Causality and learning dependences

Artificial Intelligence for Digital Health (AID)

M.Sc. in Digital Health – University of Pisa

Davide Bacciu (davide.bacciu@unipi.it)

This lecture (from yesterday lecture)

- ~~Design~~ Learn the structure of the network by identifying variable (nodes) associations (edges)
- Fit the parameters of the Bayesian Network by maximum likelihood
- Make predictions (e.g. diagnose a disease)
- Sample observations (e.g. complete missing variables)
- Reason on ~~associations~~ causal relationships

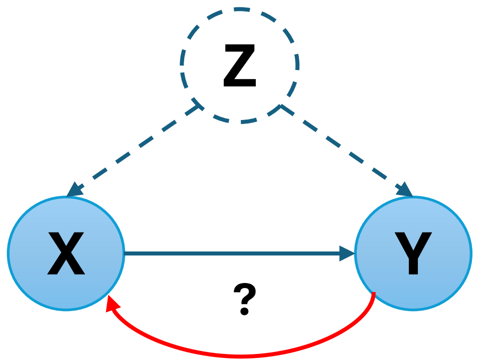
Lecture Outline

- Correlation, dependence and causation
- Causality
  - Interventions
  - Measuring causality and average treatment effects
  - Randomized control trials
- Discovering dependence in data
  - Structure learning
  - PC algorithm
  - Search-and-score
- Applications in healthcare

Correlation, Dependence and Causation

- A random variable is "**causing**" another random variable if a "**manipulation**" on the former alters the distribution of the latter.
- Correlation alone does not imply **direct causation**.
- In fact, completely **different causal structures** can entail the **same** set of conditional **independences** and dependences.

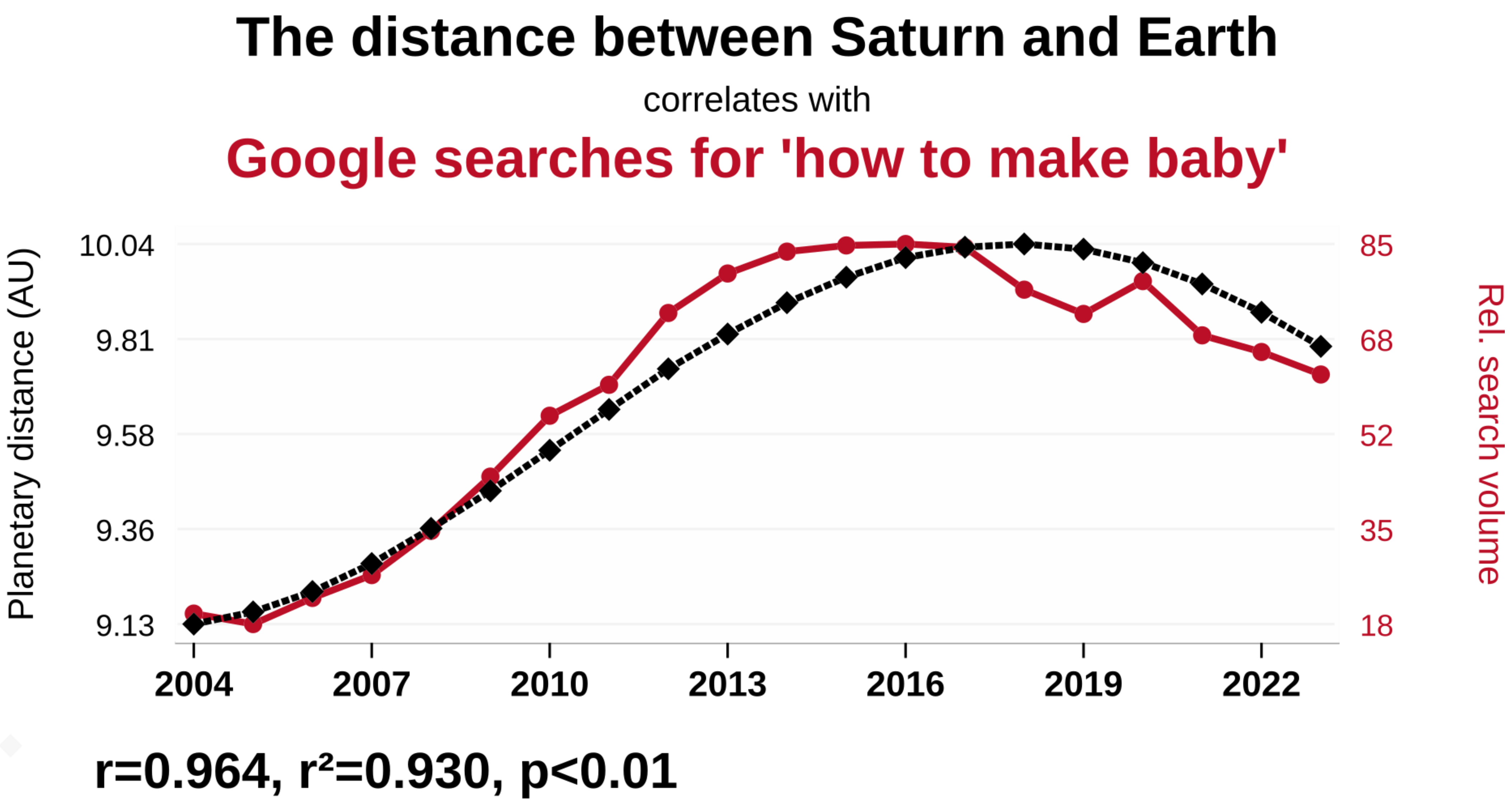
Reichenbach's Principle



Reichenbach's Common Cause Principle

Let X and Y be two variables such that X and Y are **statistically dependent**, then it holds:

i. X is indirectly causing Y, or
ii. Y is indirectly causing X, or
iii. There is a possibly unobserved common cause Z that indirectly causes both X and Y.



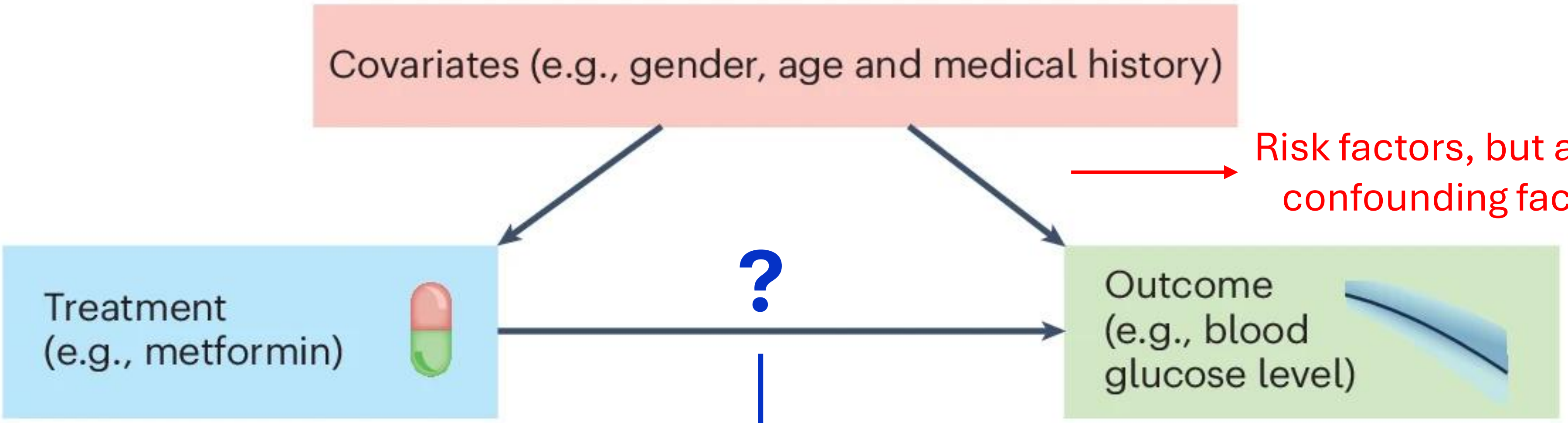
tylervigen.com/spurious/correlation/13099

Reichenbach's Principle

- The **principle** assumes that we can **perfectly** identify statistical dependence from data.
- In general, we need particular care:
  - Selection Bias
  - Small Size Datasets (Sampling Bias)
  - Common Trends
  - Data Manipulations
  - Measurement Errors

source: S Feuerriegel et al, Nature Medicine, 2024

The key (causal) question in ML for health



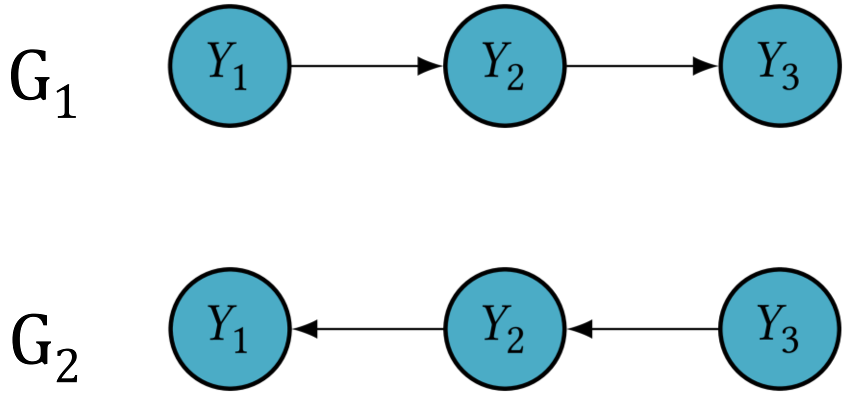
Risk factors, but also all confounding factors!

