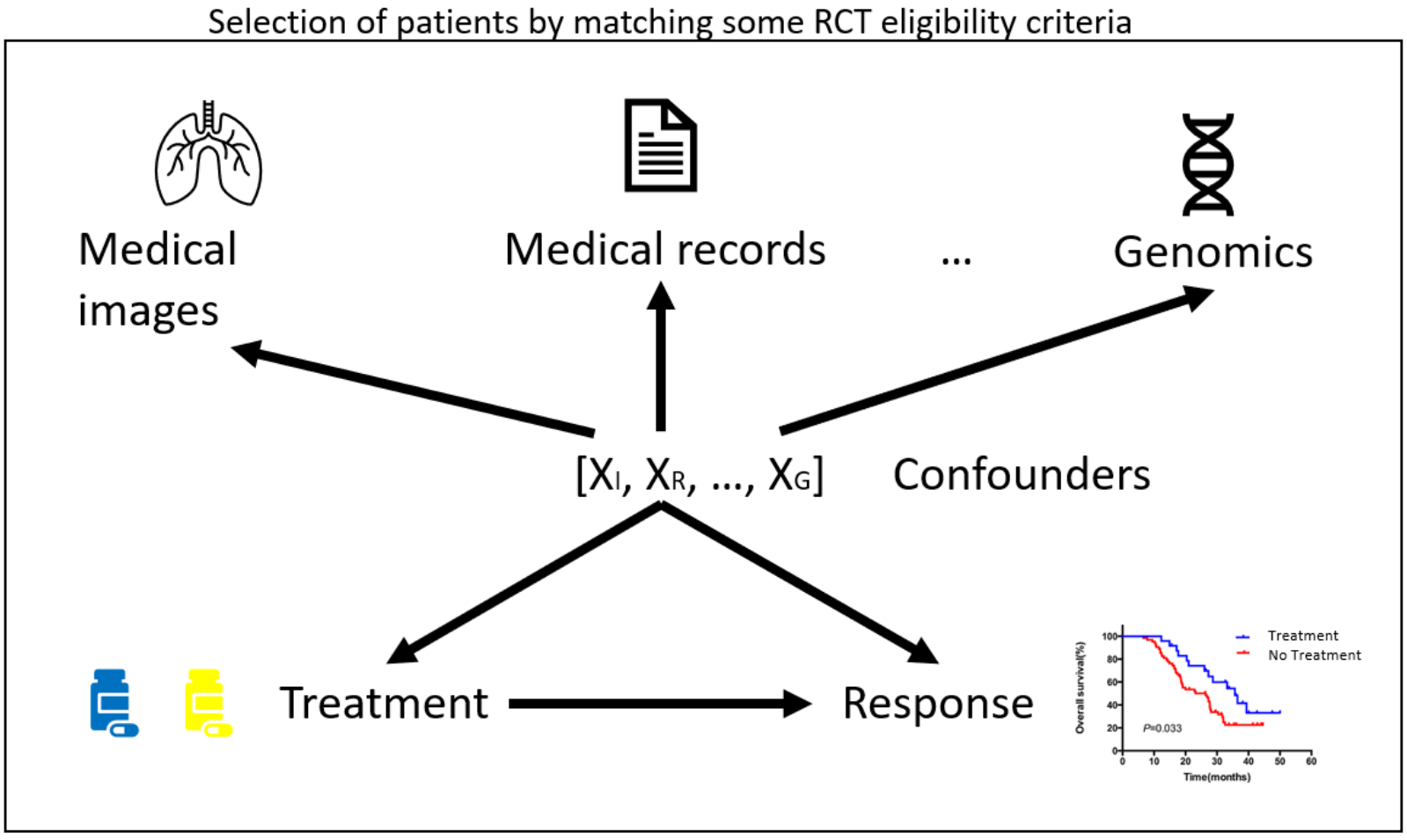
Trial name	Original trial criteria			Fully relaxed criteria		Data-driven criteria		
	No. of criteria	No. of patients	HR	No. of patients	HR	No. of criteria	No. of patients	HR
FLAURA	10	2,277	0.81	3,819	0.82	4	2,546	0.75
LUX8	11	129	0.65	1,350	0.81	5	141	0.58
Checkmate017	17	523	0.67	4,900	0.71	7	4,085	0.71
Checkmate057	19	792	0.75	4,900	0.71	9	2,594	0.66
Checkmate078	18	1,509	0.74	4,900	0.71	9	3,348	0.68
Keynote010	13	806	0.56	1,950	0.51	1	1,948	0.51
Keynote189	15	4,066	0.88	8,818	0.94	7	4,595	0.85
Keynote407	13	2,031	1.13	10,437	1.07	4	9,173	1.04
BEYOND	12	2,902	1.09	9,310	1.14	4	3,043	1.08
OAK	19	493	0.88	1,288	0.87	6	620	0.80
Average	15	1,553	0.82	5,167	0.83	6	3,209	0.77