Is the treatment substantially influencing (causing) the outcome?

Causality

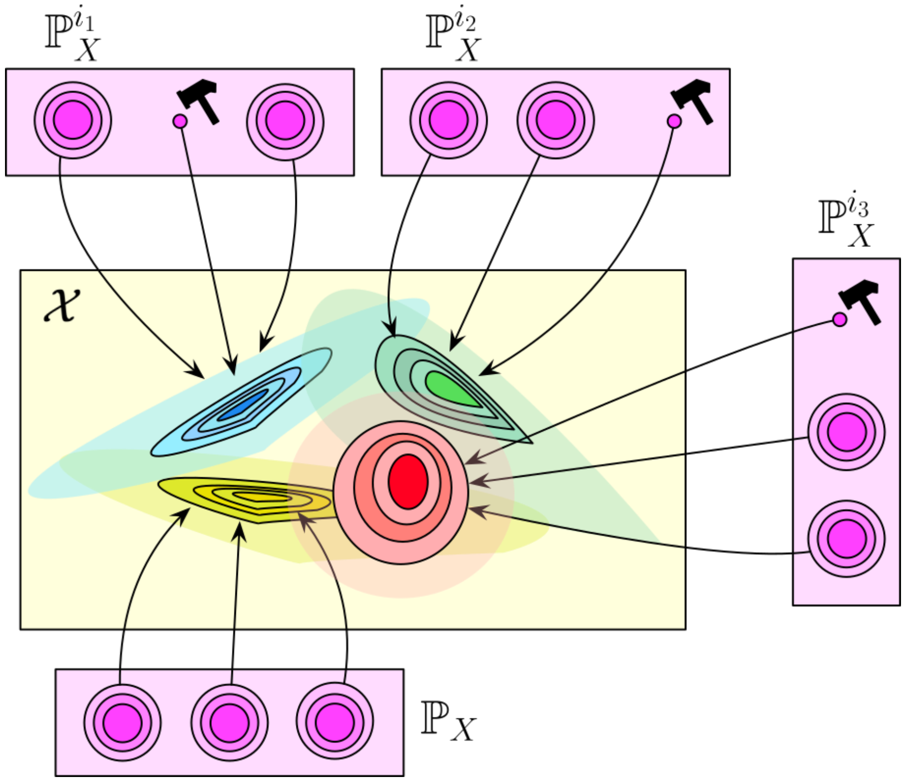
Causal Bayesian Networks

- A **Causal Bayesian Network** is a Bayesian Network where each edge $\mathrm{Y}_1 \rightarrow \mathrm{Y}_2$ represents that Y_1 **directly causes** a variable Y_2.
- The two models G_1 and G_2 denote equivalent Bayesian Networks but distinct Causal Bayesian Networks.



Intervening on Causal Models

- **Interventions** are the main operations on causal models.
- While different probabilistic models can express the same conditional distributions, different causal models entail different **interventional distributions**.



Causal Machine Learning (ML)

Estimate treatment effects to generate clinical evidence

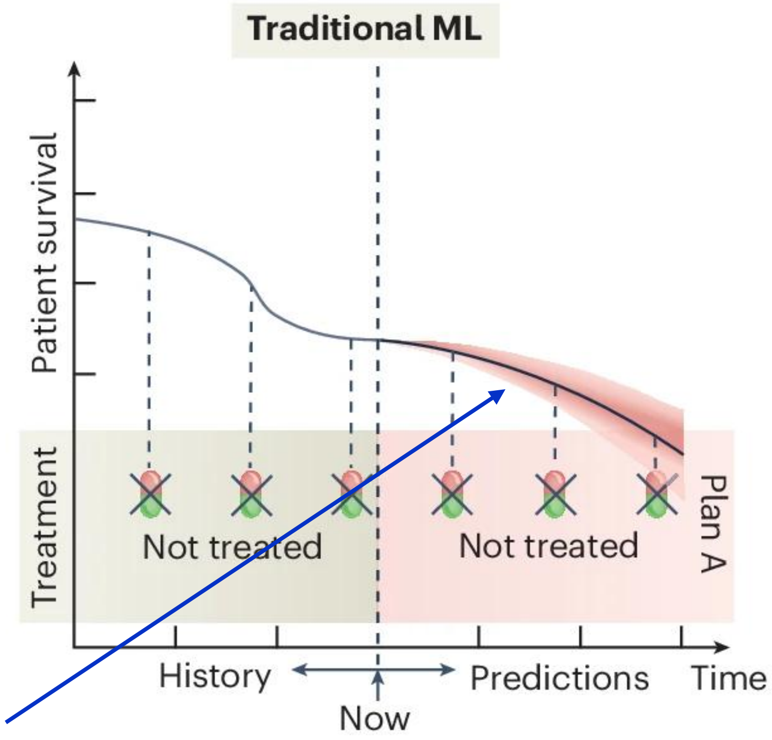
Individualized treatment effects and personalized predictions of potential patient outcomes under different treatment scenarios

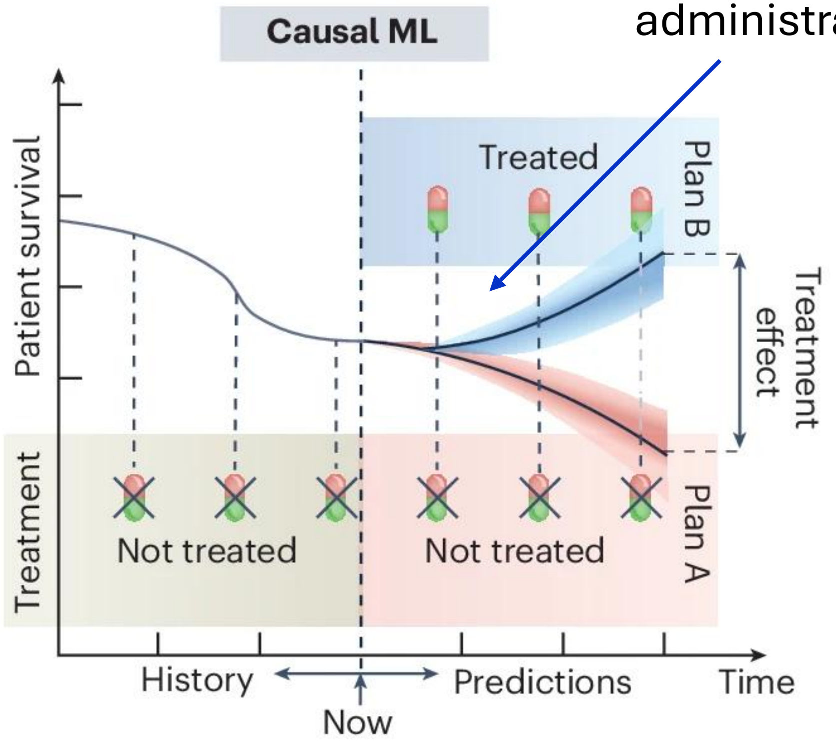
Understanding when treatments are effective or harmful

Patient care can be personalized to individual patient profiles.

Traditional Vs Causal ML

Ability to reason on what-if scenario on treatment administration





Goal is predicting outcomes with a certain confidence

source: S Feuerriegel et al, Nature Medicine, 2024

Goal is quantifying changes in outcomes due to treatment

The Fundamental Problem of Causal Inference

not all potential outcomes can be observed

Data

Traditional ML

Patient	Covariates	Treatment	Patient outcome
1	Age, sex, etc.	0	–1.0
2	↓	1	2.3
3		1	0.3

Causal ML

Patient	Covariates	Treatment	Patient outcome If not treated	Patient outcome If treated
1	Age, sex, etc.	0	–1.0	☐
2	↓	1	☐	2.3
3		1	☐	0.3

Task

Traditional ML

Patient	Covariates	Treatment	Patient outcome
1	Age, sex, etc.	1	?
2	↓	0	?

Causal ML

Patient	Covariates	Potential outcomes If not treated	Potential outcomes If treated	Treatment effect If treated → If not treated
1	Age, sex, etc.	?	?	?
2	↓	?	?	?