The number of inclusion/exclusion criteria, the number of eligible patients and the hazard ratio of the overall survival of emulated aNSCLC trials with eligibility criteria under three scenarios: the original criteria used in the trial, fully relaxed criteria and data-driven criteria. The fully relaxed criteria correspond to evaluating the hazard ratio of the overall survival of all of the patients in the Flatiron database who took the treatments in the relevant line of therapy. The data-driven criteria were selected by Shapley values. HR, hazard ratio.

Results of Liu et al. 2021 using the Flatiron database



Looking into the future



Some recent works useful in this direction

Variational auto-encoders to combine multiple sources of information.
The goal is learning a system of structural equations.

DEEP MULTI-MODAL STRUCTURAL EQUATIONS FOR CAUSAL EFFECT ESTIMATION WITH UNSTRUCTURED PROXIES

Shachi Deshpande, Kaiwen Wang
Dhruv Sreenivas, Zheng Li, Volodymyr Kuleshov
shachi@cs.cornell.edu, wangkaiwen998@gmail.com,
ds844@cornell.edu, zl634@cornell.edu, kuleshov@cornell.edu
Department of Computer Science, Cornell Tech
New York, NY 10044

ABSTRACT

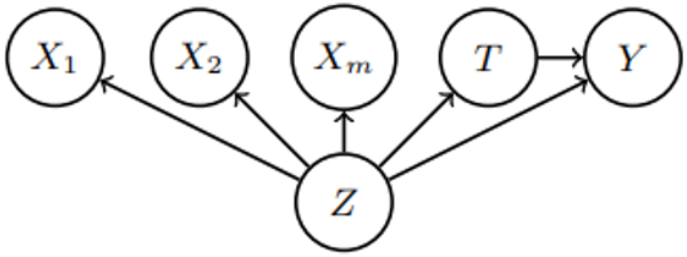
Estimating the effect of an intervention while accounting for confounding variables is a key task in causal inference. Oftentimes, the confounders are unobserved, but we have access to large amounts of unstructured data (images, text) that contain valuable proxy signal about the missing confounders. This paper demonstrates that leveraging unstructured data that is often left unused by existing algorithms improves the accuracy of causal effect estimation. Specifically, we introduce deep multi-modal structural equations, a generative model in which confounders are latent variables and unstructured data are proxy variables. This model supports multiple multi-modal proxies (images, text) as well as missing data. We empirically demonstrate on tasks in genomics and healthcare that our approach corrects for confounding using unstructured inputs, potentially enabling the use of large amounts of data that were previously not used in causal inference.

1 Introduction

An important goal of causal inference is to understand from observational data the causal effect of performing an intervention—e.g., the effect of a behavioral choice on an individual's health (Pearl, 2009). As an initial motivating example for this work, consider the problem of determining the effect of smoking on an individual's risk of heart disease.

$$Z \sim \mathcal{N}(0_p, I_p) \qquad X_j \sim \mathcal{P}_{X_j}(\theta_{X_j}(Z))) \;\; \forall j$$

$$T \sim \text{Ber}(\pi_T(Z)) \qquad Y \sim \mathcal{P}_Y(\theta_Y(Z,T))),$$



Some recent works useful in this direction

Generative models for counterfactual generation for survival analysis.