☐ Missing observations ? Prediction targets

Focus on estimating treatment effect rather than potential outcomes

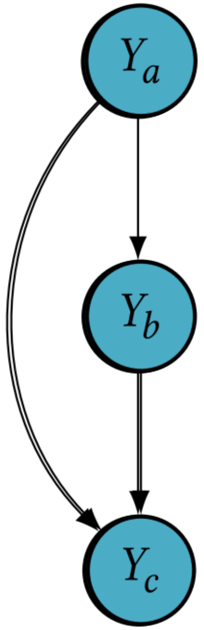
source: S Feuerriegel et al, Nature Medicine, 2024

Ideal Interventions

- Given a variable Y and a value k, we denote an **ideal intervention**, also known as *hard* or *perfect*, as

$$\mathrm{do}(\mathrm{Y} \coloneqq \mathrm{k})$$

- The intervention replaces the variable of the model with the constant value.
- In general,
$\mathrm{P}(\mathrm{Y}_2 \mid \mathrm{Y}_1 = \mathrm{k}) \neq \mathrm{P}(\mathrm{Y}_2 \mid \mathrm{do}(\mathrm{Y}_1 \coloneqq \mathrm{k}))$



$P(Y_a, Y_b, Y_c \mid do(Y_b := k))$

Truncated Factorization

- Let V be a set of variables and k a set of values.
- Then, the intervention $\mathrm{do}(\mathrm{V} \coloneqq \mathrm{k})$ assigns a value $\mathrm{k_j}$ to each $\mathrm{Y_j} \in \mathrm{V}$.
- Then, the **joint interventional distribution** factorizes as follows

$$P(Y_1, Y_2, \dots, Y_n \mid \mathrm{do}(V \coloneqq k))$$
$$= \prod_{Y_i \notin V} P(Y_i \mid \mathrm{Pa}(Y_i)) \cdot \prod_{Y_j \in V} \mathbb{I}(Y_j = k_j)$$

Measuring Causality

Average Treatment Effect (ATE)

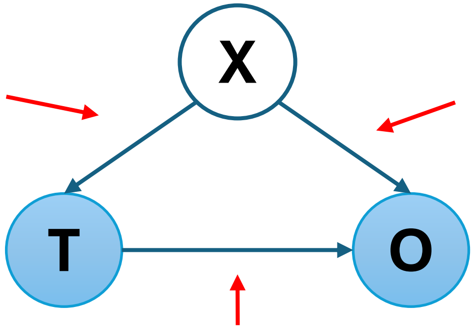
- Interventions are fundamental to study **causal effects**.
  - Does smoking causes cancer?
  - Will the vaccine avoid long-term infection?
  - How does the education level influence the average salary?
- Given a binary treatment variable T and an outcome variable O, the **average treatment effect** of T on O is

$$ATE(T, O) = \mathbb{E}_o[O|do(T = true)] - \mathbb{E}_o[O|do(T = false)]$$

ATE Interpretation in Treatment-Outcome-Covariate settings

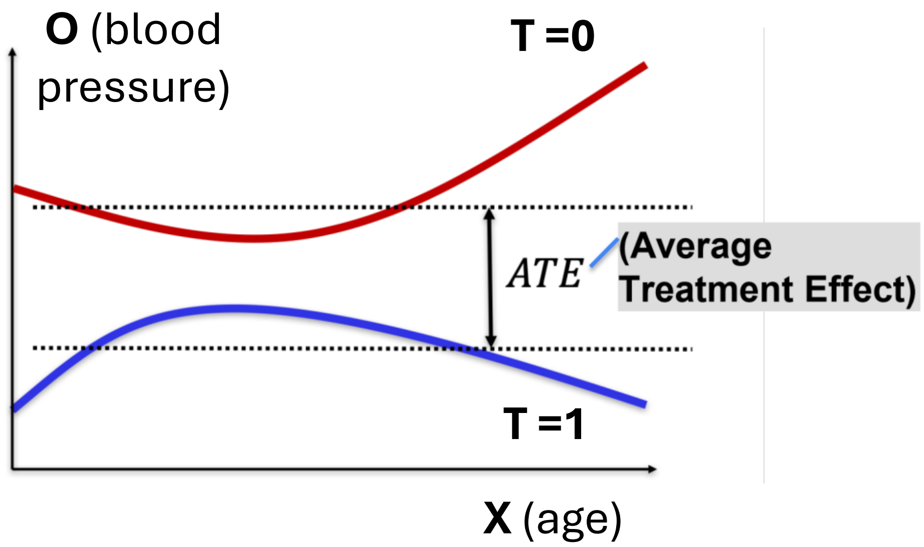
Health condition (observed or not)

Treat only upon certain health conditions



Overall health condition influences specific outcome

Treatment influences outcome



Is treatment effective? $\Rightarrow$ $ATE(T,O) = \mathbb{E}_o[O|do(T = true)] - \mathbb{E}_o[O|do(T = false)]$?

We need to cancel out the effect of X on O so that we can only focus on the $T \rightarrow O$ relationship, by computing the expectations above as:

$$\mathbb{E}_o[O|do(T = v)] = \mathbb{E}_o[\mathbb{E}_X[O|do(T = v), X]]$$

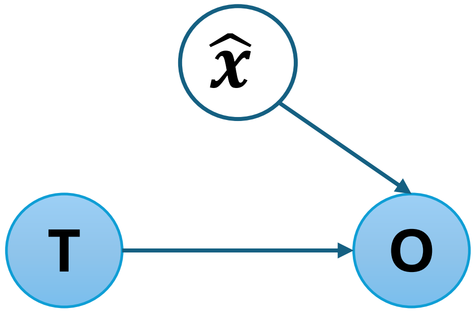
Need to observe X or to know P(X)

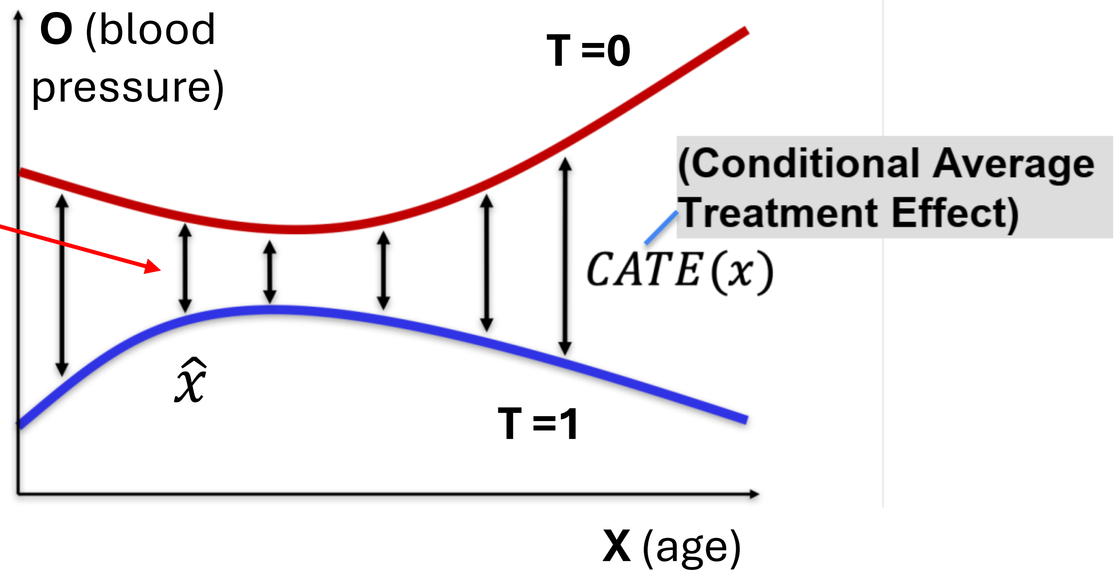
Conditional Average Treatment Effect (CATE)

Instead of marginalizing the covariate X we can fix to a value $\hat{x}$ (the full covariates or a part of it)

$$CATE(T,O) = \mathbb{E}_o[O|do(T = true), X = \hat{x}] - \mathbb{E}_o[O|do(T = false), X = \hat{x}]$$

Like computing ATE but only among people in a specific age group





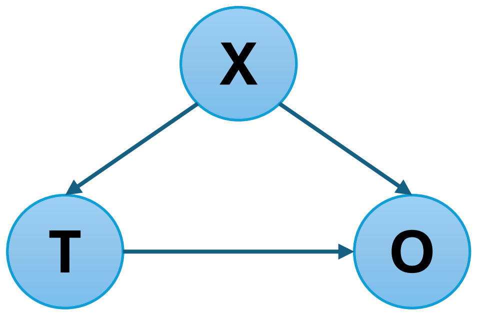
Causal Effect Identifiability

- The **causal effect** of a treatment T on an outcome O is **identifiable** whenever there exists an adjustment set X such that

$$\mathrm{P}(\mathrm{O} \mid \mathrm{do}(\mathrm{T})) = \mathrm{P}(\mathrm{O} \mid \mathrm{T}, \mathrm{X})$$