Enabling Counterfactual Survival Analysis with Balanced Representations

Paidamoyo Chapfuwa
Duke University
USA
paidamoyo.chapfuwa@duke.edu

Serge Assaad
Duke University
USA
serge.assaad@duke.edu

Shuxi Zeng
Duke University
USA
zengshx777@gmail.com

Michael J. Pencina
Duke University
USA
michal.pencina@duke.edu

Lawrence Carin
Duke University
USA
lcarin@duke.edu

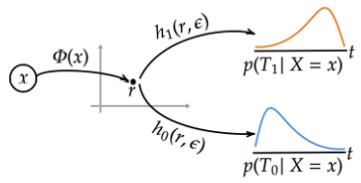
Ricardo Henao
Duke University
USA
ricardo.henao@duke.edu

ABSTRACT

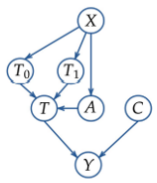
Balanced representation learning methods have been applied successfully to counterfactual inference from observational data. However, approaches that account for survival outcomes are relatively limited. Survival data are frequently encountered across diverse medical applications, *i.e.*, drug development, risk profiling, and clinical trials, and such data are also relevant in fields like manufacturing (*e.g.*, for equipment monitoring). When the outcome of interest is a time-to-event, special precautions for handling censored events need to be taken, as ignoring censored outcomes may lead to biased estimates. We propose a theoretically grounded unified framework for counterfactual inference applicable to survival out-

1 INTRODUCTION

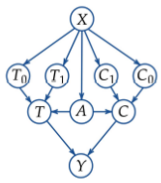
Survival analysis or time-to-event studies focus on modeling the time of a future event, such as death or failure, and investigate its relationship with covariates or predictors of interest. Specifically, we may be interested in the *causal effect* of a given intervention or treatment on survival time. A typical question may be: will a given therapy increase the chances of survival of an individual or population? Such causal inquiries on survival outcomes are common in the fields of epidemiology and medicine [22, 47, 65]. As an important current example, the COVID-19 pandemic is creating a demand for methodological development to address such questions, specifically, when evaluating the effectiveness of a potential vaccine



(a) Counterfactual inference



(b) Non-informative



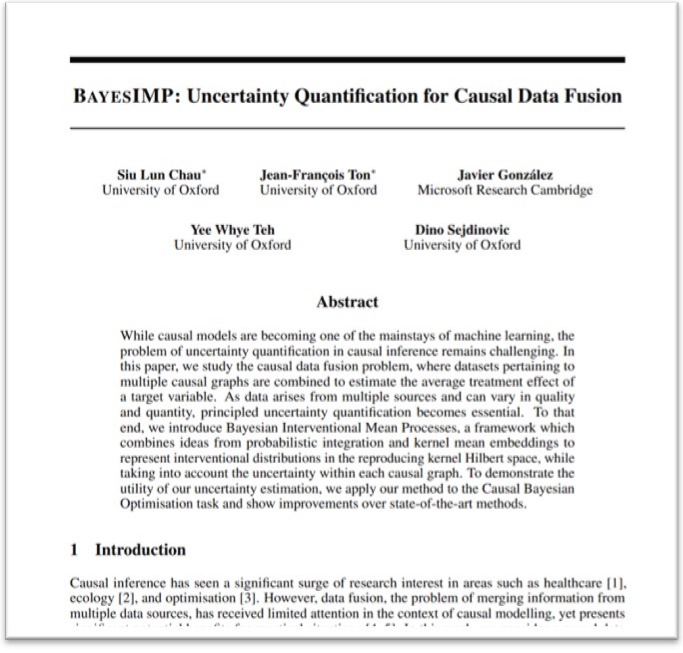
(c) Informative

Table 1: Performance comparisons on ACTG-SYNTHETIC data, with 95% HR(t) confidence interval. The ground truth, test set, hazard ratio is $\text{HR(t)} = 0.52_{(0.39,0.71)}$.

Method	Causal metrics			Factual metrics		
	ϵ_{PEHE}	ϵ_{ATE}	HR(t)	C-Index (A=0, A=1)	Mean COV	C-Slope (A=0, A=1)
CoxPH-Uniform	NA	NA	$0.97_{(0.86,1.09)}$	NA	NA	NA
CoxPH-IPW	NA	NA	$0.48_{(0.03,7.21)}$	NA	NA	NA
CoxPH-OW	NA	NA	$0.60_{(0.53,0.68)}$	NA	NA	NA
Surv-BART	352.07	77.89	0.0(0.0, 0.0)	(0.706, 0.686)	0.001	(0.398, ∞)
AFT-Weibull	367.92	133.93	$0.47_{(0.47,0.47)}$	(0.21, 0.267)	6.209	(0.707, 0.729)
AFT-log-Normal	377.76	157.64	$0.47_{(0.47,0.47)}$	(0.675, 0.556)	6.971	(0.707, 0.729)
SR	369.47	88.55	$0.38_{(0.33,0.65)}$	(0.791, 0.744)	0	(0.985, 1.027)
CSA (proposed)	358.72	**0.8**	$0.45_{(0.39,0.65)}$	(0.787, 0.767)	0.131	(0.985, 1.026)
CSA-INFO (proposed)	**344.3**	31.19	$\mathbf{0.53}_{(0.41,0.67)}$	(0.78, 0.764)	0.13	(0.999, 1.029)

Some recent works useful in this direction

Uncertainty quantification for causal data fusion.



BAYESIMP: Uncertainty Quantification for Causal Data Fusion

Siu Lun Chau* University of Oxford

Jean-François Ton* University of Oxford

Javier González Microsoft Research Cambridge

Yee Whye Teh University of Oxford

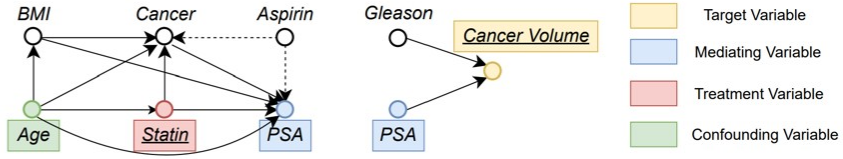
Dino Sejdinovic University of Oxford

Abstract

While causal models are becoming one of the mainstays of machine learning, the problem of uncertainty quantification in causal inference remains challenging. In this paper, we study the causal data fusion problem, where datasets pertaining to multiple causal graphs are combined to estimate the average treatment effect of a target variable. As data arises from multiple sources and can vary in quality and quantity, principled uncertainty quantification becomes essential. To that end, we introduce Bayesian Interventional Mean Processes, a framework which combines ideas from probabilistic integration and kernel mean embeddings to represent interventional distributions in the reproducing kernel Hilbert space, while taking into account the uncertainty within each causal graph. To demonstrate the utility of our uncertainty estimation, we apply our method to the Causal Bayesian Optimisation task and show improvements over state-of-the-art methods.

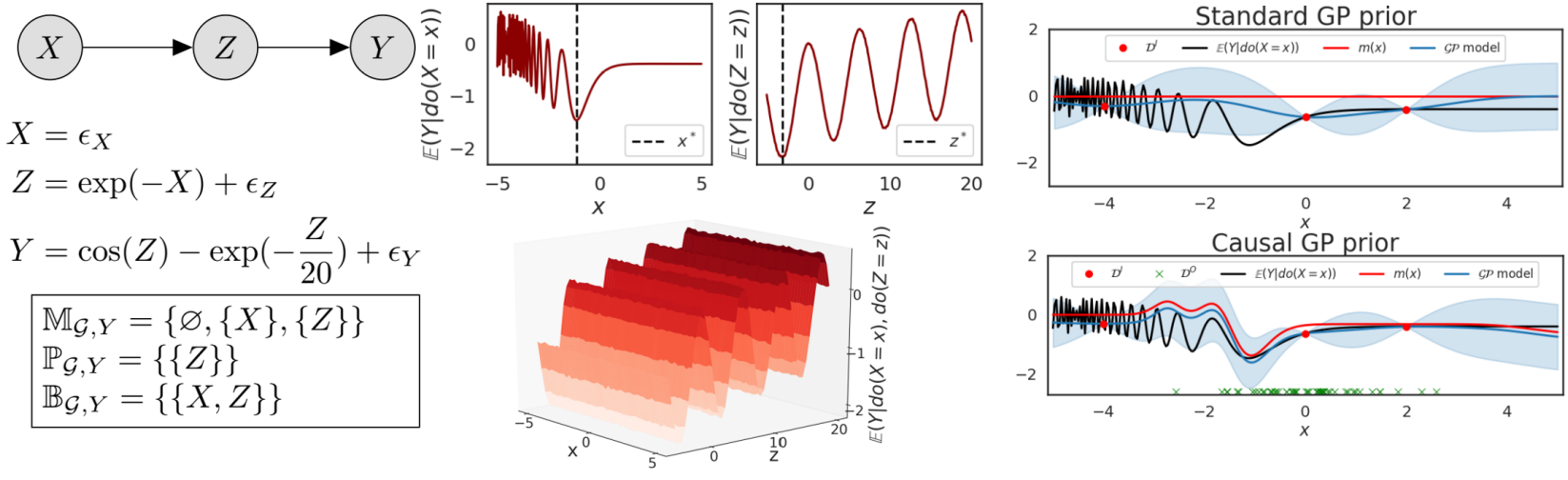
1 Introduction

Causal inference has seen a significant surge of research interest in areas such as healthcare [1], ecology [2], and optimisation [3]. However, data fusion, the problem of merging information from multiple data sources, has received limited attention in the context of causal modelling, yet presents



Some recent works useful in this direction

Causal Gaussian processes that can combine observational and experimental data.