A.k.a: under what condition we are allowed to **measure causality from observed data**

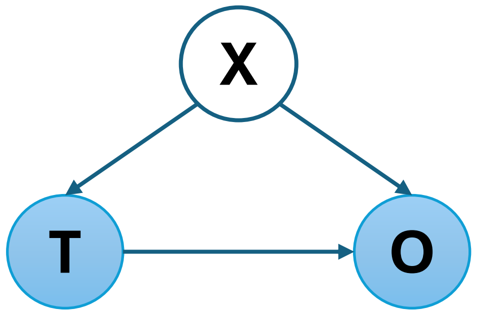
Measuring ATE/CATE



If all confounders X are observable and we have enough data, we can fit the probabilities needed to compute ATE/CATE

- $P(X)$
- $P(O|\, A, X)$

Using a neural network or other learning model

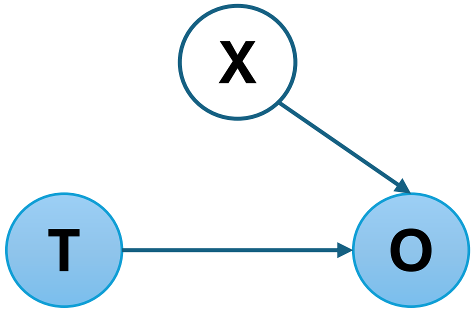


If there are (some) unobserved confounders we need a different strategy

- Unobserved confounders prevent a learner from identifying $P(O|\, A, X)$

Randomized control trials: simple and effective practice to allow causal effect estimation

Randomized Control Trial (RCT)



- To estimate the causal effect of T on the outcome O, we need to severe the edge from the confounder X to T
- Collect data according to the graph
  - The treatment decision T needs to be taken from a prior distribution P(T) (typically uniform)
  - Subjects X need to come from a general population without filtering or selection (prior P(X)), although there may be inclusion criteria to estimate CATE

Running the RCT

- Randomization - For each recruited subject X, we assign the treatment T, drawn from P(T) that is independent of X
- Trial - For the randomly assigned pair (T,X), we observe the outcome O by letting the world run and give us a sample (t, x, o)
  - In other words, we treat subject x with t and observe o
- We get a dataset $\mathrm{D} = \{(t_1, o_1), \dots, (t_N, o_N)\}$ and use it to estimate the ATE as

$$ATE(T = \hat{t}, O) \approx \frac{\sum_n \mathbb{I}\,(t_n = \hat{t})o_n}{\sum_n \mathbb{I}\,(t_n = \hat{t})}$$

- Something similar can be done for CATE

Causal Effect Identifiability

- The **causal effect** of a treatment T on an outcome O is **identifiable** whenever there exists an adjustment set X such that

$$P(O \mid do(T)) = P(O \mid T, X)$$

- The **do-calculus** is a complete system to find an adjustment set.
- From do-calculus, we can derive two fundamental adjustments:
  - The **back-door** criterion to handle observable confounders, and
  - The **front-door** adjustment to handle latent confounders.

Counterfactual Reasoning

- **Counterfactual** queries naturally occurs when we retrospectively reason on alternative outcomes **after** an intervention.
  - If the patient had received a placebo instead, would their recovery have been the same?
  - If the student had not studied the night before, would they still have passed the exam?
- Causal Bayesian Networks **cannot answer** counterfactual queries.
  - **Structural Causal Models can**, but we are leaving them out of the course for your own sake

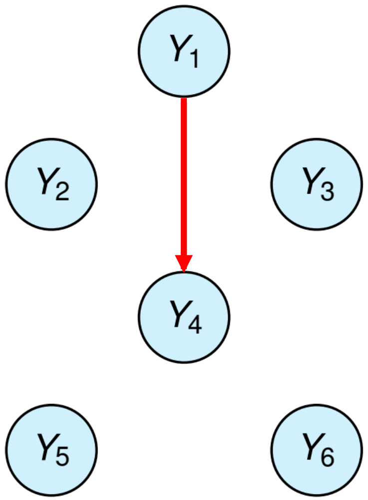
Structure Learning

Learning with Bayesian Networks

		Structure	
		Fixed Structure $P(Y\|X)$ $X \rightarrow Y$	Fixed Variables $P(X, Y)$ $X \quad Y$
Data	Complete	Naive Bayes Calculate Frequencies (ML)	Discover dependencies from the data Structure Search Independence tests
	Incomplete	Latent variables EM Algorithm (ML) MCMC, VBEM (Bayesian)	Difficult Problem Structural EM
		Parameter Learning	**Structure Learning**

The Structure Learning Problem

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1	2	1	0	3	4
4	0	0	0	1	2
…	…	…	…	…	…
…	…	…	…	…	…
0	0	1	3	2	1



- Observations are given for a set of fixed random variables
- Network structure is not specified
  - Determine which arcs exist in the network (causal relationships ⇒ causal discovery)
  - Compute Bayesian network parameters (conditional probability tables) or SCM parameters (structural functions)
- Determining the graph entails
  - Deciding on arc presence
  - Directing edges

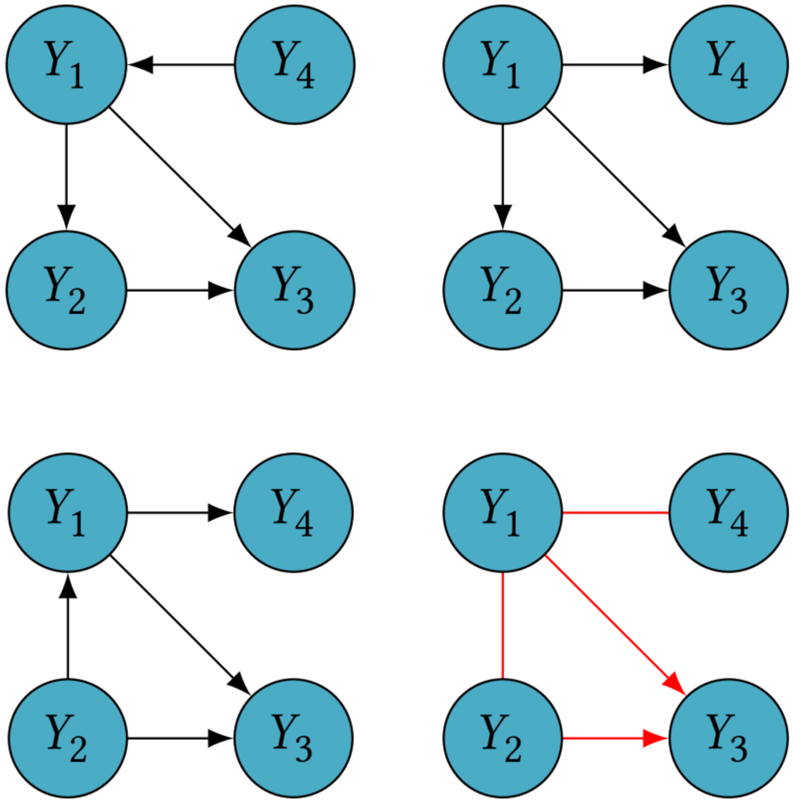
Structure Finding Approaches

- Constraint Based
  - Use **tests of conditional independence**
  - Constrain the network
- Search and Score
  - **Model selection** approach
  - Search in the space of the graphs

Markov Equivalence Class

- A **Markov Equivalence Class** (MEC) is a set of DAGs encoding the same set of conditional independences.

- Two DAGs are **Markov equivalent** if and only if they have the same **skeleton** and the same set of colliders (**v-structures**).



Constraint-Based Methods

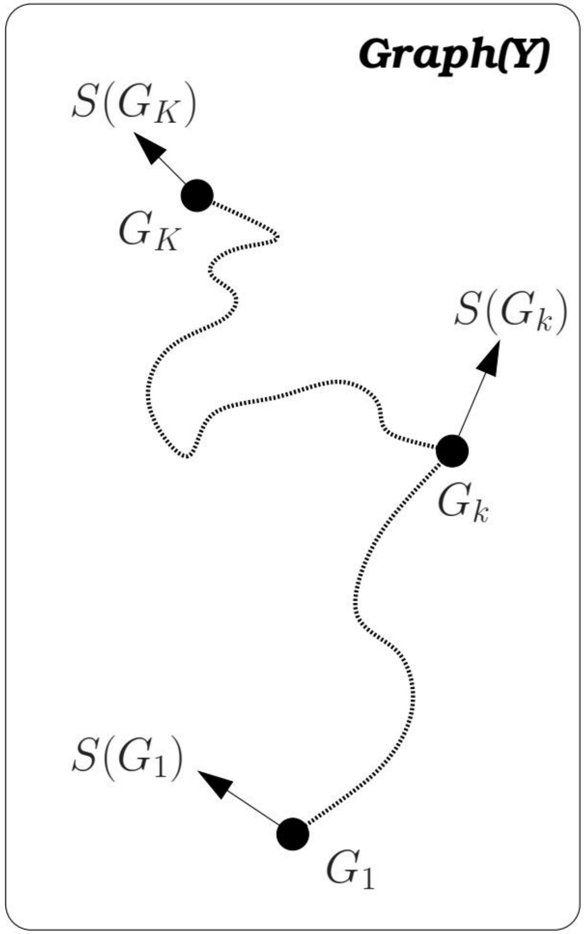
- We can reconstruct the Markov Equivalence Class by iteratively performing **conditional independence testing** $I(X_i, X_j|Z)$ (χ2-test, KCI-test, Fisher z-test, G-square test, …)
- Choice of the testing order is fundamental for avoiding a super-exponential complexity
- Level-wise testing - PC algorithm
  - Tests $I(X_i, X_j|Z)$ are performed in order of increasing size of the conditioning set Z (starting from empty Z)
  - Nodes that enter Z are chosen in the neighborhood of X_i and X_j

PC Algorithm: Skeleton

- The **PC** algorithm considers **separating sets** of **increasing size**.
- Worst case exponential cost, but much **better on average**!

$\mathcal{G} \leftarrow$ Fully connected CPDAG over V.
$K = 0$
while $K \leq |V|$ **do**
  for all Pairs (X, Y) in $\mathcal{G}$ **do**
    $A = \{Z \mid X - Z \text{ in } \mathcal{G}\} \setminus \{Y\}$
    for all $Z \subseteq A, |Z| \leq K$ **do**
      if $X \perp Y \mid Z$ **then**
        Prune $X - Y$ in $\mathcal{G}$.
      end if
    end for
  end for
  $K \leftarrow K + 1$
end while

Search & Score



- Search the space *Graph*(**Y**) of graphs G_k that can be built on the random variables $\boldsymbol{Y} = Y_1, \dots, Y_N$
- Score each structure by $S(G_k)$
- Return the highest scoring graph G^*
- Two fundamental aspects
  - Scoring function
  - Search strategy

Scoring Function

- Fundamental properties
  - Consistency - Same score for graphs in the same equivalence class
  - Decomposability - Can be locally computed
- Approaches
  - Information theoretic - Based on data likelihood plus some model-complexity penalization terms (AIC, BIC, MDL, …)
  - Bayesian – Score the structures using a graph posterior (likelihood + proper prior choice)

$$\log P(D|G) \approx \sum_{D} \sum_{X} \log \tilde{P}(x|\boldsymbol{pa}(x)) + \log P(G)$$

Search Strategy

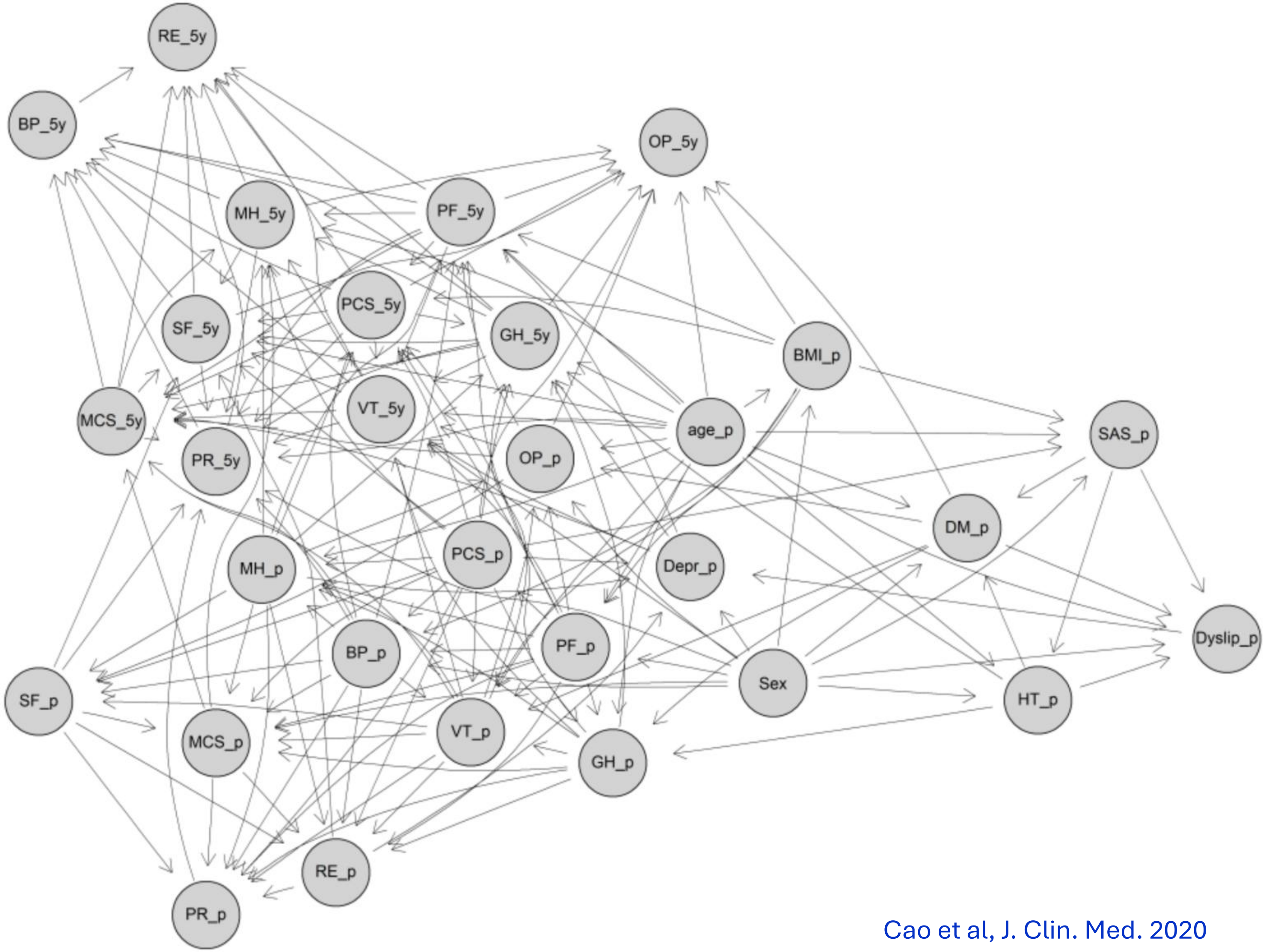
- Finding maximal scoring structures is NP complete (Chickering, 2002)
- Constrain search strategy
  - Starting from a candidate structure modify iteratively by local operations (edge/node addition or deletion)
  - Each operation has a cost
  - Cost optimization problem: greedy hill-climbing, simulated annealing, …
- Constrain search space
  - Known node order – Can reduce the search space to the parents of each node (Markov Blanket)
  - Search in the space of structure equivalence classes (GES algorithm)
  - Search in the space of node orderings (Friedman and Koller, 2003)

Hybrid Models

- Multi-stage algorithms combining previous approaches
- Independence tests to find a sub-optimal skeleton (good starting point)
- Search and score starting from the skeleton
  - Skeleton refinement
  - Edge orientation
- Max-Min Hill Climbing (MMHC) model
  - Optimized constraint-based approach to reconstruct the skeleton (Max-Min Parents and Children)
  - Use the candidate parents in the skeleton to run a search and score approach