Causal Bayesian Optimization. V Aglietti, X Lu, A Paleyes, J Gonzalez. Artificial intelligence and Statistics, AISTATS, 2020

Dynamic Causal Bayesian Optimization. V Aglietti, N Dhir, J González, T Damoulas. Neural Information Processing Systems, NeurIPS, 2021

Better modelling of the patient journey

Uncertainty quantification in patient level counterfactuals.

Predicting the impact of treatments over time with uncertainty aware neural differential equations.

Edward De Brouwer[†]
ESAT-STADIUS
KU Leuven

Javier González Hernández
Microsoft Research
Cambridge, UK

Stephanie Hyland
Microsoft Research
Cambridge, UK

Abstract

Predicting the impact of treatments from observational data only still represents a major challenge despite recent significant advances in time series modeling. Treatment assignments are usually correlated with the predictors of the response, resulting in a lack of data support for counterfactual predictions and therefore in poor quality estimates. Developments in causal inference have lead to methods addressing this confounding by requiring a minimum level of overlap. However, overlap is difficult to assess and usually not satisfied in practice. In this work, we pro-

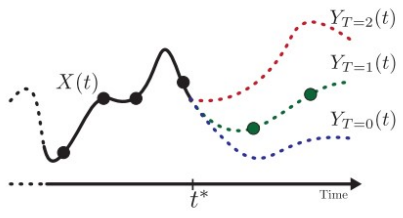


Figure 1: Based on available trajectory information $X(t)$, we aim at predicting in continuous time the potential outcomes of applying treatment regime T at time t^* (dotted lines). As with the fundamental problem of causal inference, a single outcome is available for each instance in the dataset (solid green dots).

Wrap up: Machine learning and precision medicine

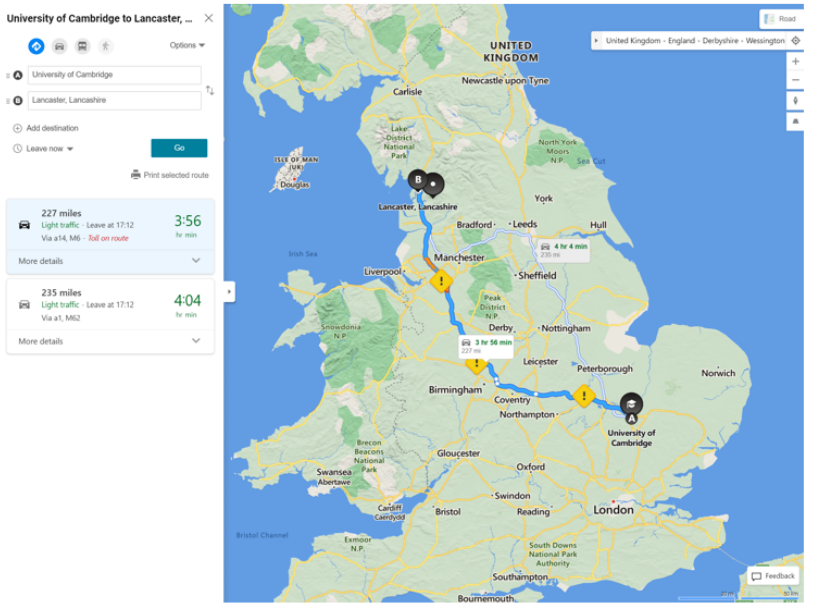
- Exciting area of research with a potentially large impact in the medical domain.
- Causal methods have a key role to play in the field.
- RCT emulation is just the first step but there are multiple opportunities.
- Multimodal learning in the context of causal methods open several opportunities for impact and research.

Better modelling of the patient journey

Better maps and better GPS for health?

Source: mapmania.com

1962



2022

Questions?