Structure Learning in Healthcare

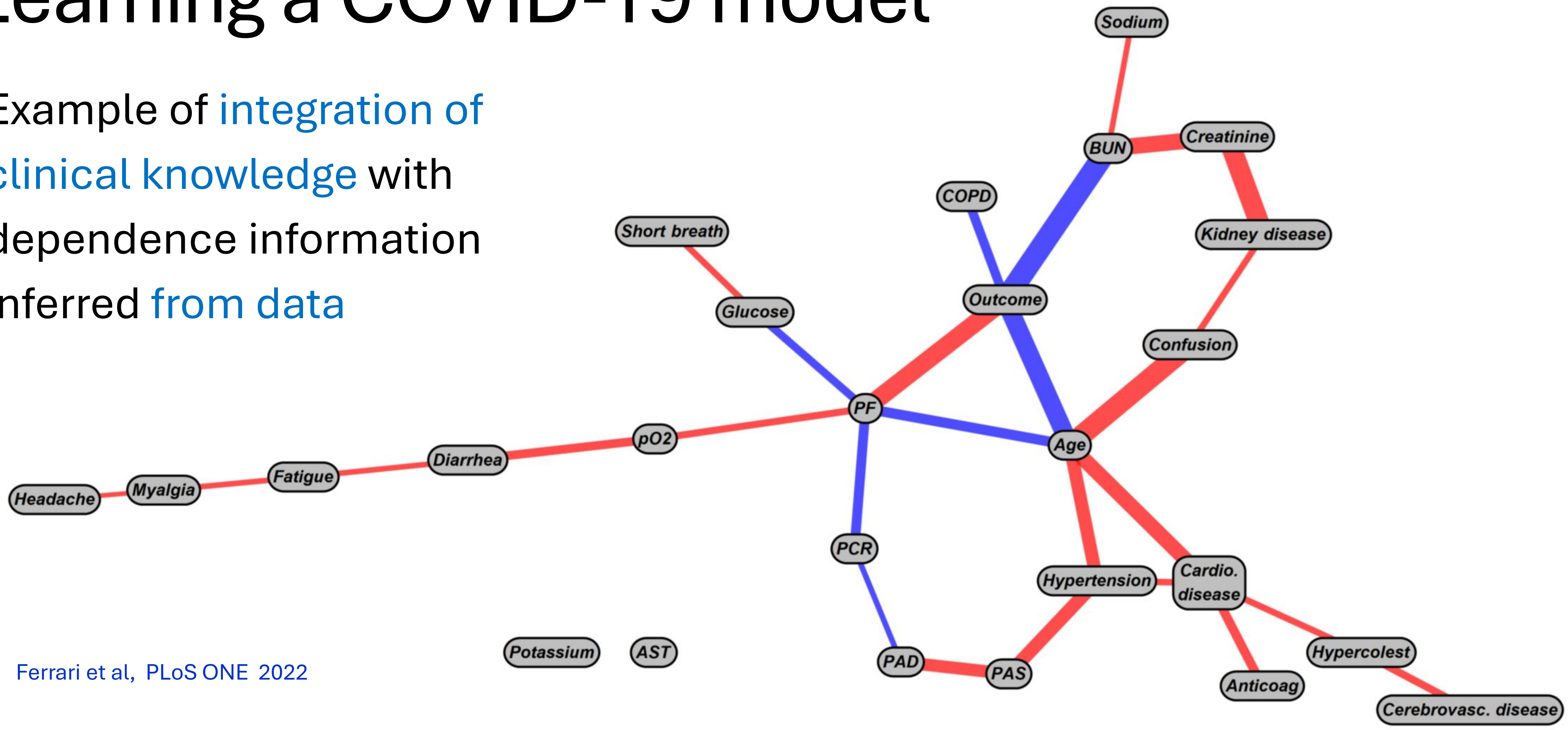
Learned structures: how interpretable?



Cao et al, J. Clin. Med. 2020

Learning a COVID-19 model

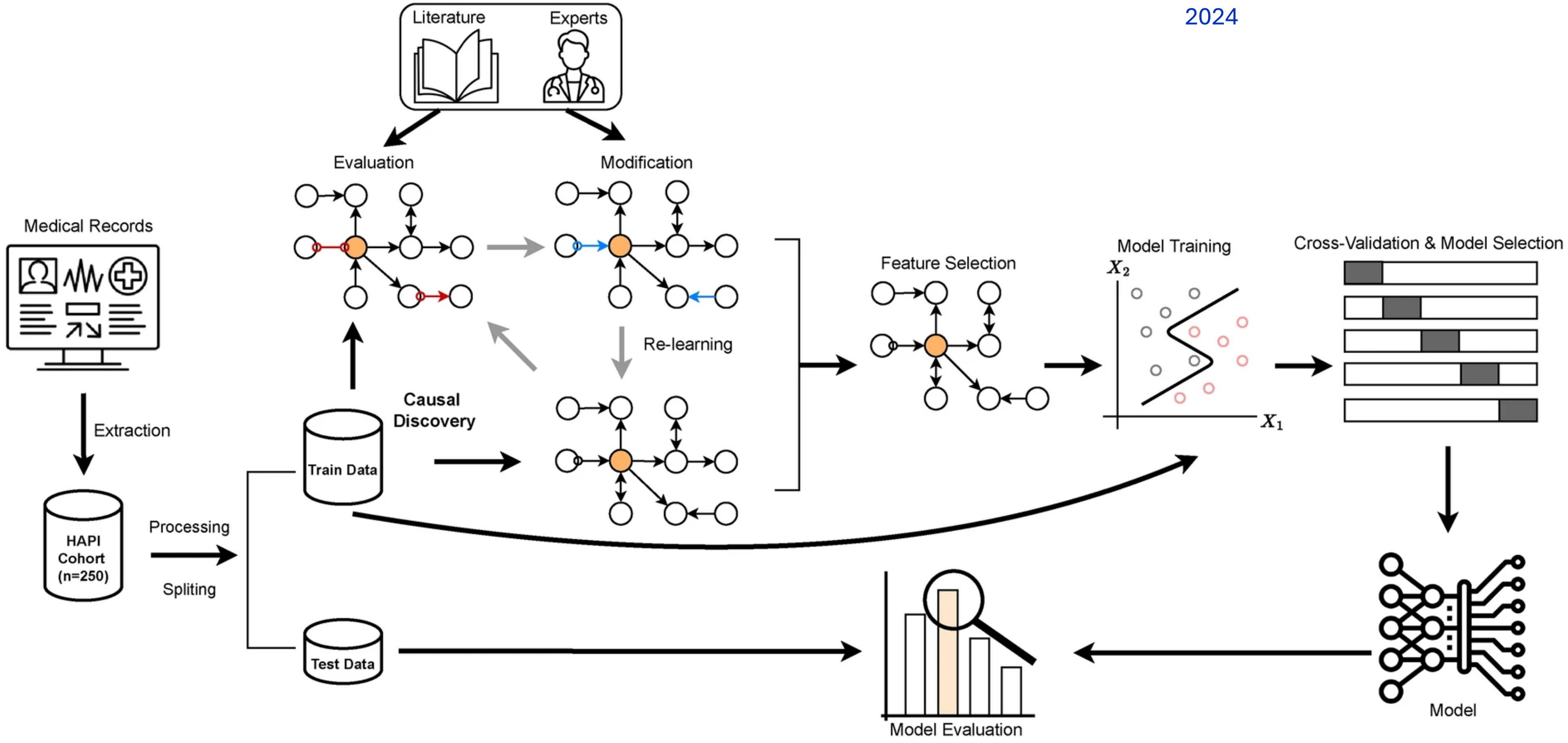
Example of integration of clinical knowledge with dependence information inferred from data



Ferrari et al, PLoS ONE 2022

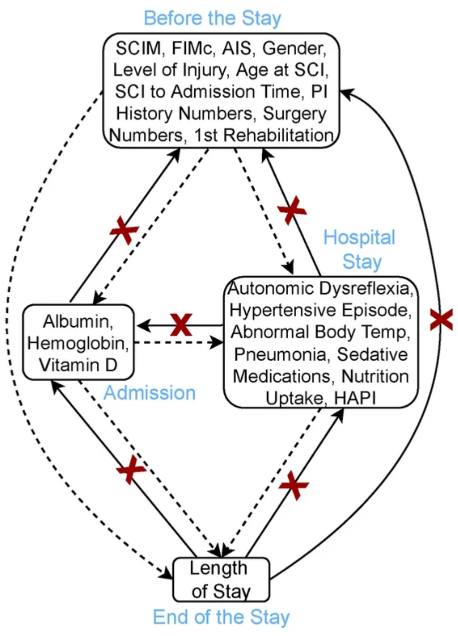
Prior/Expert Knowledge Incorporation

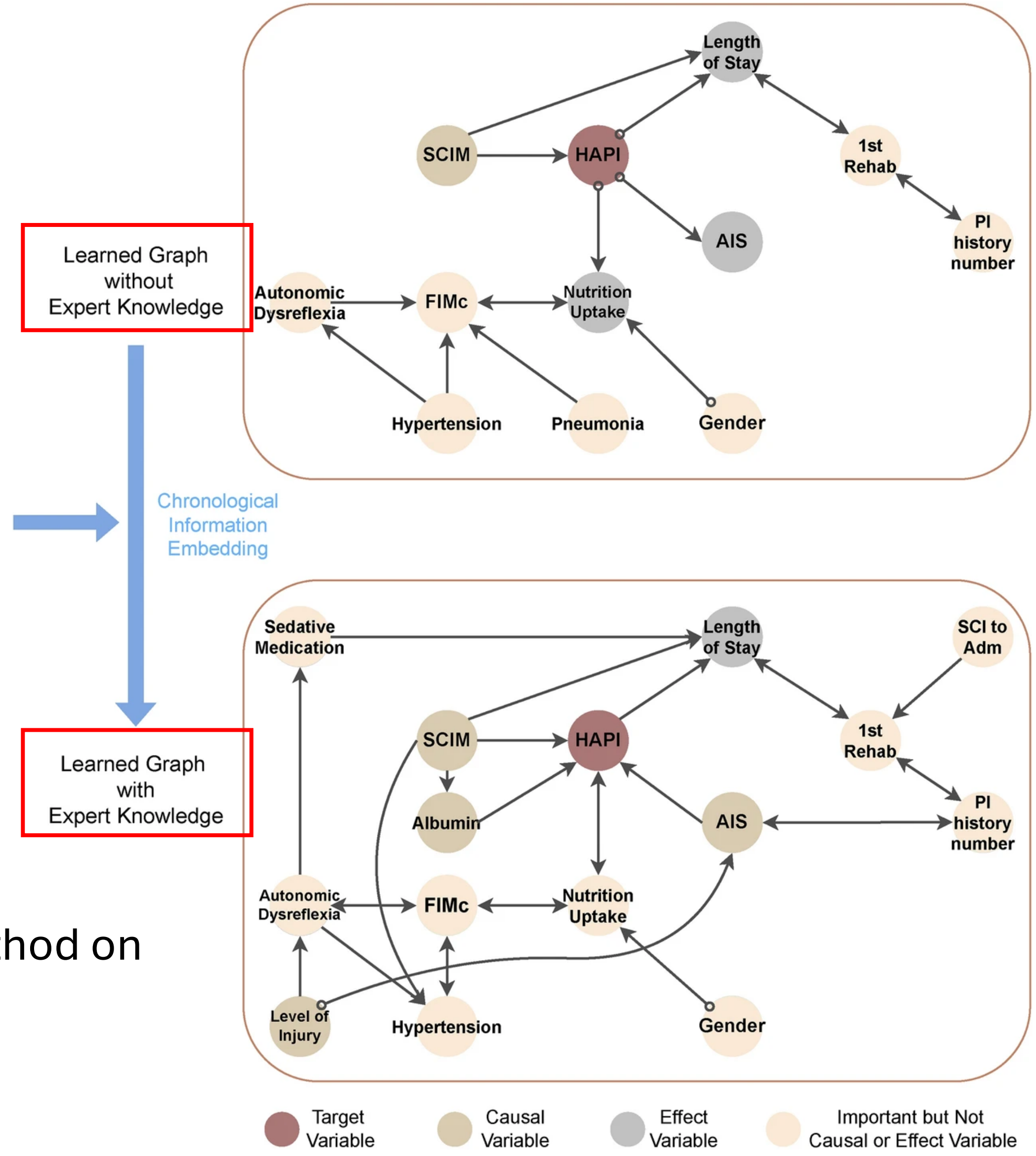
Li et al, Nature Sci. Reports, 2024



Embedding chronological knowledge

Expert knowledge on temporal ordering of variables (causal!)



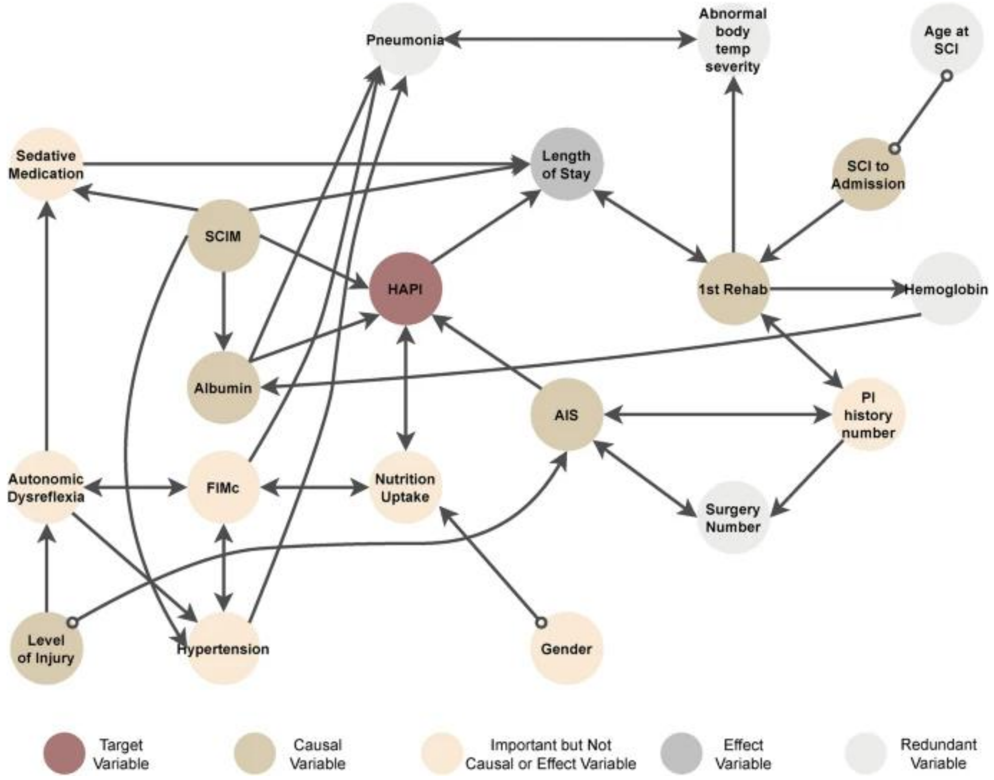


Constraint-based method on spinal cord injury data

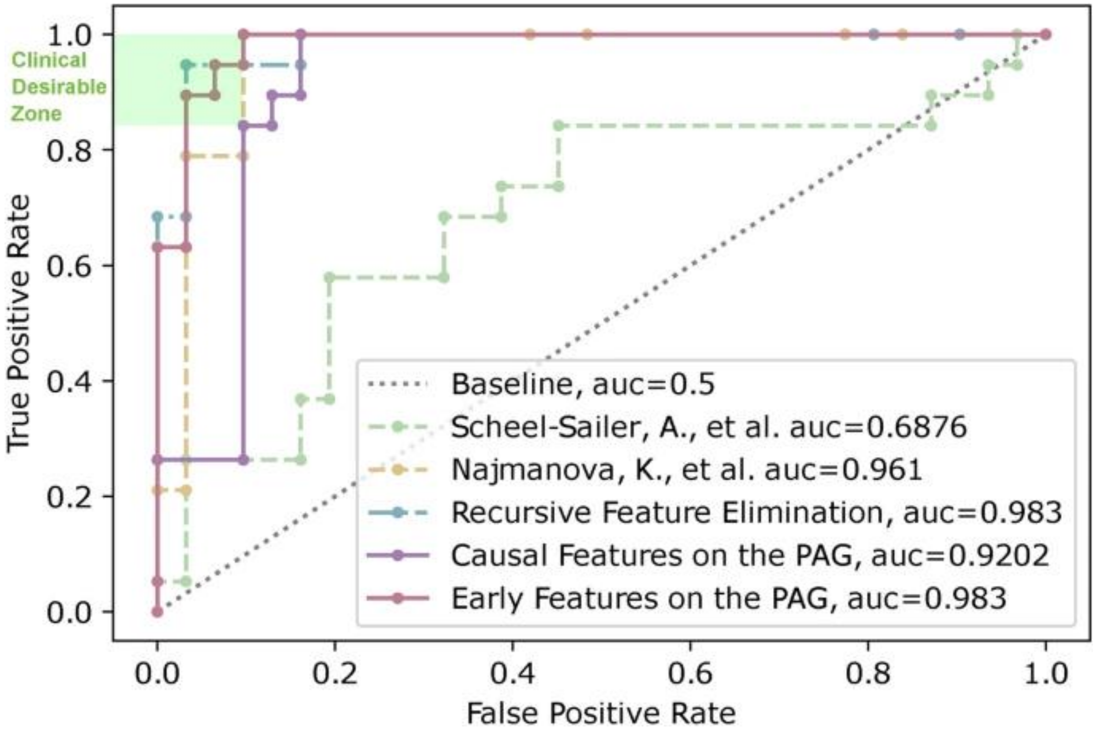
Li et al, Nature Sci. Reports, 2024

From Interpretation to Prediction

Final learned causal network



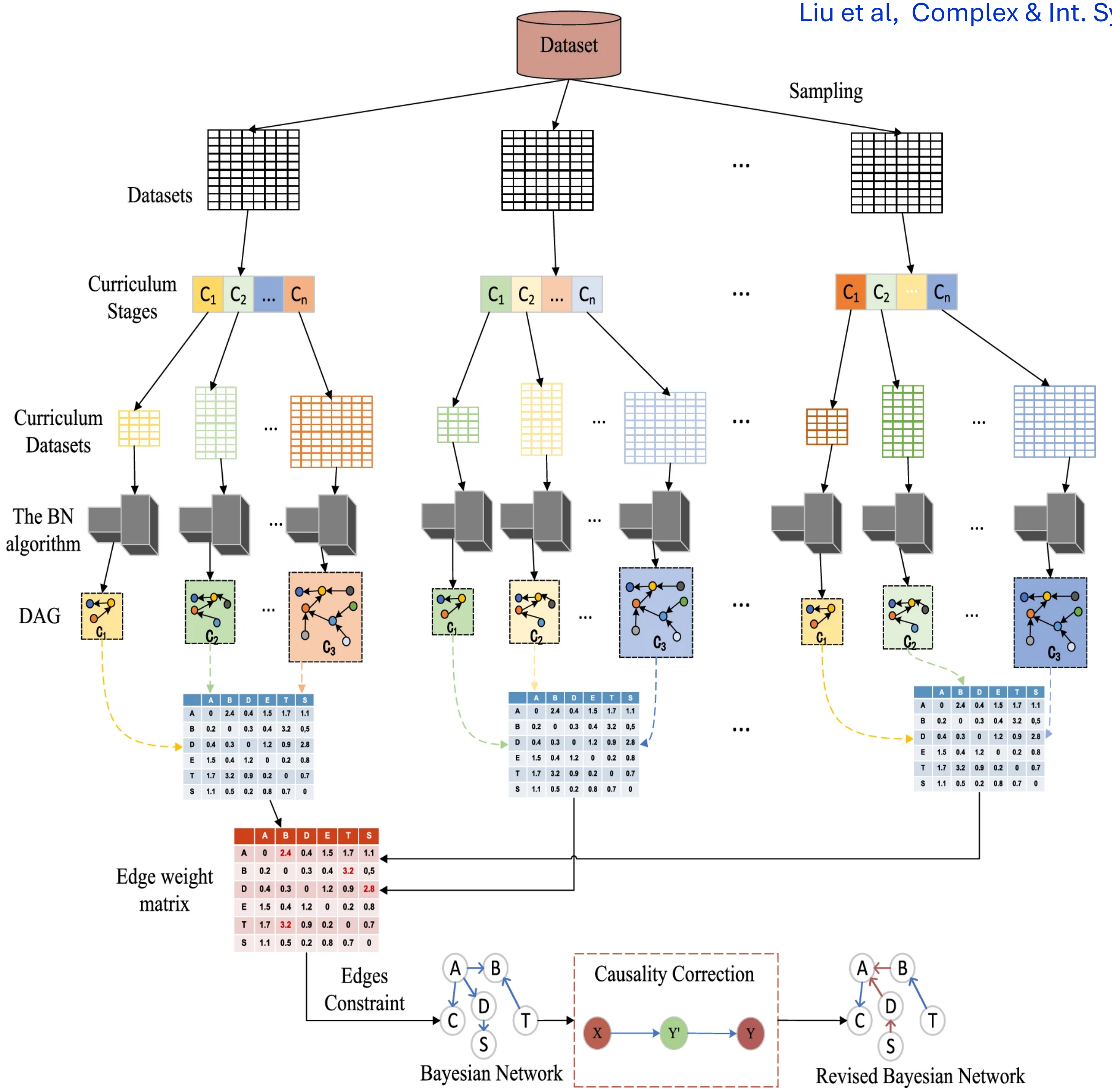
Predictive accuracy comparison



Li et al, Nature Sci. Reports, 2024

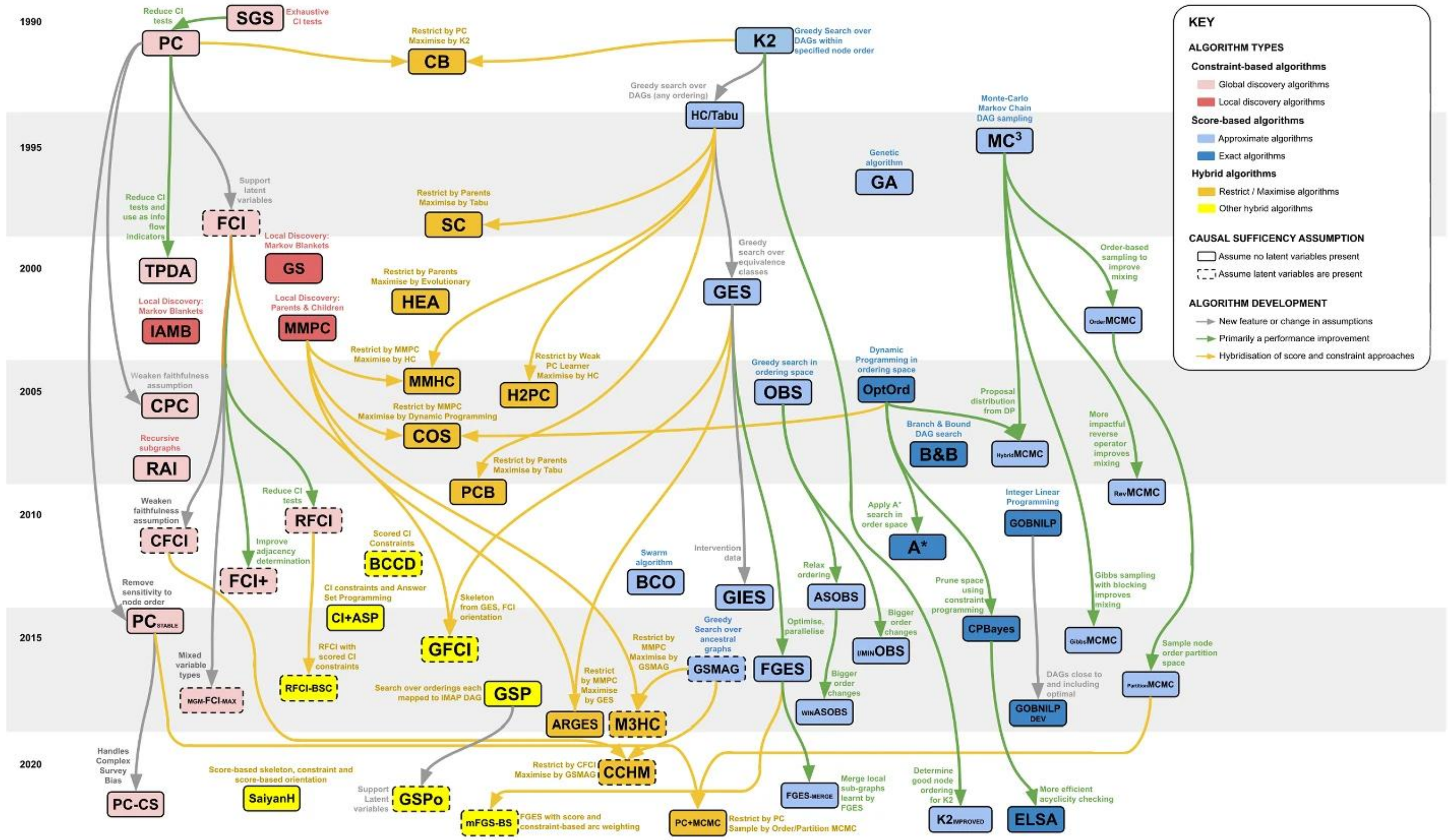
Structure Learning with Causality Correction

Liu et al, Complex & Int. Sys, 2024



Wrap-up

Structure Learning is a Bustling Field



Kitson et al, Artificial Intelligence Review, 2023

With Much Code Support

- PyWhy – A full Python-based ecosystem for causal learning
- CausalLearn – Python-based structure learning package
- DoWhy - Causal effect estimation and causal reasoning in Python
- pgmpy - Python package for causal inference and probabilistic inference with Bayesian Networks
- Bnlearn - the most consolidated and efficient library for BN structure learning (in R)

Take Home Messages

- The "**ladder of causation**" determines the relation between models and queries on a system:
  - **Probabilistic** Queries $P(Y_2|Y_1)$ → **Bayesian** Networks
  - **Interventional** Queries $P(Y_2|do(Y_1))$ → **Causal Bayesian** Networks
  - **Counterfactual** Queries $P(Y_2|do(Y_1), Y)$ → **Structural Causal** Models
- When they are **identifiable**, different causal models provides a solution to **answer causal queries** such as the **effect of treatments on outcomes**
- Randomized control trials provide a straightforward way to estimate causal effects
  - Costly and ethically challenged
  - Causal ML as a sustainable way forward
- **Learn** Bayesian/causal networks from data
  - Can lead to informative and effective models
  - Need to be integrated with constraints/prior knowledge from healthcare specialists

Next Lectures

- Lab tutorial
- Fundamentals of deep learning
  - Learning to represent
  - Enabling factors (tricks and else) for deep learning
- Neural Autoencoders
  - The first deep neural network
  - Unsupervised learning with deep neural networks
  - AE tasks: anomaly detection, compression, denoising